2013

School Quality Capitalization Into Housing Prices In Minnesota And Pennsylvania

Jonathan Taylor Smith
University of Mississippi

SCHOOL QUALITY CAPITALIZATION INTO HOUSING PRICES IN MINNESOTA AND PENNSYLVANIA

A dissertation
submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in the Department of Economics
The University of Mississippi

by

JONATHAN TAYLOR SMITH

August 2013

ABSTRACT

This dissertation develops and estimates a spatial autoregressive with autoregressive errors model of housing prices that accounts for both the endogeneity of spatially-lagged housing prices and local school quality measured by performance on state standardized tests. Two datasets are used from Boyertown, PA and Minneapolis, MN. Homes are spatially weighted against each other using a k nearest-neighbor approach. School quality is thought to be endogenous because unobserved neighborhood amenities in the error term of a hedonic regression are very likely positively correlated with local elementary, middle, and high school quality. Following previous literature, the optimal instrument matrix is constructed as the Cochrane-Orcutt tranformed conditional means of the spatially-lagged housing prices and quality measures. As school quality is observed on a much lower frequency than housing prices, it is not possible to estimate the conditional mean of school quality using non-parametric methods as proposed previously in the literature. So in order to instrument the school quality variables, a parametric model in which school quality is a function of average home prices within its attendance zone and average home prices outside its attendance zone but still within the same school district is used. Three different methods are presented for estimating the conditional mean of the spatiall-lagged housing prices, one of which is new to the literature. I find that parametrically estimating school quality can cause issues when the number of observations on quality are low as in the PA dataset. Also results are not robust to different specifications of W as small changes in k can affect the estimates by a large amount.

DEDICATION

To my loving wife, Suzy, and my two beautiful daughters, Avery and Violet, whose patience, encouragement, and love have made this degree possible.

To my parents, David and Marla Smith, who instilled in me a passion for learning and sacrified so much so that I could be given every opportunity possible.

LIST OF ABBREVIATIONS

2SLS	Two Stage Least Squares
OLS	Ordinary Least Squares
GMM	Generalized Method of Moments
MDE	Minnesota Department of Education
ML	Maximum Likelihood
MN	Minnesota
NCLB	No Child Left Behind
PA	Pennsylvania
RHS	Right Hand Side
SAR	Spatial Autoregressive Model
SARAR	Spatial Autoregressive Model with Autoregressive errors
SEM	Spatial Error Model

ACKNOWLEDGMENTS

I will be forever indebted to my dissertation committee chairman, Dr. Walter Mayer. He has, with great patience and understanding, guided me through the graduate program and dissertation process. For his willingness to carefully answer my many questions and lead me to being a better scholar, I am very grateful.

Dr. John Conlon has offered many suggestions and improvements over the past months to make this a better scholarly work. His need to understand economics in the most intuitive way possible has forced me to become a better economist and I thank him for pushing me to a better understanding of the material.

The other members of my committee, Dr. William Chappell and Dr. Matthew Hill, have offered invaluable feedback on the research and writing process. I greatly appreciate their willingness to serve on my dissertation committee and how they have supported me these last few years.

I am thankful to FNC, Inc. of Oxford, MS for supplying the housing data for this project. They have been very generous with their data and incredibly responsive with requests for more data.

The Boyertown Area School District transportation office and the GIS office of the Minnesota Department of Education made the process of generating school zone boundaries a much easier process than it could have been.

TABLE OF CONTENTS

ABSTRACT ii

DEDICATION iii

LIST OF ABBREVIATIONS iv

ACKNOWLEDGMENTS v

LIST OF FIGURES ix

LIST OF TABLES x

1 INTRODUCTION 1

1.1 Brief History 2

1.2 Dissertation Description 3

2 LITERATURE 5

2.1 School Quality Literature 5

2.1.1 Early Work 5

2.1.2 Fixed Effect Approach 6

2.1.3 Repeat-Sales Approach 8

2.1.4 Two-stage Least Squares Approach 9

2.2 Spatial Literature 10

2.2.1 Kelejian-Prucha Approach 10

2.2.2 The Lee Critique 12

2.2.3 Additional Endogenous Variables 13

2.2.4 Spatial and School Quality Literature 14

3 MODEL **15**

3.1 A Naïve Model 15

3.2 A Spatial Model 16

3.2.1 Spatial Weights 17

3.2.2 Zone-Time Effects 18

3.3 Instrumental Variables Estimation 18

3.4 Estimating The Conditional Mean of School Quality 20

3.5 Estimating the Conditional Mean of Spatially Lagged Prices 22

4 DATA **24**

4.1 Pennsylvania Data 24

4.1.1 Quality Variable 26

4.1.2 Summary Statistics 27

4.1.3 Zone Heterogeneity 27

4.2 Minnesota Data 29

4.2.1 Quality Variable 31

4.2.2 Summary Statistics 34

5 RESULTS **36**

5.1 Non-spatial Models 37

5.1.1 OLS 37

5.1.2 Boundary Fixed Effect Model 39

5.2 Spatial Models 44

5.2.1 SARAR with Exogenous Quality 46

5.2.2 SARAR with Endogenous Quality and Parametric Estimates of $E(S|X)$ . . . 47

6 CONCLUSION **56**

BIBLIOGRAPHY **61**

Appendices **65**

A FULL RESULT TABLES **67**

A.1 MN Data 67

A.1.1 KP Model with Exogenous SQ 67

A.1.2 Method A 73

A.1.3 Method B 79

A.1.4 Method C 85

A.2 PA Data 91

A.2.1 KP Model with Exogenous SQ 91

A.2.2 Method A 97

A.2.3 Method B 103

A.2.4 Method C 109

B PSEUDO-REPEAT SALES APPROACH **116**

B.1 A Basic Repeat-Sales Model 116

B.1.1 A Numerical Example 117

B.2 The Problem with the Repeat-Sales Model 118

B.3 How Rezoning Helps 119

B.3.1 A Numerical Example 120

B.4 Pseudo-Repeat Sales 122

B.5 Monte-Carlo Results 123

C HOUSING SUBMARKET IDENTIFICATION **126**

C.1 Literature 126

C.2 Methodology 133

C.2.1 Genetic Algorithm 136

C.2.2 Initial Results 137

VITA **141**

LIST OF FIGURES

Figure Number		Page
1	Home Sales Jan. 2004 - Oct. 2008 (Pre-Rezoning)	25
2	Home Sales Nov. 2008 - Jan. 2011 (Post-Rezoning)	25
3	Map of MN Attendance Zones	31
4	MN Home Sales by School District	32
5	MN Home Sales by Attendance Zone	33
6	Solutions to equation (B.6) when there is a shock to SQ ($\nu = 50$) in the second period.	118
7	Solutions to equation (B.6) when $U_2 = 1.1U_1$.	119
8	Solutions to equation (B.6) when U, the utility the home draws from the park, increases by a factor of 1.1 in the second period and there is a rezoning of school attendance zones.	121

LIST OF TABLES

Table Number		Page
1	PA Data Set Variable Definitions	27
2	PA Dataset Summary Statistics	28
3	PA Data Set Zone Demographics	29
4	PA Data Set Summary Statistics by Attendance Zone	30
5	Test Score Measure and Category Measure Correlations	34
6	MN Data Set Variable Definitions	35
7	MN Dataset Summary Statistics	35
8	OLS Regression of ln($SalesPrice$) on Housing Characteristics and Exogenous School Quality Variables in PA Dataset	39
9	OLS Regression of ln($SalesPrice$) on Housing Characteristics and Exogenous School Quality Variables in MN Dataset	40
10	OLS Regression of ln($SalesPrice$) on Housing Characteristics, Boundary Fixed Effects, and Exogenous School Quality Variables in PA Dataset from Jan. 2004 - Oct 2008	41
11	OLS Regression of ln($SalesPrice$) on Housing Characteristics, School Boundary Fixed Effects, and Exogenous School Quality Variables in MN Dataset	42
12	OLS Regression of ln($SalesPrice$) on Housing Characteristics, Elementary School Boundary Fixed Effects, and Exogenous School Quality Variables in MN Dataset	43
13	OLS Regression of ln($SalesPrice$) on Housing Characteristics, Middle School Boundary Fixed Effects, and Exogenous School Quality Variables in MN Dataset	44

14 OLS Regression of ln($SalesPrice$) on Housing Characteristics, High School Boundary Fixed Effects, and Exogenous School Quality Variables in MN Dataset . 45

15 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Exogenous School Quality Variables in PA Dataset when $k = 20$ 47

16 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Exogenous School Quality Variables in MN Dataset when $k = 20$ 48

17 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in PA Dataset when $k = 20$ with $E(WP|X)$ as Method A 49

18 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in MN Dataset when $k = 20$ with $E(WP|X)$ as Method A 50

19 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in PA Dataset when $k = 20$ with $E(WP|X)$ as Method B 51

20 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in MN Dataset when $k = 20$ with $E(WP|X)$ as Method B 52

21 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in PA Dataset when $k = 20$ with $E(WP|X)$ as Method C 53

22 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in MN Dataset when $k = 20$ with $E(WP|X)$ as Method C 54

23 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Exogenous School Quality Variables in MN Dataset when $k = 5$ 67

24 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Exogenous School Quality Variables in MN Dataset when $k = 10$ 68

25 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Exogenous School Quality Variables in MN Dataset when $k = 15$ 69

26 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Exogenous School Quality Variables in MN Dataset when $k = 20$ 70

27 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Exogenous School Quality Variables in MN Dataset when $k = 25$ 71

28 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Exogenous School Quality Variables in MN Dataset when $k = 30$ 72

29 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in MN Dataset when $k = 5$ with $E(WP|X)$ as Method A 73

30 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in MN Dataset when $k = 10$ with $E(WP|X)$ as Method A 74

31 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in MN Dataset when $k = 15$ with $E(WP|X)$ as Method A 75

32 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in MN Dataset when $k = 20$ with $E(WP|X)$ as Method A 76
33 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in MN Dataset when $k = 25$ with $E(WP|X)$ as Method A 77
34 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in MN Dataset when $k = 30$ with $E(WP|X)$ as Method A 78
35 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in MN Dataset when $k = 5$ with $E(WP|X)$ as Method B 79
36 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in MN Dataset when $k = 10$ with $E(WP|X)$ as Method B 80
37 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in MN Dataset when $k = 15$ with $E(WP|X)$ as Method B 81
38 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in MN Dataset when $k = 20$ with $E(WP|X)$ as Method B 82
39 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in MN Dataset when $k = 25$ with $E(WP|X)$ as Method B 83
40 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in MN Dataset when $k = 30$ with $E(WP|X)$ as Method B 84

41 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in MN Dataset when $k = 5$ with $E(WP|X)$ as Method C 85
42 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in MN Dataset when $k = 10$ with $E(WP|X)$ as Method C 86
43 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in MN Dataset when $k = 15$ with $E(WP|X)$ as Method C 87
44 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in MN Dataset when $k = 20$ with $E(WP|X)$ as Method C 88
45 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in MN Dataset when $k = 25$ with $E(WP|X)$ as Method C 89
46 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in MN Dataset when $k = 30$ with $E(WP|X)$ as Method C 90
47 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Exogenous School Quality Variables in PA Dataset when $k = 5$ 91
48 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Exogenous School Quality Variables in PA Dataset when $k = 10$ 92
49 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Exogenous School Quality Variables in PA Dataset when $k = 15$ 93

50 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Exogenous School Quality Variables in PA Dataset when $k = 20$ 94
51 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Exogenous School Quality Variables in PA Dataset when $k = 25$ 95
52 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Exogenous School Quality Variables in PA Dataset when $k = 30$ 96
53 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in PA Dataset when $k = 5$ with $E(WP|X)$ as Method A 97
54 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in PA Dataset when $k = 10$ with $E(WP|X)$ as Method A 98
55 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in PA Dataset when $k = 15$ with $E(WP|X)$ as Method A 99
56 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in PA Dataset when $k = 20$ with $E(WP|X)$ as Method A 100
57 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in PA Dataset when $k = 25$ with $E(WP|X)$ as Method A 101
58 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in PA Dataset when $k = 30$ with $E(WP|X)$ as Method A 102

59 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in PA Dataset when $k = 5$ with $E(WP|X)$ as Method B 103
60 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in PA Dataset when $k = 10$ with $E(WP|X)$ as Method B 104
61 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in PA Dataset when $k = 15$ with $E(WP|X)$ as Method B 105
62 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in PA Dataset when $k = 20$ with $E(WP|X)$ as Method B 106
63 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in PA Dataset when $k = 25$ with $E(WP|X)$ as Method B 107
64 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in PA Dataset when $k = 30$ with $E(WP|X)$ as Method B 108
65 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in PA Dataset when $k = 5$ with $E(WP|X)$ as Method C 109
66 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in PA Dataset when $k = 10$ with $E(WP|X)$ as Method C 110
67 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in PA Dataset when $k = 15$ with $E(WP|X)$ as Method C 111

68 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in PA Dataset when $k = 20$ with $E(WP|X)$ as Method C 112

69 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in PA Dataset when $k = 25$ with $E(WP|X)$ as Method C 113

70 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in PA Dataset when $k = 30$ with $E(WP|X)$ as Method C 114

71 Changes in Price, Quality, and Unobserved Variables in Figure 8 122

72 Results of Monte Carlo Experiment on 2 Repeat Sales Estimators and 2 Pseudo-Repeat Sales Estimators 124

73 K-Means Algorithm Results 138

74 Genetic Algorithm Results 139

75 GA with K-Means Pre-cluster Results 139

Chapter 1

INTRODUCTION

This dissertation investigates the link between public school quality and housing prices using hedonic regressions that control for spatial correlation and endogenous school quality measures. School quality is a key determinant of housing demand at a given location and, consequently, housing prices. Anyone who has purchased a home in the last decade can attest to the role school quality plays in both the pricing and selection of homes, whether the buyer has children or not. Partly fueling this obsession with school quality is the No Child Left Behind Act (NCLB) that was passed in 2001. One of the major benefits of the law is the requirement of data on student progress and school performance to be provided to parents yearly. This data has enviably made its way to researchers. The Minnesota Department of Education has a section of their website devoted to multiple datasets ranging from performance to financial information to transportation[1] (as do most state departments of education). However, all this data means little if potential buyers have no access to it. Websites[2] have come to parents' rescue by aggregating the mountains of data and ranking schools based on different metrics (usually based on yearly standardized test score data). In areas where data is available, sites even match addresses to attendance zones so perspective home buyers can quickly sort their potential purchases by their assigned school's quality. It then stands to reason that school quality could potentially play a large determining role in

[1] http://education.state.mn.us/MDE/Data/

[2] http://www.schooldigger.com and http://www.greatschools.org are great examples.

a home's price.

1.1 Brief History

There have been numerous attempts to estimate the impact of school quality on housing prices since the pioneering study of Oates (1969). A standard approach uses a hedonic model in which individual housing prices are regressed on measures of school quality such as standardized test scores and on controls such as the observed characteristics of houses and the neighborhood. As argued by Black (1999), one challenge is that school quality is likely correlated with the error term which reflects neighborhood and demographic characteristics not fully captured by the sample data. For example, residents in better neighborhoods may be more willing and able to pay for better schools. Following Black (1999), one line of research controls for the endogeneity by confining the sample observations to houses near the boundaries of the school zones. The idea is that unobserved neighborhood quality should vary less than school quality along zone boundaries.

An alternative to the boundary approach is to find instrumental variables that are correlated with school quality but not housing prices. This is the approach taken by Gibbons and Machin (2003, 2006), Downes and Zabel (2002) and others. As noted by Nguyen-Hoang and Yinger (2011), however, the instruments used by these studies are difficult to defend. For example, Gibbons and Machin (2003) include school type while Downes and Zabel (2002) include the proportion of the population renting and school-aged. The validity of these variables as instruments is questionable since they may affect the demand for housing and, consequently, be correlated with price.

Another approach that can be used to motivate instrumental variables is based on models of spatial dependence. This approach has been largely neglected by the school capitalization literature. Spatial models have been widely used in a variety of fields since the early work of Anselin (1988). One advantage of spatial models is that spatially lagged exogenous cross-sectional variables can serve as instruments analogous to time-lagged exogenous variables in time-series regressions. The properties of estimators based on these instruments

have been extensively studied most notably by Kelejian and Prucha (1998, 2004, 2007), and Lee (2003, 2007). Such instruments may be easier to defend than instruments used in the non-spatial models in the capitalization literature. Justification of spatially lagged variables as instruments rests on assumptions that: a) prices in a specified neighborhood are correlated, and b) the routinely assumed exogenous characteristics of each house are also exogenous with respect to other houses. The "neighborhood" is specified by the choice of a spatial weight matrix which can be based on a variety of metrics including physical distance.

1.2 Dissertation Description

Using housing data from FNC, Inc. and standardized test results collected from the Boyertown School District in Pennsylvania for 2004 to 2010 and the Minneapolis - St. Paul metro area from 2009 to 2013, we estimate spatial hedonic regressions of housing prices in which school quality is treated as endogenous. This study appears to be the first to treat school quality as endogenous in spatial models of individual housing prices. One issue we encounter not addressed in previous papers is that one endogenous variable (school quality) is observed for a small number of clusters (school attendance zones) while the other (prices) is observed for a large number of individual units (houses). Consequently there are only a small number of observations to estimate the conditional mean of school quality for the optimal instrumental variable matrix. This also affects the estimation of the conditional mean of prices which depends on the conditional mean of school quality through a restriction of the type noted by Lee (2003). We address the problem with different estimation strategies that vary in their dependence on the conditional mean of school quality.

I found that the method used to estimate the conditional mean of the spatially-lagged housing prices along with the number of neighbors used in the weighting matrix can have a large effect on the size, sign, and significance of key estimates. Ranges of the spatial autocorrelation coefficient in the Minnesota dataset are from about 0.4 to 0.9 depending on the specification used. The effect of elementary school quality ranges from -0.22 to 0.77 in the Minnesota dataset. This can be interpreted as the effect of a one standard deviation change

in school quality affecting housing prices in its attendance zone by decreasing them 22 percent or increasing them 77 percent. As for the Pennsylvania dataset, spatial autocorrelation is found to be in the about 0.25 in specifications where it is significant. Elementary school quality is found to have a negative but insignificant effect in almost every specification in the PA dataset.

The remainder of the dissertation proceeds as follows: chapter 2 covers the school quality capitalization and spatial literature in more detail while chapter 3 specifies the econometric models and describes the estimators. Chapter 4 describes the two datasets and chapter 5 presents the estimation results. Finally chapter 6 concludes and points to possible future areas of research and extensions of the current work. Also included are several appendices that contain extra estimation result tables, preliminary work for possible future research, or estimation programs written for other sections.

Chapter 2

LITERATURE

This chapter endeavors to review and combine two streams of literature. The first section begins with a discussion of the early work in the school quality capitalization literature. It then examines the different estimation methods used in the literature to measure the effect of school quality on housing prices. The second section is a brief review of spatial econometrics from early work to current techniques. The chapter closes with a discussion of how spatial methods have been applied to the school quality capitalization problem.

2.1 School Quality Literature

2.1.1 Early Work

The beginning of the school quality capitalization literature hearkens back to Tiebout (1956) in which Tiebout proposes his model of local expenditures. In this model local governments provide different combinations of public goods in an effort to attract or discourage individuals from living in these towns. The different combinations of public services allow sorting based on individual preferences and provides a means for local governments to supply an optimum bundle of public goods to its citizens. In an effort to test the Tiebout model empirically, Oates (1969) studies several suburban neighborhoods of New York City and examines how the different levels of public goods and services affect home prices in the area. One of the major public services Oates discusses is primary and secondary education. While Oates focuses on educational inputs (expenditures per pupil) as opposed to outputs (test scores), he mentions this is purely because of the availability of annual expenditure data where none exist for educational outputs. Using an OLS regression, Oates finds a positive

and significant relationship between school expenditures and the price of homes. Even in this early paper, Oates notes that a standard OLS regression is misleading due to simultaneous-equation bias. That is, spending in the local school system is likely a function local income and therefore correlated with the error term in the regression equation. He therefore tries to address the problem by instrumenting school expenditures with a list of local demographic variables. Using this 2SLS approach, Oates still finds a positive and significant effect.

2.1.2 Fixed Effect Approach

Another issue plaguing researchers trying to measure the effect of school quality on residential property values is that of omitted variable bias. Areas with a high measure of school quality are also likely to have other high quality public goods and amenities. This is a natural conclusion of the Tiebout model where individuals who care a great deal about education likely self-select into areas with other high quality public goods. Black (1999) confronts this problem and pioneers a boundary discontinuity design in which only homes within a narrow band surrounding the attendance zone borders are used in the regression analysis. Black looks at home sales within three suburbs of Boston, Massachusetts. She restricts the dataset to homes within 0.15 miles[1] and assigns each home to its closest boundary. Fixed effects are then included in a hedonic regression for each boundary. This has the effect of holding constant unobserved neighborhood amenities that are assumed to be constant across the boundary. After controlling for other observable housing characteristics through the hedonic regression, this results in an estimate of the effect of school quality on housing prices that accounts for the omitted variable bias that exists in other methodologies. She found that a 5 percent increase in school test scores resulted in a 2.5 percent increase in housing prices which was half the size of the effect found when not including the boundary fixed effects. This lends credibility to the idea that the boundary fixed effect method reduces the omitted variable bias (which should bias estimates away from zero due to amenities and school quality being positively related).

[1] Black (1999) also looks at homes within 0.20 and 0.35 miles, but results are similar to 0.15 miles.

Bayer *et al.* (2007) extends Black (1999) by recognizing that these zone boundaries, even if constructed through homogeneous regions at their conception, will eventually provide a natural setting for household sorting based on owner demographics. They provide strong evidence to show that on any given attendance zone boundary, households with more education and income are more likely to exist on the side of the border with the better school. While Bayer *et al.* are not using transaction data, they are able to use restricted-level census data which identifies long-form census responders to the census block level. This allows Bayer *et al.* to impute housing values based on responses to questions (which they acknowledge will include a fair amount of measurement error) and assign sociodemographic information to each home. By including sociodemographic information along with the boundary fixed effects, Bayer *et al.* maintain that the sorting behavior previously discussed is controlled for and less biased estimates can be computed. They find that like Black (1999) the simple inclusion of boundary fixed effects significantly lowers the coefficient on school quality. However, by also including the sociodemographic data the coefficient is reduced even farther. The final estimate gives an increase of one standard deviation of test scores raises housing prices only by 0.02 percent.

In essence, the boundary approach controls for the likely omitted variable bias with the assumption that unobservable factors in the error term are constant along a boundary and, thereby, uncorrelated with school quality in the restricted sample. This entails two costs. The first is a drastic reduction in the sample size. Since the majority of houses are not located along boundaries, considerable information is lost on the covariation of school quality and housing prices. Moreover, if certain unobserved neighborhood characteristics are more likely to occur along boundaries, then there is also selection bias. Second, although it may be plausible that neighborhood effects are approximately constant along zone boundaries, consistent estimation of the coefficients requires a more restrictive assumption, namely that there is no variation in factors that would be otherwise correlated with school quality. This is a very restrictive assumption. Bayer *et al.* (2007) make the point that in an ideal world,

researchers would compare two homes on opposite sides of a street that is bisected by an attendance boundary. However, in the real world researchers must look at homes within a band along the boundary up to 0.35 miles on each side in some studies (Black, 1999). While Bayer *et al.* focus on the issue of the width of the band around the boundary, another compounding problem is it's length. Some boundaries could be miles in length and therefore having to assume that neighborhood effects are constant along the *entire* border is a stretch of one's imagination.

2.1.3 Repeat-Sales Approach

One of the largest hurdles for the above fixed effect approaches to clear is that unobserved neighborhood and home characteristics can have large effects on a home's price and would therefore bias any estimates. Both Black (1999) and Bayer *et al.* (2007) try to capture the unobserved neighborhood effects by including the boundary fixed effects. While this likely captures some, if not most, neighborhood amenities, it does nothing to address the issue that homes in a better school zone may be more likely to have nicer details (granite counter-tops, crown molding, appliances, etc.) which would demand higher selling prices. In an effort to address this concern, Ries and Somerville (2010) employ a repeat-sales approach that removes time-invariant unobserved housing and neighborhood characteristics. They study home sales in Vancouver from 1996-2003. In September 2000, Vancouver schools were rezoned so that some homes in the district changed which elementary and/or secondary school children in that home would attend. This natural experiment provides Ries and Somerville their identification strategy. Imagine a home within a rezoned area that is sold once before the rezoning and once afterwards. Controlling for time trends and assuming there were no changes to the home in the time between sales, one can attribute any price change to the change in school quality. They control for neighborhood price trends through the use of localized price indexes in an effort to separate the changing neighborhood trends from changes in school quality. Repeat-sales methodology is usually plagued with small sample size issues. However, in this instance Ries and Somerville have 87,381 repeated

transactions, with 3,790 within the rezoned areas. Using this unusually large dataset, Ries and Somerville find little to no evidence for capitalization of elementary scores (and in some cases negative effects). Secondary scores do tend to have a positive effect of a 1.6 percent increase of housing prices with a one standard deviation increase in secondary quality.

2.1.4 Two-stage Least Squares Approach

Another method of controlling for omitted variable bias is to use a 2SLS approach. The trouble then becomes finding useful instruments for school quality. Gibbons and Machin (2003) face an interesting challenge in that they strive to tie school quality to housing prices using a dataset from the U.K. However, in the U.K. attendance zones are only mildly enforced and the system operates much closer to a "school choice" model. Therefore for a given postal code zone (their most disaggregated level of home information), local school quality is calculated as a weighted average of nearby primary schools based on distance to the school. Neighborhood amenities and other sociodemographic variables are not included in the regression equation directly, but are estimated non-parametrically through a methodology similar to a spatial Durbin model[2] where nearby exogenous variables are weighted based on distance or other measures and the averages are included in the regression function. In order to control for the endogenous school quality measure, Gibbons and Machin use the historical category of local schools (Community, Religious, or Controlled) and the age range of students in the schools as instruments of quality. They argue that these categories are so historical in nature they they are not affected by short-term conditions in the housing market nor by local sociodemographic conditions. They find that an increase of one percentage point of the number of children meeting specific testing goals raise surrounding home prices by 0.67 percent.

Like Gibbons and Machin (2003), Downes and Zabel (2002) do not work directly with home sales data, but with Chicago respondents to the American Home Survey which

[2] $y = X\beta + WX\theta + \epsilon$, where W is an nxn weighting matrix. See Elhorst (2010) for more on the spatial Durbin model.

are identified down to their census tract. So in order to construct a school quality measure, Downes and Zabel assign each school to a census tract and create a weighted average of the school quality measures of each school based on what percentage of the census tract their attendance zone covers. Another issue they face is that they only have respondent's self-valuation of the home's value and not a market price which can introduce bias if most people tend to overstate the price of their home. However, Downes and Zabel note that because the survey includes a random sampling of homes, their sample does not suffer from any selection bias which might exist due to qualitative differences in homes that are selling versus those that do not sell in a given time period. As instruments to school quality, they use a mixture of neighborhood characteristics and school characteristics (proportion of the tax base that is residential, proportion of the population that is school aged, per pupil assessed value, and the proportion of the population renting). As Ries and Somerville (2010) point out, these instruments may very well directly affect local home prices and therefore instrument validity is a major concern.

2.2 Spatial Literature

As the field of spatial econometrics is quite broad in scope, this review will concentrate on papers directly applicable to the model and estimation procedures put forth in later chapters.[3]

2.2.1 Kelejian-Prucha Approach

It seems intuitive that one determinant of a home's price would be the prices of the surrounding homes. With other home prices in the regression equation, however, home prices must then be considered endogenous. In that regard, spatial techniques have long been applied to the housing sector in economic research to control for spatial effects and to correct for the endogeneity. Anselin (2001) offers a thorough introduction to the topic by

[3]If the reader wants a more exhaustive view of the history of the field, Anselin (2010) covers the early development in the 1970s through modern implementations.

walking through the basic foundations of spatial econometrics and briefly discussing several different estimation procedures (ML, Spatial 2SLS, and GMM).

One of the earliest spatial models studied was that of a spatially autoregressive (SAR) model. Both Ord (1975) and Anselin (1988) study the SAR model and Kelejian and Robinson (1993) show that a natural choice of instruments to deal with the endogeneity of the spatially lagged home prices are spatial lags of the exogenous RHS variables. In order to understand their suggestion, it is useful to first examine the basic model. A SAR model has the following form:

$$y = X\beta + \lambda Wy + \epsilon. \tag{2.1}$$

Now, one can quickly solve the model for y by subtracting λWy from both sides, factoring out y and then multiplying by the inverse of the remaining term to get

$$y = (I - \lambda W)^{-1} (X\beta + \epsilon).$$

The term $(I - \lambda W)^{-1}$ is commonly referred to as the "spatial multiplier". However, to understand Kelejian and Robinson's suggestion of spatially lagged exogenous variables, it helps to not solve the equation as above, but to expand the right-hand side of the equation by recursively substituting the original model in for y. This leads to

$$y = X\beta + WX\lambda\beta + W^2X\lambda^2\beta + W^3X\lambda^3\beta + \ldots \tag{2.2}$$

and therefore their suggestion of spatially lagged exogenous variables as instruments for the endogenous Wy term makes perfect sense.

Kelejian and Prucha (1998) build on the SAR model by extending the earlier results to a spatially autoregressive with autoregressive errors (SARAR) model. Up to this point researchers had developed consistent estimators for SAR models and models with a spatial component to their error term, called spatial error models (SEM). However, there was no

consistent estimator for SARAR model which combined both of these issues (spatially lagged dependent variables and spatially lagged error disturbance) together. The SARAR model is generally

$$y = X\beta + \lambda Wy + u,\ u = \rho Wu + \epsilon. \tag{2.3}$$

Kelejian and Prucha (1998) develop a now widely used 4-step 2SLS procedure to estimate the coefficient of the spatially lagged dependent variable and spatial autoregressive parameter in the error term. This procedure is briefly

Step 1 Estimate $\tilde{\beta}$ and $\tilde{\lambda}$ using 2SLS with instrument matrix $[X, WX, W^2X, \ldots, W^pX]$.

Step 2 Estimate $\tilde{\rho}$ using GMM from the residuals of step 1.

Step 3 Perform a Cochrane-Orcutt type transformation to the variables so that error term is spherical using $\tilde{\rho}$ from step 2.

Step 4 Estimate $\hat{\beta}$ and $\hat{\lambda}$ using 2SLS on transformed variables with instrument matrix $[X, WX, W^2X, \ldots, W^pX]$.

In some instances, a fifth step is added to re-estimate the spatial error coefficient ($\hat{\rho}$) in order to get a more precise estimate, but is not required.

2.2.2 The Lee Critique

While the Kelejian and Prucha method produces asymptotically consistent estimators, they are not asymptotically efficient as shown by Lee (2003). Lee develops the best generalized spatial two-stage least squares (BGS2SLS) estimator. His procedure differs from that of Kelejian and Prucha (1998) only in the last step. In Lee (2003)

Steps 1–3 Same as Kelejian and Prucha (1998) steps 1–3.

Step 4 Estimate $\hat{\beta}$ and $\hat{\lambda}$ using 2SLS on transformed variables with instrument matrix $\tilde{H}_n^*$.

where

$$\tilde{H}_n^* = (I - \tilde{\rho}W)[X, W(I - \tilde{\lambda})^{-1}X\tilde{\beta}]. \tag{2.4}$$

Lee's instrument matrix is different from that of Kelejian and Prucha (1998) in two respects. First, he notes that the instruments should have the same Cochrane-Orcutt type transformation applied that is applied to other variables. Second, and more importantly, he notes that the SARAR model has a closed form optimum instrument matrix and so there is then no need to non-parametrically approximate it with Kelejian and Prucha's instrument matrix of $[X, WX, W^2X, \ldots, W^pX]$. It is by imposing this restriction given by the model[4] that Lee gains his efficiency.

2.2.3 Additional Endogenous Variables

Unfortunately, the literature of extending the spatial models described above to additional right-hand side endogenous variables is quite sparse. Drukker *et al.* (2013) is a very recent theoretical paper extending Kelejian and Prucha (1998) to allow for additional endogenous variables besides the spatially lagged dependent variable. They essentially handle the additional variables by using the same spatially lagged instrument matrix suggested by Kelejian and Prucha (1998). The reasoning behind the choice of also using this IV matrix for other endogenous variables is given as it "achieves a computationally simple approximation of the ideal instruments, which are given in terms of the conditional means of the RHS variables". This method is also mentioned by Elhorst (2010). Drukker *et al.* are also the author of a recently developed Stata module for spatial IV regression that implements this same methodology. It is therefore likely that there will soon be a more dense literature following this method. In a contrasting and more traditional approach, L. and N. (2008) use additional exogenous variables as instruments in addition to the already included spatially-lagged exogenous variables in their study of air quality effects on housing prices. Liu and Lee (2012) present a theoretical work studying the finite sample properties of large instrument sets that could arise when trying to control of additional endogenous variables in the spatial models and present a bias correction when instrument sets grow as the sample size increases.

[4] $(I - \rho W)y = (I - \rho W)(I - \lambda W)^{-1}X\beta + \epsilon$

2.2.4 Spatial and School Quality Literature

The neglect of the spatial approach by school capitalization literature is somewhat surprising since hedonic-regression studies of housing prices routinely adopt spatial models (see, for example, Dorsey *et al.*, 2010 and the survey by Hill, 2012). The only applications of hedonic spatial models to school quality appear to be Brasington and Haurin (2006, 2009); Sedgley *et al.* (2008). However, all of these studies treat the school quality variables as exogenous. Brasington and Haurin (2006, 2009) use a spatial hedonic model for different specifications of school quality in an effort to test which measure is being used by buyers and sellers as a measure of quality. They argue that by controlling for spatial effects, omitted variable bias is accounted for. This is because the spatial lag term "acts like a highly localized dummy variable capturing high localized influences common to just the nearest neighbors of each house." Sedgley *et al.* (2008) also investigate which measures of school quality are capitalized into home prices. They fail to address any endogeneity issue of their school quality measures. One possible explanation is that most applied spatial studies in all areas confine treatment of endogeneity to the spatially lagged variables and assume all other right hand side variables are exogenous. The only exception in the school quality literature is Fingleton and Le Gallo (2008). They specifically treat school quality as endogenous in their model. However in contrast to this study, their housing data is aggregated to the school district level and so does not contain individual transaction data.

Chapter 3

MODEL

3.1 A Naïve Model

To begin, it helps to start at perhaps the most simplistic and naïve point possible and build from there. Therefore, a good model to begin with is a simple hedonic model that ignores any endogeneity concerns. This simple model is specified as follows:

$$p_{ith} = X^*_{ith}\beta^* + s_{ht}\theta + \epsilon_{iht} \tag{3.1}$$

$$i = 1, \ldots, N(t);\ t = 1, \ldots, T;\ h = 1, \ldots, H$$

where X^*_{ith} is a vector of housing characteristics of home i in time period t in the school attendance zone h, p_{ith} is the natural logarithm of the sales price of that house, s_{ht} is the quality of the school in zone h at time t, and ϵ_{iht} is an idiosyncratic error. The vector X^*_{ith} consists of observed attributes such as the age of the home, gross living area, lot size, and the numbers of rooms, bedrooms, and bathrooms. It is important to note that in each time period a different set of homes are sold and so homes ith and ish usually correspond to different houses if $t \neq s$.

As described in the literature chapter above, this model grossly ignores two issues well known in the literature: housing values can have an effect on each other through spatial autocorrelation and the school quality variable is very likely correlated with unobserved neighborhood amenities and should therefore be treated as endogenous. By not controlling for these concerns, estimates of the coefficients β and θ will be biased and inconsistent.

Other researchers have attempted to confront these issues. For example, Black (1999) adds boundary fixed effects to the hedonic regression above (and restricts the dataset to homes within a certain distance the boundary) in the form of

$$p_{ithb} = X^*_{ithb}\beta^* + s_{ht}\theta + K_b\delta + \epsilon_{ithb} \tag{3.2}$$

where K_b is a vector of dummy variables that are equal to 1 when home i is near boundary b. This methodology is equivalent to calculating the expected average price of a home on different sides of boundaries and attributing the difference in price to the difference in school quality. As was stated above, this method still is likely to have biased coefficients due to neighborhood effects not being constant across the boundary, for example if there was gerrymandering of boundary lines around certain neighborhoods or, as Bayer *et al.* (2007) point out, sorting has occurred after the lines were drawn. Bayer *et al.* try to alleviate this problem by adding sociodemographic data to (3.2) but they still fail to account for the possibility of spatial autocorrelation in the data.

3.2 A Spatial Model

A modified version of (3.1) to account for first-order spatial autocorrelation in both the hedonic model and error term is

$$p_{ith} = X^*_{ith}\beta^* + \lambda \sum_{i \neq j} w_{ij} p_{ith} + s_{ht}\theta + \nu_{ht} + u_{iht} \tag{3.3}$$

$$i = 1, \ldots, N(t);\ t = 1, \ldots, T;\ h = 1, \ldots, H$$

where

$$u_{iht} = \rho \sum_{j \neq i} w_{ij} u_{jht} + \epsilon_{iht}. \tag{3.4}$$

In addition to the variables defined above, w_{ij} is the spatial weight corresponding to houses i and j and ν_{ht} is a "zone-time" effect that reflects unobserved factors that are common to

homes in zone h at time t. The problem is then to estimate β^*, λ, θ, and ρ from a sample of observed variables. In this model, both the prices of other homes and the school quality variables are allowed to be endogenous, while the observed attributes in X^*_{ith} are assumed to be uncorrelated with u_{iht}. Before moving on to how to address possible instruments for these endogenous variables, let us dive a bit deeper into how w_{ij} is constructed and the specification of ν_{ht}.

3.2.1 Spatial Weights

In the spatial literature, the number of methods to construct w_{ij} are almost as numerous as the papers themselves. Despite the lack of consensus of their form, the choice of spatial weights is very important. w_{ij} effectively defines the economic relationship between units (in this case, homes). Getis (2009) argues against the use of a simple contiguous weighting matrix. In a contiguous weighting matrix, all "neighboring" homes are equally weighted, irrespective of any differences in distance or other factors. In an earlier paper (Getis and Aldstadt, 2004) he favors an empirically generated weighting matrix. However, in this case we require the weights to be exogenous and so perhaps the most economically intuitive option is that of a decaying distance function, specifically,

$$w_{ij} = 1/d_{ij}^2 \tag{3.5}$$

where d_{ij} is the distance between houses i and j. In the light of a large dataset however, having every home related to every other home becomes extremely data intensive.[1] Therefore, in this project the number of "neighbors" are capped at some constant k that we allow to vary between 5 and 30 in increments of 5. This allows the use of a structure in Matlab called a "sparse" matrix which only stores non-zero elements.[2]

[1] The size of a matrix in Matlab is composed of two parts: the size of the array header and the data itself. The size of the header is equal to 112 bytes for each row in the matrix. The size of the data is 8 bytes for each cell. So therefore the total size of a matrix (in gigabytes) is $(8n^2 + 112n)/2^{30}$ which when $n = 4000$ is equal to 0.12Gb (or about 120 megabytes). However, when $n = 100000$ the size becomes 74.5Gb which is too large to store on almost any desktop computer

[2] Therefore, the size of a matrix when $n = 100000$ and $k = 5$ is only 14.5 *megabytes*.

3.2.2 Zone-Time Effects

As stated above, the zone effect ν_{ht} reflects unobserved factors that are common to houses in zone h at time t. These include unobserved aspects of school quality and neighborhood quality and demographics. Without additional restrictions, ν_{ht} cannot be distinguished from $s_{ht}\theta$ and, consequently, the main coefficients of interest are not identified. One "solution" is to relegate ν_{ht} to the error term under the assumption that it is uncorrelated with X^*_{ith}. However, this would be difficult to justify since ν_{ht} reflects in part unobserved aspects of school quality that may be correlated with variables in X^*_{iht}. For example, larger houses might be more likely have more school age children and, consequently, homeowners that are more willing to support increased funding for schools. Another problem with relegating ν_{ht} to the error term is that estimators of (3.3) would then depend on HT-asymptotics. This would be undesirable since T and H are relatively small in our Pennsylvania application.

To identify the school quality coefficients without relegating ν_{ht} to the error term, we assume additive time and zone effects:

$$\nu_{ht} = \delta_t + \eta_h. \tag{3.6}$$

Under this assumption, the coefficients can be identified without restricting the correlation between X^*_{ith} and ν_{ht} by simply adding separate time and zone dummies to (3.3).

3.3 Instrumental Variables Estimation

To discuss estimation of (3.3), it would beneficial to write it in matrix notation. Let $zone_{iht} = 1$ if house i is in zone h at time t and 0 otherwise; $N = \sum_{t=1}^{T} N(t)$; X is an N by k matrix consisting of the observations on X^*_{ith} and the zone and time dummies from (3.6); β is the coefficient vector for X; S is an N by 1 vector of elements: $\sum_{h=1}^{H} zone_{iht}s_{ht}$; P, u and ϵ are N by 1 vectors of elements p_{ith}, u_{ith}, and ϵ_{ith}; and W an N by N weighting matrix of elements w_{ij}. Then (3.3) and (3.4) can be written as:

$$P = X\beta + \lambda WP + S\theta + u,\ u = \rho Wu + \epsilon \tag{3.7}$$

We assume the variables in X are exogenous and that the regressors satisfy the usual full rank condition:

Assumption 3.1. $E(\epsilon|X) = 0$ *and rank*$([X, WP, S]) = k + 2$

Under this assumption, any N by k+2 matrix of functions of X with full column rank can be used to construct consistent and asymptotically normal IV estimator of (3.7). Using the result of Amemiya (1977)[3] that given the model

$$y_i = h(X_i, \beta) + \epsilon_i$$

where $\exists Z_i$ s.t. $E(\epsilon_i|Z_i) = 0$, the optimal instrument matrix A^* takes the form

$$A^* = E\left(\frac{\partial \epsilon}{\partial \beta}\middle| Z_i\right)\left(E(\epsilon\epsilon'|Z_i)\right)^{-1} \tag{3.8}$$

So then, the optimal IV matrix for (3.7) has the following form[4]

$$H_{AE} = -(I - \rho W)[X, E(WP|X), E(S|X)]. \tag{3.9}$$

Using the same argument as Lee (2003), (3.7) implies the following form for $E(WP|X)$[5]:

$$E(WP|X) = W(I - \lambda W)^{-1}[X\beta + E(S|X)\theta] \tag{3.10}$$

The more closely the IV matrix approximates H_{AE}, the lower asymptotic variance of the estimator. Estimation of H_{AE} requires preliminary estimates of β, λ, ρ, θ, and $E(S|X)$. Under Assumption 3.1, consistent estimates of β, λ, ρ, and θ can be obtained from the 2SLS estimator of Kelejian and Prucha (2004, 2007), for example. Estimation of

[3] See also Newey (1990) and Wooldridge (2010, pg. 542)

[4] Equation (3.7) implies $\epsilon = (I - \rho W)(P - X\beta - \lambda WP - S\theta)$.

[5] Solving the reduced form of (3.7) gives $E(P|X) = (I - \lambda W)^{-1}(X\beta + E(S|X)\theta)$

$E(S|X)$ is less straightforward. Kelejian and Prucha (2004, 2007) propose approximating the conditional means of the endogenous variables with fitted regression values based on the cross products of the exogenous variables and spatial weights. To justify this approach, they note that if an equation such as (3.7) is part of a system of linear equations[6] then, given regularity conditions, the conditional mean vector of the endogenous variables has the form: $\sum_{j=0}^{\infty} W^j X \Pi_j$ where Π_j are reduced-form coefficients. The fitted regression values can be interpreted as nonparametric series estimates of the conditional means. Results on consistent series estimation are given by Newey (2007), for example. Recently, Lee and Liu (2010) derive the asymptotic distribution for the 2SLS estimator of a spatial autoregressive model when the number of instruments grows with the sample size. Their results cover the case in which a series of base functions is used to consistently estimate the conditional expectations of the endogenous variables.

3.4 Estimating The Conditional Mean of School Quality

The problem in the present case is that there is not enough observations to reliably estimate $E(S|X)$ non-parametrically. It is well known that nonparametric estimators converge slowly and, consequently, require very large sample sizes. Since school quality varies only over zone and time, there are only $HT = 49$ observations in the Pennsylvania dataset available to estimate these conditional means. For this reason, we will adopt simple parametric model for $E(S|X)$. The Minnesota dataset has a much larger value of H with 347 elementary schools and so $HT = 1041$ and while it might then be possible to estimate $E(S|X)$ non-parametrically, this is left for future research.

Housing characteristics in X affect school quality indirectly through housing prices which generate property taxes., a major source of revenue for schools. Hence, we will assume that school quality in zone h depends on the averages of within-zone and outside-zone prices:

$$s_{ht} = \pi_0 + \pi_1 \bar{p}_{h,t} + \pi_2 \bar{p}_{\sim h,t} + \xi_{ht} \tag{3.11}$$

[6] See equation (2.2).

where $\bar{p}_{h,t}$ and $\bar{p}_{\sim h,t}$ the sample averages of prices at time t within and outside, respectively, zone h, and the π's are unknown constants. Assuming $E(\xi_{ht}|X) = 0$, we have:

$$E(s_{ht}|X) = \pi_0 + \pi_1 E(\bar{p}_{h,t}|X) + \pi_2 E(\bar{p}_{\sim h,t}|X) \tag{3.12}$$

Consistent estimation of (3.12) requires consistent estimates of $E(\bar{p}_{h,t}|X)$, $E(\bar{p}_{\sim h,t}|X)$ and the π's. In order to create estimates of $E(\bar{p}_{h,t}|X)$ and $E(\bar{p}_{\sim h,t}|X)$, first substitute (3.11) into (3.3) to get the following

$$p_{ith} = X_{ith}\beta + \lambda \sum_{i \neq j} w_{ij} p_{ith} + \pi_1 \theta \bar{p}_{h,t} + \pi_2 \theta \bar{p}_{\sim h,t} + \theta \xi_{ht} + u_{ith} \tag{3.13}$$

In matrix notation the above becomes

$$P = X\beta + \lambda WP + \pi_1 \theta \bar{P}_{h,t} + \pi_2 \theta \bar{P}_{\sim h,t} + u,\ u = \rho Wu + \epsilon \tag{3.14}$$

or more compactly,

$$P = X\beta + \lambda WP + \pi_1^* \bar{P}_{h,t} + \pi_2^* \bar{P}_{\sim h,t} + u,\ u = \rho Wu + \epsilon \tag{3.15}$$

It is then possible to estimate the π^*'s using the above equation.

If then M and R are defined as weighting matrices such that $\bar{P}_{h,t} = MP$ and $\bar{P}_{\sim h,t} = RP$, (3.15) can be rewritten as

$$P = (I - \lambda W - \pi_1^* M - \pi_2^* R)^{-1} X\beta \tag{3.16}$$

Estimates of $E(\bar{P}_{h,t}|X)$ and $E(\bar{P}_{\sim h,t}|X)$ can then be computed by

$$\hat{E}(\bar{P}_{h,t}|X) = M(I - \hat{\lambda} W - \hat{\pi}_1^* M - \hat{\pi}_2^* R)^{-1} X\hat{\beta} \tag{3.17}$$

and

$$\hat{E}(\bar{P}_{\sim h,t}|X) = R(I - \hat{\lambda}W - \hat{\pi}_1^* M - \hat{\pi}_2^* R)^{-1} X\hat{\beta}. \tag{3.18}$$

Using these estimates, it is then possible to directly estimate the πs from (3.12) by regressing S on $\hat{E}(\bar{P}_{h,t}|X)$ and $\hat{E}(\bar{P}_{\sim h,t}|X)$. Doing so then allows the computation of $\hat{E}(S|X)$ as

$$\hat{E}(S|X) = \hat{\pi}_0 + \hat{\pi}_1 \hat{E}(\bar{p}_{h,t}|X) + \hat{\pi}_2 \hat{E}(\bar{p}_{\sim h,t}|X) \tag{3.19}$$

Now having estimates of $E(S|X)$, constructing estimates of $E(WP|X)$ is as simple as substituting the initial estimates of β, θ and λ along with the estimates of $E(S|X)$ into (3.10) and finally it is possible to construct the optimal instrument matrix H_{AE}

It is important to note that consistency of the IV estimator of (3.7) is robust to the misspecification of (3.11) and (3.12). In particular, when (3.11) and (3.12) are misspecified they still generate consistent estimates of the linear projections of S. Since the latter does not, in general, coincide with the conditional mean when (3.11) and (3.12) are misspecified, the IV estimator of (3.7), while still consistent, is asymptotically less efficient.

3.5 Estimating the Conditional Mean of Spatially Lagged Prices

While $E(S|X)$ must be estimated parametrically because HT is relatively small, there are more options for the estimation of $E(WP|X)$ in (3.10). They include:

A Estimate $E(WP|X)$ non-parametrically (using cross-products-series regression of WP on linearly independent columns of $[X, WX, W^2X, \ldots]$) without restriction (3.10) imposed.

B Estimate $E(WP|X)$ parametrically with restriction (3.10) imposed and using the parametric estimates of E(S|X).

C Regress $WP - W(I - \hat{\lambda}W)^{-1}X\hat{\beta}$ on linearly independent columns of $[X, WX, W^2X, \ldots]$ and compute the predicted values from this regression, $\hat{\Gamma}$ say. Then estimate $E(WP|X)$ using $WP = W(I - \hat{\lambda}W)^{-1}X\hat{\beta} + \hat{\Gamma}$

The advantage of method A is robustness since the estimator does not depend on the specifications of $E(S|X)$. Even if the specification is correct, the estimate of $E(S|X)$ might not be precise because HT is relatively small. The disadvantage of method A is that it neglects restriction (3.10) which might result in an imprecise estimate. The advantage of method B is that it incorporates more restrictions into the estimation while the disadvantage is that it depends on the estimates of $E(S|X)$.

Method C is a nonparametric estimator of $E(WP|X)$ that incorporates the restriction (3.10). To motivate method C, note that (3.10) implies:

$$WP - W(I - \lambda W)^{-1}X\beta = W(I - \lambda W)^{-1}E(S|X)\theta + \mu$$

where by construction $E(\mu|X) = 0$ and, therefore, given consistent estimates for λ and β

$$WP - W(I - \hat{\lambda} W)^{-1}X\hat{\beta} = W(I - \lambda W)^{-1}E(S|X)\theta + \mu + o_p(1)$$

Consequently, the regression estimator in C can be viewed as a series estimator of

$$W(I - \lambda W)^{-1}E(S|X)\theta \tag{3.20}$$

and the $\hat{\Gamma}$ corresponding predicted values. In contrast to the estimates of the conditional means based on (3.12) there are N observations available to estimate (3.20). Since $(I - \lambda W)^{-1}$ is the spatial multiplier, (3.20) can be interpreted as the average total spatial effect of expected school quality. It consists of linear combinations of the elements of $E(S|X)$ with coefficients that vary over houses.

Chapter 4

DATA

4.1 Pennsylvania Data

Boyertown Area School District is located about an hour northwest of Philadelphia, Pennsylvania. The school district itself covers over 100 square miles and services over 3000 elementary school students. The data set is comprised of 3,728 home sales from January 2004 through January 2011. In October 2008 school district officials announced to the parents that they would be redrawing attendance zone boundaries for the seven elementary schools in the Boyertown Area School District. This rezoning was necessary due to overcrowding in some schools due to population growth. By rezoning students from these schools to ones with in the district with slower growth, the district was able to more effectively use its existing physical resources. This change creates two sets of attendance zone boundaries for this district over the timespan of the data. As the rezoning announcement was very public and well dispersed, it then stands to reason that buyers after this announcement would have the post-rezoning boundaries in mind when purchasing homes. Figures 1 and 2 show home sales from January 2004 - October 2008 and November 2008 - January 2011 respectively.

The data for this project was gathered from several different sources. The dataset of home sales and characteristics was provided by FNC, Inc. of Oxford, MS. FNC, Inc. is a technology services company that specializes in software and data related to the mortgage industry. School quality data was downloaded from the Pennsylvania State Board of Education's website. Attendance zone maps were obtained from the Boyertown Area School District Superintendent and Transportation offices directly.

Figure 1: Home Sales Jan. 2004 - Oct. 2008 (Pre-Rezoning)

Figure 2: Home Sales Nov. 2008 - Jan. 2011 (Post-Rezoning)

4.1.1 Quality Variable

Pennsylvania measures student achievement by a yearly standardized test. This test measures both mathematical and reading skills. Unfortunately in this data set the average test score is unavailable. However, each student is ranked as one of the following categories in both reading and mathematics: Advanced, Proficient, Basic, or Below Basic. The data obtained from the State of Pennsylvania gave the percentage of students in each school who scored at each level by subject area. There is no overall measure of school quality defined by the state. Ries and Somerville (2010) confront this same situation and they construct a measure of school quality based on these achievement bins. Vancouver schools only have three levels of success and so Ries and Somerville assign the highest achievement level 1 point, the lowest level -1 point and the middle level 0 points. First, a criticism of this approach is that it allows radically different schools to have the same score of quality. For example, a school with 50 percent of it's students in the low group and 50 percent in the high group would have a score of 0, as would any school where the same percentage of students scored in the high and low group (including when all students score in the middle group). So to tackle this problem, it was originally thought that changing the bin weights to all positive would solve issue. Unfortunately, this is not the case. As an example, for each percentage point of Advanced students, let the school receive 4 points. For scores of Proficient, Basic, and Below Basic, the schools receives 3, 2, or 1 point respectively. Therefore, this measure has a minimum score of 100 (all students scored "Below Basic") and a maximum score of 400 (all students scored "Advanced"). Now imagine a school with 25 percent of its students in each category. This school has a quality score of 250. This same score could be from a school with 50 percent of students scoring "Below Basic" (1 point) and the other 50 percent scoring "Advanced" (4 points). It seems then that any system that uses these categories to construct a school quality measure will fall to such criticism. However, as shown below in the Minnesota data section, when these categories exist alongside actual test data, the measures constructed as above are very highly correlated with the average test score of the

school which gives confidence in using the constructed measures in instances where actual test data does not exist. In this data set's case, the quality measure of both reading and math exams are constructed using the "4,3,2,1" weighting method previously described. Schools' math and reading summary scores are then averaged to obtain an overall measure of school quality. Finally, the scores are standardized so that the final quality measure can be viewed as deviations from the mean.

4.1.2 Summary Statistics

Definitions for the variables used in the econometric models are given in Table 1. Basic data quality procedures were followed to eliminate unusual observations. For example, if the age of the home was negative or if there were 0 rooms, the observation was dropped. This resulted in dropping 172 observations. Summary statistics for the remaining 3728 observations are shown below in Table 2.

Table 1: PA Data Set Variable Definitions

Variable	Definition
Sales Price	The sales price of the home in dollars
Age	The age of the home in years at time of sale
Rooms	The number of rooms in the home including living rooms, kitchen, etc.
Beds	The number of bedrooms in the home
Baths	The number of bathrooms in the home
GLA	The gross living area in square feet
Lot Size	The size of the housing lot in square feet
School Quality	A variable constructed from test scores on state reading and math standardized exams. A normalized version of this variable was used in the regression analysis.

4.1.3 Zone Heterogeneity

In order to study the heterogeneity of the school attendance zones within the Boyertown Area School District, data on educational attainment and median income were collected from the 2000 U.S. census at the block group level. These block groups were then matched

Table 2: PA Dataset Summary Statistics

Variable	Mean	Std. Dev.	Min	Max	Median	25th	75th
Sales Price	239691.3	93117.9	50000	1258073	225000	170000	300355
Age	29.97	36.18	0	259	16	3.5	46
Rooms	6.93	1.56	3	16	7	6	8
Beds	3.33	0.72	1	8	3	3	4
Baths	2.33	0.85	1	8	2	2	3
GLA	1959.5	755.6	576	5764	1789	1366	2415
Lot Size	28110.0	37083.5	144	217800	15000	7090.5	29185
School Quality	322.2	17.5	251.2	347.4	324.9	314.8	334.2
Std. School Quality	0	1	-4.06	1.44	0.15	-0.43	0.68

The standardized version of the school quality variable was used in the regression analysis.

to school district attendance zones and a weighted average based on the block group population was used to aggregate the census data into zone averages. These zone demographics are given in Table 3. The percentages in the educational attainment cells give the percentage of the zone population with each attainment level as their highest educational attainment respectively. We notice that even within this relatively small geographical area, there seem to be rather large disparities in median household income and educational attainment. The Gilbertville Elementary School (GES) zone has a median household income of $64,324 and 23 percent of its population has a college degree or higher. However, the Boyertown Elementary School (BES) zone has a median income of $43,655 and only 13 percent of its population has a college degree or higher.

In view of such large gaps amongst zones, we break our summary statistics down by attendance zone in Table 4. Again, there seem to be large differences between school attendance zones. Median sales price ranges from $174,000 in the Colebrookdale Elementary School zone (CES) to $278,000 in the GES zone. These median home prices are also strongly correlated with the census demographics given in Table 3. However, basic home characteristics (rooms, bedrooms, etc.) appear to remain fairly stable across the zones except for age for which there seems to be some zones with fairly new housing (GES and NHUF). The median age also appears strongly correlated with median sales price. This demonstrates the

Table 3: PA Data Set Zone Demographics

School	< HS	High	< College	College	Graduate	Median Household Inc.
BES	22%	46%	18%	10%	3%	$43,655.16
CES	21%	45%	21%	8%	5%	$46,306.04
EES	22%	45%	21%	8%	5%	$52,115.99
GES	12%	43%	21%	16%	7%	$64,234.47
NHUF	15%	42%	22%	15%	6%	$58,478.15
PFES	21%	45%	18%	10%	6%	$51,172.74
WES	19%	46%	18%	13%	4%	$54,221.68

All data is pulled from the 2000 US Census.

"< HS" refers to the percentage of the population who did not graduate high school.

"High" refers to those who graduated high school but did not attend any college.

"< College" refers to those who finished high school but did not graduate with a 4-year degree.

"College" refers to those who graduated from college with a 4-year degree but not a graduate degree.

"Graduate" refers to the percentage of the population with a graduate degree.

need for a hedonic pricing model which can control for age and other characteristics in order to measure the effect of school quality on housing prices.

4.2 Minnesota Data

The second and much larger data set used in this dissertation is from the Minneapolis - St. Paul (MSP) metro area and consists of homes sold from June 2009 through January 2013. The traditional MSP metro area consists of seven counties (Anoka, Carver, Dakota, Hennepin, Scott, Ramsey, and Washington) and this data set includes two additional counties (Sherburne and Wright) on the northwest corner of the metro area where growth has been high the last two decades. These nine counties cover 4142.36 square miles and have a combined population of 3,062,766[1]. Within this area exists 62 school districts, of which the Minneapolis School District and St. Paul School District are obviously the largest. These districts combine to include 347 elementary schools, 133 middle schools, and 92 high schools[2]. These schools combine to create 436 distinct attendance zones within the study area[3]. Figure

[1]From 2010 U.S. Census

[2]These counts are of the public schools that homes are zoned for and do not include magnet schools, special education centers, private schools, etc.

[3]There are more zones than elementary schools due to some elementary zones being split between different middle or high schools

Table 4: PA Data Set Summary Statistics by Attendance Zone

School	N		Price	Age	Rooms	Beds	Baths	GLA	Lot Size	Quality
		Mean	204705	41.8	6.7	3.3	2.2	1835	13089	315.2
BES	680	S.D.	81477	35.5	1.5	0.79	0.86	685	14830	6.72
		Median	184950	30	6	3	2	1600	9147	314
		Mean	184558	50.5	6.2	3.1	1.8	1572	17161	321.1
CES	288	S.D.	56983	38.2	1.2	0.64	0.75	468	27264	8.22
		Median	174000	47	6	3	2	1535	7840	320
		Mean	208790	39.7	6.2	3.1	1.9	1699	61132	327.0
EES	258	S.D.	81984	42.7	1.2	0.71	0.78	757	57441	14.06
		Median	192250	25	6	3	2	1531	36155	333
		Mean	271594	18.43	7.5	3.5	2.6	2190	19939	328.4
GES	869	S.D.	74545	25.71	1.5	0.64	0.78	725	19967	10.97
		Median	278000	8	8	4	3	2172	15245	327
		Mean	275205	15.88	7.3	3.4	2.71	2174	30585	333.5
NHUF	946	S.D.	96317	30.1	1.5	0.61	0.67	758	40188	6.90
		Median	270000	7	7	3	3	2036	14000	336
		Mean	206997	48.4	6.1	3.0	1.8	1608	43525	299.4
PFES	213	S.D.	135938	37.0	1.7	0.91	0.85	748	41790	30.23
		Median	179900	45	6	3	2	1447	27878	318
		Mean	225524	36.2	6.6	3.2	2.1	1820	41450	306.78
WES	474	S.D.	833365	40.6	1.5	0.75	0.88	776	46073	24.24
		Median	210000	22	6	3	2	1672	22651	323

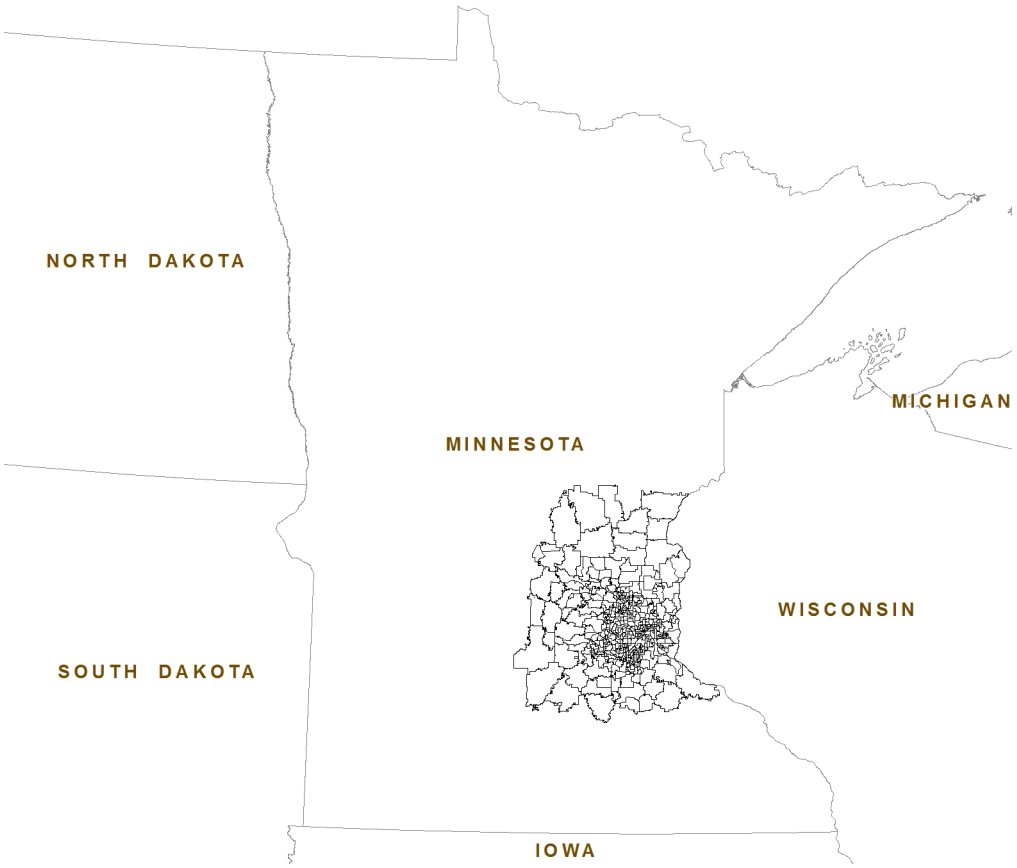

Figure 3: Map of MN Attendance Zones. This map shows the relative location of the study area within the state of Minnesota.

3 gives a reference to the size of the sample area in relation to the rest of the state. While only looking at relatively short timespan, this dataset includes a total of 101,993 observations. As it would not be feasible to plot each sale as was done in the previous section, figure 4 shows the number of home sales in each of the school districts and figure 5 shows home sales by attendance zone. Figure 5 is quite interesting as it gives a good picture of growth areas within the metro area which can be seen in the darker areas around the perimeter of the study area and in the center of the city.

4.2.1 Quality Variable

Unlike the Pennsylvania dataset above, this Minnesota dataset covers multiple school districts and so one must now worry not only about the school quality of the elementary schools but also middle and high schools as well. This introduces two new quality variables

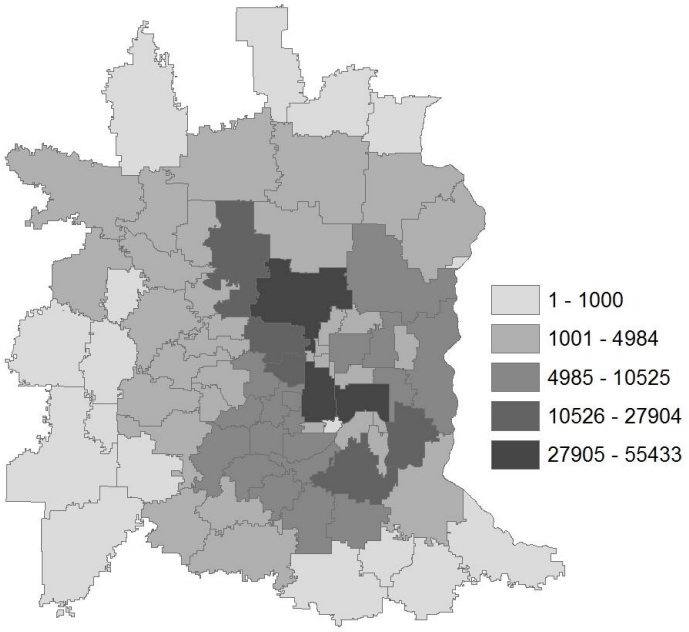

Figure 4: MN Home Sales by School District. The shade of gray indicates the number of home sales that occurred in each school district from June 2009 - January 2013.

into the mix and so S is no longer Nx1, but Nx3. One of the biggest benefits of this dataset however is that not only does it include the No Child Left Behind category percentages (Below basic, Basic, Proficient, and Advanced)[4] but it also includes average test score data at the school level. This allows us to not rely on the generated quality measures as in the previous section but to use actual test data as a quality measure.

Score Reporting

In Minnesota test score data is reported in an unusual format. Standardized test scores for each school are reported by grade on a 0-99 scale with the grade number prefixed to the score. So for example second grade test scores range from 200-299 while third grade scores range from 300-399. Also, within each of these ranges, scores have been scaled so that a score of 50 or above is "passing" by state requirements. Therefore the method to obtain an overall school score is a bit more complex than just averaging across the grade

[4]Although, in MN these categories are labeled "Does not meet standards", "Partially meets standards", "Meets standards", and "Exceeds standards".

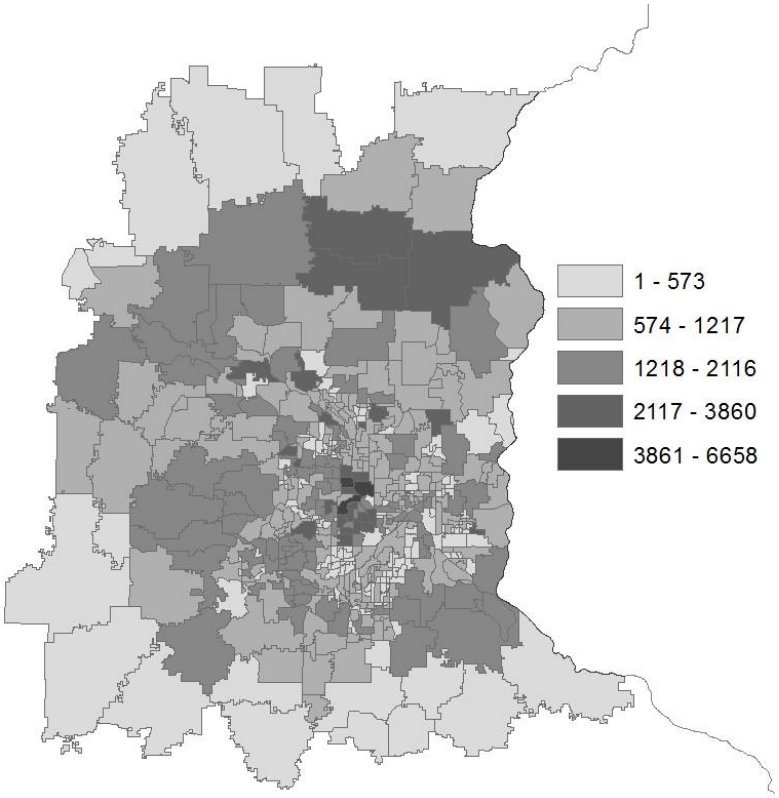

Figure 5: MN Home Sales by Attendance Zone. The shade of gray indicates the number of home sales that occurred in each attendance zone from June 2009 - January 2013.

scores. After discussing this project with test data specialists at the Minnesota Department of Education, the following process was suggested: first standardize each grade's score against the statewide average score for that grade then secondly average those standardized scores across all grades within each school. This puts each grade on equal footing (being compared to the state average) before being averaged together.

Categories Vs. Test Scores

Another major benefit of having both test score data along with the No Child Left Behind category data is that we can examine if the data-generated quality measure used in the Pennsylvania dataset adequately reflects the school's quality. One way to do this is to generate the same measure of quality used in the Pennsylvania dataset in the Minnesota dataset. Correlation coefficients can then be calculated between the test score measure and the category generated measure. Results from doing just this are found in table 5. The fact that the test score data is very highly correlated with the generated measure should give comfort when actual score data is unavailable as in the Pennsylvania dataset case.

Table 5: Test Score Measure and Category Measure Correlations

School Level	Correlation
Elementary	0.9842
Middle	0.9925
High	0.9929

4.2.2 Summary Statistics

Variable definitions for the dataset can be found in table 6 while table 7 gives the basic summary statistics for the 101,993 observations in the Minneapolis - St. Paul dataset. Standard data cleaning procedures were followed to remove incorrectly entered date (where rooms = 0 for example). Observations were also dropped when homes were not assigned to a middle school. This occurs in some districts where the high school includes the middle school, but resulted in missing values for the middle school quality variable.

Table 6: MN Data Set Variable Definitions

Variable	Definition
Sales Price	The sales price of the home in dollars
Age	The age of the home in years at time of sale
Rooms	The number of rooms in the home including living rooms, kitchen, etc.
Beds	The number of bedrooms in the home
Baths	The number of bathrooms in the home
GLA	The gross living area in square feet
Lot Size	The size of the housing lot in square feet
Elementary Quality	A variable constructed from test scores on state reading and math standardized exams at the home's zoned elementary school
Middle Quality	A variable constructed from test scores on state reading and math standardized exams at the home's zoned middle school
High Quality	A variable constructed from test scores on state reading and math standardized exams at the home's zoned high school

Table 7: MN Dataset Summary Statistics

Variable	Mean	Std. Dev.	Min	Max	Median	25th	75th
Sales Price	217300.38	189711.09	15003	4859000	173900	119000	258500
Age	38.028	31.102	0	285	30	13	55
Rooms	6.78	2.157	2	141	6	5	8
Beds	2.991	0.965	1	18	3	2	4
Baths	1.985	0.998	1	12	2	1	3
GLA	1699.213	1955.999	164	401771	1456	1123	2018
Lot Size	108342.736	5725437.03	435	6.64e8	10454	5663	15246
Elementary Quality	0.116	0.945	-3.159	2.058	0.31	-0.34	0.75
Middle Quality	0.246	0.882	-2.817	2.363	0.41	-0.31	0.84
High Quality	0.504	0.753	-1.547	1.814	0.63	0.17	1.0

Chapter 5

RESULTS

The following chapter is divided into two main sections. The first section examines results from non-spatial models in an effort to place the later spatial results into context within existing literature. The second section examines the SARAR mode and the different specifications of $E(WP|X)$.

Before diving into results, it is important to quickly discuss what we expect to see below. As far as expected signs, both the previous literature and basic economic intuition suggests that a home's age would have a negative effect However, we would also expect that effect to be increasing (moving in the positive direction) as the age of the home increases and eventually turning positive. That is when homes are relatively young, slightly older homes will sell for less than newer homes. However, there reaches an age when buyers are actively seeking older homes and so they able to demand a premium. We would also expect the number of bedrooms, bathrooms, and total rooms to have a positive effect but one that diminishes as the number increases. Buyers are likely willing to pay extra for a third bedroom while a tenth does not likely add much value to the home, especially while holding the square footage of the home constant. Gross living area and lot size are both expected to have a positive effect. As discussed before, school quality is expected to have a positive effect on a housing prices. Finally, the spatial lag coefficient (λ) is expected to be positive as being surrounded by higher priced homes should raise the price a given home. The spatial error autoregressive coefficient (ρ) is a bit more nuanced. There are some shocks to neighboring homes that could have competing effects on my home like new landscaping, for example. It beautifies the neighborhood and gives curb appeal to neighboring houses (mine included).

However, it also increases the quality of the newly landscaped home in relationship to mine and so it could have a negative effect on my home's value if they were being compared for sale. In this case it is not clear whether ρ would have a positive or negative sign. There are other cases where improvements to my neighbor's home only affects its value (i.e. new granite counter-tops) and so ρ would be negative in this case. Therefore, I hesitate about placing any prior expectations on the sign of ρ.

5.1 Non-spatial Models

Perhaps the most obvious choice for a baseline case is a naïve OLS regression that ignores more complicated endogeniety issues and concerns of spatial autocorrelation. The OLS estimates are biased and inconsistent. As mentioned before, there is a myriad of concerns with the model. First it ignores the omitted neighborhood amenities that are likely correlated with school quality. That is, schools with higher test scores are likely located in areas with better neighborhood public goods and services like parks, community programs, a safer environment, etc, all of which also affect housing prices. Therefore by ignoring these unobserved variables we are likely overestimating the effect school quality will have on housing prices. However the model does give information on the basic relationships between price and the exogenous RHS variables, and so we begin there.

5.1.1 OLS

PA Data

The results of an OLS regression of $\ln(SalesPrice)$ on housing characteristics and elementary school quality are found in table 8. All signs are as they were expected to be above (ignoring the spatial components which are not a part of this model). As this is a log-linear model, partial effects are a bit more cumbersome to calculate than in a linear model. However, using age as an example:

$$\frac{\partial E(P|X)}{\partial Age} = \exp\left(X\hat{\beta} + S\hat{\theta}\right)\left(\hat{\beta}_{Age} + 2\hat{\beta}_{Age2}Age\right). \tag{5.1}$$

Iit is good to note that because the first term in the above equation is always positive, the critical point (when the partial effect is equal to zero) is just as easily calculated as in the linear case by the second term ($\hat{\beta}_{Age} + 2\hat{\beta}_{Age2}Age$) above. So given the estimates in table 8, the partial effect is zero when a home is about 175 years old. So we would expect a premium to be paid on homes over 175 years old, of which there are only 18 observations in our sample. Bedrooms is perhaps a more interesting example. Given the estimates, adding a fifth bedroom would actually decrease the value of a home (again, holding other characteristics like GLA constant). The total number of rooms is found to be insignificant which makes sense as we are already controlling for the gross living area and number of bedrooms and bathrooms, so changing the number of rooms is just rearranging the floorplan without adding any value to the home. The school quality variable is actually insignificant with this specification, but as we believe there to be spatial effects present, this is a most likely a misspecification anyway and estimates will likely change below with the inclusion of the spatial lag terms.

MN Data

As in the Pennsylvania dataset, the estimates from an OLS estimation of prices on housing characteristics and school quality in table 9 all have the expected signs. There is a slight difference in these results from those above because the study area is larger in the Minnesota dataset and spans over several school districts, so quality variables for elementary, middle, and high schools must be included as opposed to only including elementary in the previous model. Only two variables are statistically insignificant: lot size and elementary school quality. In table 8, lot size has the highest t-statistic in the model (excluding the constant) and so finding it insignificant here is a bit perplexing. One hypothesis is that buyers in a metro area like MSP do not put much value into the lot size as it is a very urban population whereas in a suburb in Pennsylvania, lot size likely matters a great deal. The fact that elementary school quality is insignificant is a bit more puzzling although consistent with the Pennsylvania results above. The fact that middle and high school quality both have

Table 8: OLS Regression of ln(*SalesPrice*) on Housing Characteristics and Exogenous School Quality Variables in PA Dataset

Variable	Coefficient	t-Statistic
Age	-0.00389***	(-11.69)
Age2	0.0000111***	(5.63)
Beds	0.104**	(3.06)
Beds2	-0.0133**	(-2.79)
Baths	0.0567**	(3.01)
Baths2	-0.00329	(-0.91)
Rooms	0.0235	(1.16)
Rooms2	-0.000239	(-0.18)
GLA	0.000205***	(22.24)
Lot Size	0.00000260***	(23.04)
School Quality	0.00326	(0.54)
Constant	11.27***	(174.06)
N	3728	
R^2	0.660	

Includes 28 quarterly time fixed effects and 7 elementary school fixed effects.
t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

a positive and significant effect on housing prices is encouraging though. Perhaps there is a multicollinearity issue present here. If elementary, middle, and high school quality are all highly correlated then the analysis would be unable to separate the effects causing some of the variables to be insignificant. Aside from that issue, the coefficients on middle and high school quality seem plausible and consistent in magnitude with previous literature. However, as noted above this model ignores both the spatial nature of the housing market and any endogeniety issues with the school quality variable and so any estimates must be viewed with caution.

5.1.2 Boundary Fixed Effect Model

PA Data

In an effort to replicate the methodology of Black (1999), boundary fixed effects were created in an effort to control for unobserved neighborhood characteristics which likely change little over a boundary line but where school quality makes a distinct jump. Having

Table 9: OLS Regression of ln(*SalesPrice*) on Housing Characteristics and Exogenous School Quality Variables in MN Dataset

Variable	Coefficient	t-Statistic
Age	-0.00736***	(-36.86)
Age2	0.0000305***	(18.01)
Beds	0.0814***	(16.11)
Beds2	-0.0108***	(-16.40)
Baths	0.160***	(32.97)
Baths2	-0.00331***	(-4.02)
Rooms	0.0421***	(26.70)
Rooms2	-0.000399***	(-15.08)
GLA (in thousands)	0.0228***	(28.13)
Lot Size (in thousands)	2.88e-7	(1.11)
Elem. SQ	0.0107	(1.21)
Midd. SQ	0.0622***	(11.96)
High SQ	0.0310***	(4.69)
Constant	14.06	(0.00)
N	101993	
R^2	0.526	

Includes 16 quarterly time fixed effects and 347 elementary school fixed effects.
t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

a much smaller dataset than in Black (1999), the method was quite simple. First, as the need for stationary boundaries is evident, only data before the October 2008 redistricting of the Boyertown Area School District (BASD) was available. Observations within 0.15 miles of each zone boundary are pulled into a new dataset and assigned to a dummy variable for their closest boundary. This reduces the number of observations from 3728 to 529, an almost 86 percent decrease. These fixed effects are then included in the OLS regressions above. Like Black (1999), robust standard errors clustered at the attendance zone are used. Results can be found in table 10 below. One first notices that almost none of the variables are significant in this specification. This is likely due to the decrease in sample size and likely low variability in housing characteristics within such a narrow band around zone borders. Secondly, elementary school quality is found to have a negative and significant (at the 10 percent level) relationship. Again, this model does not account for any spatial effects and so

this could be due to a misspecification problem. However, Ries and Somerville (2010) also found a negative and significant relationship between elementary school quality and housing prices in the Vancouver school district. Likewise, they were unable to offer any possible justification.

Table 10: OLS Regression of ln(*SalesPrice*) on Housing Characteristics, Boundary Fixed Effects, and Exogenous School Quality Variables in PA Dataset from Jan. 2004 - Oct 2008

Variable	Coefficient	t-Statistic
Age	-0.00209	(-0.86)
Age2	0.00000196	(0.11)
Rooms	0.0158	(0.17)
Rooms2	-0.00000682	(-0.00)
Beds	0.00936	(0.06)
Beds2	0.000629	(0.04)
Baths	0.0829	(1.52)
Baths2	-0.00511	(-0.54)
GLA (in thousands)	0.175***	(3.77)
Lot Size (in thousands)	0.00359**	(3.44)
School Quality	-0.00176*	(-2.03)
Constant	12.33***	(23.42)
N	529	
R^2	0.631	

Includes 18 quarterly time fixed effects and 11 boundary fixed effects.
t statistics in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

MN Data

In replicating the Black (1999) methodology for the Minnesota dataset, I quickly ran into an issue that she only briefly mentions in her data appendix: some boundary lines serve as a boundary to more than one school type. For example a given boundary may serve as the edge of elementary school A's attendance zone while also serving as the boundary line for middle school B's attendance zone. This also happens at the high school level and the school district level. Black mentions within the paper that she removes boundaries that also serve as school district borders, but fails to mention the situation where a boundary serves multiple school zones. In the data appendix she mentions that in rare situations an

elementary school might share a border with a middle school and in that situation the quality of both schools are included (Black, 1999, pg. 596). However, as seen below, this is not a rare occurrence in this dataset. Table 11 presents estimates when no borders are removed due to overlapping. All significant variables have their respective expected signs but none of the quality variables are significant. Again, this makes sense if there are a large number of boundaries that are shared and so the disjoint jump in quality that Black (1999) relies on for identification is happening in multiple variables at once.

Table 11: OLS Regression of ln(*SalesPrice*) on Housing Characteristics, School Boundary Fixed Effects, and Exogenous School Quality Variables in MN Dataset

Variable	Coefficient	t-Statistic
Age	-0.00432***	(-7.48)
Age2	0.0000139***	(2.76)
Beds	0.0800***	(5.21)
Beds2	-0.0105***	(-5.18)
Baths	0.102***	(8.12)
Baths2	-0.0110***	(-5.83)
Rooms	-0.00883	(1.34)
Rooms2	-0.000315	(-1.12)
GLA (in thousands)	0.271***	(16.00)
Lot Size (in thousands)	3.30e-7*	(2.11)
Elem. SQ	0.0127	(1.46)
Midd. SQ	-0.00201	(-0.19)
High SQ	0.0201	(1.26)
Constant	10.58***	(137.9)
N	29058	
R^2	0.638	

Includes 16 quarterly time fixed effects and 1014 boundary fixed effects.
Robust standard errors adjusted for attendance zone clusters.
t statistics in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

In an effort to combat this situation, I wanted to solely examine the effect a change in elementary school quality would have on prices. In order to correct the situation of shared boundaries, any elementary school boundary that is shared with a middle school, high school, or school district was removed. This reduces the number of boundaries used from 1,014 to

338. So roughly two-thirds of the elementary school boundaries were shared with other school types or other districts. Dropping these additional boundaries resulted in a lowering of the number of observations from 29,058 in table 11 to 14,692 in table 12. Now, not only do we see very similar estimates of the housing characteristics when compared to table 11, but also significant elementary school quality effects. By these estimates an increase in test scores of one standard deviation (compared to the state average) raises home prices by 2.25 percent, or $4,889 for the average home in the dataset.

Table 12: OLS Regression of ln($SalesPrice$) on Housing Characteristics, Elementary School Boundary Fixed Effects, and Exogenous School Quality Variables in MN Dataset

Variable	Coefficient	t-Statistic
Age	-0.00458***	(-5.44)
Age2	0.0000110	(1.55)
Beds	0.0748***	(3.04)
Beds2	-0.00942***	(-3.05)
Baths	0.0952***	(4.79)
Baths2	-0.0110***	(-2.71)
Rooms	0.0243*	(1.89)
Rooms2	-0.00106	(-1.64)
GLA (in thousands)	0.273***	(8.80)
Lot Size (in thousands)	2.91e-7*	(1.71)
Elem. SQ	0.0225*	(1.88)
Constant	11.28***	(131.92)
N	14692	
R^2	0.607	

Includes 16 quarterly time fixed effects and 338 boundary fixed effects.
Robust standard errors adjusted for elementary attendance zone clusters.
t statistics in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Similar data techniques were used for estimates of the effect of middle school and high school quality in tables 13 and 14, respectively. However, because elementary schools are the most spatially disaggregated school level, far fewer boundaries exist when elementary boundaries are removed as they constitute the majority of all boundaries. In fact, for the middle school estimation only 25 boundaries are not shared by either elementary or high

schools while in the high school estimation only 15 boundaries were not shared. The lack of significance of most housing characteristics in tables 13 and 14 are likely due to severely diminished sample sizes of 1,254 and 1,216. Also as the vast majority of middle and high school boundaries consist of boundaries of elementary school zones, it begs the question as to why these few are not. Is there something special going on in these areas that makes them different than other places in the study area? If so, results from this limited estimation obviously can't be extrapolated to explain the entire dataset. Any one of these concerns are likely behind the estimate of a negative effect of high school quality on home prices found in table 14.

Table 13: OLS Regression of ln(*SalesPrice*) on Housing Characteristics, Middle School Boundary Fixed Effects, and Exogenous School Quality Variables in MN Dataset

Variable	Coefficient	t-Statistic
Age	-0.00760***	(-3.66)
Age2	0.0000178	(0.60)
Beds	0.120	(1.29)
Beds2	-0.0290**	(-2.26)
Baths	0.0782	(1.49)
Baths2	-0.0166*	(-1.98)
Rooms	0.0258**	(2.22)
Rooms2	-0.000163**	(-2.09)
GLA (in thousands)	0.304***	(10.59)
Lot Size (in thousands)	0.0100***	(7.03)
Middle SQ	-0.0419	(-1.04)
Constant	11.10***	(79.77)
N	1254	
R^2	0.634	

Includes 16 quarterly time fixed effects and 25 boundary fixed effects.
Robust standard errors adjusted for middle attendance zone clusters.
t statistics in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

5.2 Spatial Models

While the above specifications give some context in which to evaluate any spatial models that are encountered in this section, the first model spatial presented in this section

Table 14: OLS Regression of ln($SalesPrice$) on Housing Characteristics, High School Boundary Fixed Effects, and Exogenous School Quality Variables in MN Dataset

Variable	Coefficient	t-Statistic
Age	-0.00434**	(-2.41)
Age2	0.00000247	(0.12)
Beds	0.154***	(2.89)
Beds2	-0.0192***	(-3.58)
Baths	0.0627	(1.27)
Baths2	-0.00860	(-1.18)
Rooms	0.0462	(1.19)
Rooms2	-0.00277	(-1.36)
GLA (in thousands)	0.289***	(7.64)
Lot Size (in thousands)	-2.62e-06	(-0.71)
High SQ	-0.0623*	(-1.72)
Constant	10.86***	(77.07)
N	1216	
R^2	0.526	

Includes 16 quarterly time fixed effects and 15 boundary fixed effects.
Robust standard errors adjusted for high school attendance zone clusters.
t statistics in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

should be viewed as the baseline to compare all other spatial models against. Spatial models are commonly used in the housing literature and so the standard SARAR, a generalization of both the SAR and SEM models, with all exogenous variables other than the spatially lagged dependent variable provides a starting point for future discussions on how estimation procedures effect results. Spatial models require the use of a weighting matrix to create the spatially lagged variables. In all the models presented in this dissertation, W is constructed as a row-normalized inverse squared distance of the k-nearest neighbors. Specifically, the k-nearest neighbors are found for each home and the distance between these k neighbors and the home are computed. These distances are then invented and squared to create the weights between the home and its k neighbors. Because row-normalized weighting matrices have desirable properties for spatial analysis[1], W is then normalized so that each row sums to one. Since the choice of k is arbitrary, I rely on previous literature to guide the choice.

[1]See Anselin (1988)

Most papers that use the k-nearest neighbor approach tend to use values of k that range between five and twenty. I therefore estimate each specification below using values of k that run from five to thirty in steps of five. Results reported in this section will all have $k = 20$ as it appears to be the most widely used in the literature, but the other five sets of estimates ($k = 5, 10, 15, 25, 30$) can be found in Appendix A.

5.2.1 SARAR with Exogenous Quality

PA Data

Table 15 presents estimates of a SARAR model with housing prices as the dependent variable and elementary school quality being treated as exogenous. This model is analogous to the OLS model in table 8 and should be thought of as a baseline against which to compare the other spatial models in this section. In the estimates presented in table 15, the effects of housing characteristics are similar (both in magnitude and sign) to table 8. However, we now are presented with strong evidence for spatial autocorrelation in both the main regression equation and the error term. Both λ and ρ are positive and significant with point-estimates of 0.243 and 0.0576 respectively. The estimates of the coefficients on housing characteristics are of the correct sign and their significance is similar to those in table 8. The effect of elementary school quality is found to be negative which is again, a strange result and one hard to defend.

MN Data

The Minnesota data in table 16, however, seems to be much more "well-behaved" in that its estimates are all of the expected sign, mostly significant [2], and with strong spatial effects. The negative nature of ρ gives insight that in the Minnesota dataset, a positive shock to a neighboring house tends to be viewed a negative shock by a given home. This would be true in a market where homes can be seen more as competitors with their neighbors than anything else. In spatial models, these point estimates cannot be thought of as marginal

[2]Lot size is insignificant here, but a similar argument can be made to that from section 5.1.1

Table 15: 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Exogenous School Quality Variables in PA Dataset when $k = 20$

Variable	Coefficient	t-Statistic
Constant	7.81***	(30.72)
Age	-0.00282***	(-8.51)
Age2	0.000008***	(4.05)
Beds	0.1129***	(3.50)
Beds2	-0.0151***	(-3.32)
Baths	0.0708***	(3.93)
Baths2	-0.00507	(-1.47)
Rooms	0.0185	(0.96)
Rooms2	-0.000408	(-0.32)
GLA (in thousands)	0.1744***	(19.04)
Lot Size (in thousands)	0.00222 ***	(19.41)
Elem. SQ	-0.009244	(-0.416)
Spatial Lag (λ)	0.243***	(12.03)
Spatial Error Lag (ρ)	0.0576***	(2.35)
N	3728	
R^2	0.6937	

Includes 28 quarterly time fixed effects and 7 elementary zone fixed effects.
t statistics in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

effects, but as a first-degree approximation to the true marginal effects as a rise in given homes price will raise other prices (assuming $\lambda > 0$) which in turn "re-raise" this home's price again. This multiplier process can play itself out over time, but the first major change is given by the point estimates in table 16.

5.2.2 SARAR with Endogenous Quality and Parametric Estimates of $E(S|X)$

The following three sections compare and contrast the results of the three different estimation procedures presented in chapter 3. In all three methods $E(S|X)$ is being estimated parametrically through the method explained in chapter 3 and only differentiate from each other in one key aspect. As the methods are essentially different approaches to estimating $E(WP|X)$, discussion will focus on how each method effects estimates of λ, ρ, and θ while primarily skipping any discussion of the β estimates.

Table 16: 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Exogenous School Quality Variables in MN Dataset when $k = 20$

Variable	Coefficient	t-Statistic
Age	-0.00322***	(-21.5)
Age2	0.000009***	(7.02)
Beds	0.0564***	(14.06)
Beds2	-0.00683***	(-12.92)
Baths	0.0818***	(20.76)
Baths2	-0.000310	(-0.47)
Rooms	0.0269***	(20.85)
Rooms2	-0.000249***	(-11.34)
GLA (in thousands)	0.0149***	(21.85)
Lot Size (in thousands)	-1.26e-06	(-0.14)
Elem. SQ	0.0541***	(6.40)
Midd. SQ	0.0659***	(7.74)
High SQ	0.0225**	(2.51)
Constant	4.45***	(46.42)
Spatial Lag (λ)	0.708***	(139.65)
Spatial Error Lag (ρ)	-0.382***	(-70.64)
N	101993	
R^2	0.667	

Includes 16 quarterly time fixed effects and 347 elementary zone fixed effects.
t statistics in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Method A

Recall that the method A estimator is estimating $E(WP|X)$ non-parametrically as suggested by Kelejian and Prucha (1998). This is essentially the Kelejian and Prucha estimator that has been altered to treat school quality as endogenous. The benefit of method Al is that it the estimate of $E(WP|X)$ does not depend on the estimation of $E(S|X)$ and so any misspecification of that model should have no effect on its estimation. However, because this method does not impose the restriction on WP imposed by the model, asymptotic efficiency is lost.

PA Data Before discussing λ, ρ, or θ, it is obvious that something has drastically changed between tables 15 and 17. However, the only estimation difference is that school quality

Table 17: 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in PA Dataset when $k = 20$ with $E(WP|X)$ as Method A

Variable	Coefficient	t-Statistic
Constant	17.57**	(2.25)
Age	-0.00218**	(-2.32)
Age2	0.000001	(0.16)
Beds	-0.00612	(-0.55)
Beds2	0.00314	(0.19)
Baths	0.0686*	(1.69)
Baths2	-0.00481	(-0.61)
Rooms	0.0639	(1.25)
Rooms2	-0.00312	(-1.00)
GLA (in thousands)	0.1543***	(5.38)
Lot Size (in thousands)	0.00195 ***	(6.07)
Elem. SQ	-0.0342	(-1.31)
Spatial Lag (λ)	0.274***	(5.45)
Spatial Error Lag (ρ)	0.0454	(0.95)
N	3728	
R^2	-0.235	

Includes 28 quarterly time fixed effects and 7 elementary zone fixed effects.
t statistics in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

is no longer being treated as exogenous and is therefore being instrumented as described in previous chapters. It seems this has a disastrous effect on the significance levels of the housing characteristics in the models. This most likely stems from the fact that there are only 49 observations on school quality in the PA dataset and so any estimation of $E(S|X)$ will likely be of poor quality due to the small sample size, even if the model is correctly specified. Here λ is very similar to previous estimates but its t-statistic has lowered from the previous table and ρ has also become insignificant. These results are robust across the specifications of k.

MN Data The major effect of treating school quality as endogenous in this specification, shown in table 18, is that the point estimate on elementary school quality has jumped from 0.0541 to an amazingly high estimate of 0.786. Even as a first approximation, an increase of

Table 18: 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in MN Dataset when $k = 20$ with $E(WP|X)$ as Method A

Variable	Coefficient	t-Statistic
Constant	2.88***	(21.27)
Age	-0.00605***	(-28.50)
Age2	0.000027***	(14.71)
Beds	0.0654***	(8.27)
Beds2	-0.00700***	(-6.00)
Baths	0.1054***	(21.42)
Baths2	-0.00312***	(-3.24)
Rooms	0.00290***	(11.53)
Rooms2	-0.000269***	(-3.57)
GLA (in thousands)	0.0139*	(1.68)
Lot Size (in thousands)	-1.57e-06	(-0.10)
Elem. SQ	0.786***	(3.84)
Midd. SQ	-0.0258	(-1.06)
High SQ	0.0516***	(3.11)
Spatial Lag (λ)	0.609***	(62.92)
Spatial Error Lag (ρ)	-0.423***	(-25.36)
N	101993	
R^2	0.556	

Includes 16 quarterly time fixed effects and 347 elementary zone fixed effects.
t statistics in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

78 percent of a home's price due to a change in school quality seems unrealistic. In the other versions of this model (seen in the appendix in tables ?? to ??) where the effect of elementary school quality is significant, the estimate didn't jump as high, but still goes to the 0.3 or 0.4 range. Surprisingly the change did not propagate through to the other school quality measures. Middle school quality is now negative and insignificant. This could be due to high correlation between middle and high school scores. The effect of middle school quality also change by the value of k. In some instances ($k = 5, 30$), it is positive and significant. While in other specifications it is negative and significant ($k = 15, 20$). The inclusion of endogenous school quality seems to have had little effect on λ which is encouraging as $E(WP|X)$ is being estimated separately and non-parametrically and it changes little under different values of

Table 19: 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in PA Dataset when $k = 20$ with $E(WP|X)$ as Method B

Variable	Coefficient	t-Statistic
Constant	19.18**	(2.66)
Age	-0.00269**	(-2.34)
Age2	0.000003	(0.36)
Beds	-0.0137	(-0.12)
Beds2	0.00440	(0.27)
Baths	0.0674	(1.51)
Baths2	-0.00424	(-0.52)
Rooms	0.0674	(1.34)
Rooms2	-0.00319	(-1.00)
GLA (in thousands)	0.1686***	(5.06)
Lot Size (in thousands)	0.00212 ***	(5.43)
Elem. SQ	-0.0350	(-1.36)
Spatial Lag (λ)	0.153	(1.24)
Spatial Error Lag (ρ)	0.0775	(1.00)
N	3728	
R^2	-0.287	

Includes 28 quarterly time fixed effects and 7 elementary zone fixed effects.
t statistics in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

k.

Method B

In this section, the previously non-parametrically estimated $E(WP|X)$ is now given a functional form similar to the one expressed in Lee (2003). Method B is then just in fact the Lee (2003) estimator that has been modified to allow for endogenous, parametrically-instrumented school quality. The benefit of this methodology is through the gains in asymptotic efficiency made by parametrically estimating using the closed-form of E(WP|X) instead of non-parametrically through cross-products of the spatially lagged exogenous variables. However, now the $E(WP|X)$ is dependent on the specification of $E(S|X)$ and so if the estimate of $E(S|X)$ has issues, so will $E(WP|X)$.

Table 20: 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in MN Dataset when $k = 20$ with $E(WP|X)$ as Method B

Variable	Coefficient	t-Statistic
Constant	0.697	(0.894)
Age	-0.00530***	(-16.97)
Age2	0.000023***	(12.31)
Beds	0.0613***	(8.32)
Beds2	-0.00622***	(-5.54)
Baths	0.0927***	(11.13)
Baths2	-0.000314***	(-3.43)
Rooms	0.0269***	(9.73)
Rooms2	-0.000246***	(-3.43)
GLA (in thousands)	0.0120*	(1.67)
Lot Size (in thousands)	-2.19e-06	(-0.25)
Elem. SQ	0.529***	(3.48)
Midd. SQ	-0.0458**	(-2.05)
High SQ	0.0161	(0.86)
Spatial Lag (λ)	0.869***	(13.77)
Spatial Error Lag (ρ)	-0.663***	(-21.61)
N	101993	
R^2	0.538	

Includes 16 quarterly time fixed effects and 347 elementary zone fixed effects.
t statistics in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

PA Data Now that $E(WP|X)$ is dependent on the estimation of $E(S|X)$ which is based on 49 observations, both λ and ρ have become insignificant in this specification shown in table 19. Evidence suggests that the parametric model is either misspecified or small sample size issues are causing $E(S|X)$ to be poorly estimated, which is in turn causing the the rest of the variables to be insignificant. A change that was made in the name of asymptotic efficiency is ending up causing a lot of problems in the finite sample realm. We see the same results in the other specifications of k found in the appendix as well.

MN Data The Minnesota dataset continues to perform well under the different specifications. In table 20, the data shows no sign of the issues plaguing the Pennsylvania dataset

Table 21: 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in PA Dataset when $k = 20$ with $E(WP|X)$ as Method C

Variable	Coefficient	t-Statistic
Constant	17.61***	(2.25)
Age	-0.00219***	(-2.34)
Age2	0.000001	(0.17)
Beds	0.00632	(-0.05)
Beds2	-0.00697	(0.19)
Baths	0.00317*	(1.68)
Baths2	-0.00479	(-0.61)
Rooms	0.0640	(1.25)
Rooms2	-0.00316	(-1.00)
GLA (in thousands)	0.1547***	(5.42)
Lot Size (in thousands)	0.00196 ***	(6.11)
Elem. SQ	0.0.0343	(-1.31)
Spatial Lag (λ)	0.270***	(5.48)
Spatial Error Lag (ρ)	0.0461	(0.96)
N	3728	
R^2	-0.236	

Includes 16 quarterly time fixed effects and 7 elementary zone fixed effects.
t statistics in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

which likely stems from the fact that the parametric estimation of school quality in the Minnesota data set is using many more observations and so can more accurately estimate $E(S|X)$. The Minnesota dataset is not without its problems however. A point estimate of 0.529 is still an order of magnitude larger than any other school quality effect reported in the literature. Also, when $k = 15$ elementary school quality has a negative and significant estimate of -0.22. This is obviously a troubling result. The estimate of λ is fairly consistent through the different specifications of k ranging from 0.66 to 0.87 when significant.

Method C

In a new method not previously used in the literature, $E(WP|X)$ is estimated using both parametric and non-parametric methods. $E(WP|X)$ is separated into two parts: one that is influenced by the estimation of $E(S|X)$ and one that is not. The part that is not

Table 22: 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in MN Dataset when $k = 20$ with $E(WP|X)$ as Method C

Variable	Coefficient	t-Statistic
Constant	2.79***	(17.00)
Age	-0.00602***	(-28.41)
Age2	0.000027***	(14.71)
Beds	0.0653***	(8.28)
Beds2	-0.00697***	(-6.00)
Baths	0.105***	(21.15)
Baths2	-0.00313***	(-3.25)
Rooms	0.0290***	(11.53)
Rooms2	-0.000268***	(-3.57)
GLA (in thousands)	0.0138*	(1.68)
Lot Size (in thousands)	-1.69e-06	(-0.11)
Elem. SQ	0.7763***	(3.82)
Midd. SQ	-0.0266	(-1.10)
High SQ	0.0502***	(3.04)
Spatial Lag (λ)	0.619***	(50.92)
Spatial Error Lag (ρ)	-0.438***	(-25.99)
N	101993	
R^2	0.556	

Includes 16 quarterly time fixed effects and 347 elementary zone fixed effects.
t statistics in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

affected by $E(S|X)$ is estimated parametrically in order to capture as much efficiency as possible. However, the part that is influenced by $E(S|X)$ is regressed on the KP instrument matrix of cross-products of spatially lagged exogenous variables[3] which prevents $E(WP|X)$ from being influenced by misspecification in $E(S|X)$. Surprisingly, both the Pennsylvania data in table 21 and the Minnesota data in table 22 are extremely close, if not almost exact replicas of tables 17 and 18

In the MN data, the estimate on elementary school quality fluctuates heavily based on k. It ranges from 0.34 when $k = 30$ to 0.78 when $k = 20$. Lambda and rho are fairly stable across the specifications of k. In the PA dataset, elementary school quality is insignificant

[3]For details, see the end of chapter 3

in every specification of k whereas lambda is positive and significant in all specifications.

Chapter 6

CONCLUSION

This dissertation develops and estimates a spatial autoregressive with autoregressive errors model of housing prices that accounts for both the endogeneity of spatially-lagged housing prices and local school quality measured by performance on state standardized tests. By using two separate datasets, one in Minnesota and one in Pennsylvania, I am able to test the model in a dataset with a small number schools and home sales (PA) and one with a large number of both (MN). Homes are spatially weighted against each other using a k nearest-neighbor approach mixed with an inverse distance approach. That is, the k nearest neighbors are weighted against each other by the inverse of the distance between them. As the choice of k is then arbitrary, models are estimated by varying k from 5 to 30 in steps of 5.

School quality is thought to be endogenous because unobserved neighborhood amenities in the error term of a hedonic regression are very likely positively correlated with local elementary, middle, and high school quality. Following previous literature, the optimal instrument matrix is constructed as the conditional means of the spatially-lagged housing prices and quality measures which have had a Cochrane-Orcutt like transformation applied to them. However, as school quality is observed on a much lower frequency than housing prices, it is not possible to estimate the conditional mean of school quality using non-parametric methods as proposed previously in the literature. So in order to instrument the school quality variables, a parametric model in which school quality is a function of average home prices within its attendance zone and average home prices outside its attendance zone but still within the same school district is used.

Three methods are used to estimate the conditional mean of the spatially-lagged housing prices. First, the most well known and used method of estimating the conditional mean non-parametrically using the cross-products of spatially lagged exogenous variables is used. Second, it is shown that the SARAR model imposes a restriction on the conditional mean so that it has a closed form and can be estimated parametrically resulting in gains in asymptotic efficiency. However, this estimate depends on the estimation of the conditional mean of school quality and therefore is sensitive to the functional form chosen to represent school quality. Finally, we present a method previously unknown in the literature where the part of the conditional mean that depends on the conditional mean of school quality is estimated non-parametrically while the part that does not is estimated parametrically.

Using datasets from Boyertown Area School District in Pennsylvania and the Minneapolis - St. Paul metro area, each of these models were estimated and compared to results found previously in the literature. In addition, estimates were computed using the boundary fixed effect approach pioneered by Black (1999) and a SARAR model that treats school quality as exogenous in order for comparison. I found that the method used to estimate the conditional mean of the spatially-lagged housing prices along with the number of neighbors used in the weighting matrix can have a large effect on the size, sign, and significance of key estimates.

One of the biggest results was that there was very little difference in the results of the first and third method of estimating the conditional mean of the spatially-lagged housing prices. In fact, estimates using the first method tended to have slightly higher t-statistics on all variables even though the third method made some attempt to incorporate a model restriction to improve asymptotic efficiency. Estimates using the second method were different, yet not so dissimilar from the other methods to pose too much concern. However, the largest issue raised through the estimation results is the apparently heavy sensitivity of the results on the specification of W. As W is used as both the weighting matrix in the main regression equation as well as the weighting matrix in the error term, it permeates

every area of the model. Therefore it makes sense that it would have a large effect on the results. However, as there is no economic intuition available for the correct choice of k, the widely varied results are troublesome as there is not an easy way to distinguish the correct specification from an incorrect one.

Ideas for future research abound from this model. I believe one of the first areas that needs study is Monte Carlo experiments that look at a few different aspects of this study. First, a study of the difference between the 3 estimation methods stated above and their finite sample properties. This could be conducted assuming the researcher knows the correct specification of W and $E(S|X)$ so that the basic finite sample properties could be studied. Second, testing how different misspecifications in W affect estimators would be a worthwhile study and could have far-reaching implications for future spatial research. Thirdly, seeing how the three estimation methods are affected by misspecification of the additional endogenous variables and what finite sample ramifications those might have would a nice follow-up to the proposed research in the first point.

Other possible future research ideas include using this data to study the possible existence and identification of housing submarkets. A preliminary look at this research, including literature and initial results using a genetic algorithm to identify the submarkets can be found in Appendices C and D. Another possible research topic is using a identification method not discussed in this dissertation to estimate the effect of school quality on housing prices. It would be similar to combining the methods of Black (1999) and Ries and Somerville (2010). That is, by pairing homes very close to each other and treating them as (pseudo) repeat-sales, it could be possible to identify the effect of school quality on housing prices using rezoning similar to the one found in Ries and Somerville (2010) (and the Pennsylvania dataset here) while not being nearly as constrained by sample size. Preliminary results and discussion of this methodology can be found in Appendix B.

Overall, the estimation of the effect that school quality has on housing prices is a rich and interesting question. Unfortunately, the nature and the the interconnectivity of the economic systems and the large amount of unobserved characteristics of homes and schools

pose serious problems to any estimation procedure. However, the continued dissemination of school data and the increasing quality and availability of housing data provide ample opportunity and means for future research on the topic.

BIBLIOGRAPHY

BIBLIOGRAPHY

Akaike, H. (1974) "A New Look At The Statistical Model Identification." *IEEE Transactions on Automatic Control*, Vol. 19(6), pp. 716–723.

Amemiya, T. (1977) "The maximum likelihood and the nonlinear three-stage least squares estimator in the general nonlinear simultaneous equation model." *Econometrica: Journal of the Econometric Society*, Vol. pages 955–968.

Anselin, L. (1988) *Spatial econometrics: methods and models.*, volume 4 Kluwer Academic Pub.

Anselin, L. (2001) *Spatial econometrics.* Blackwell Publishing.

Anselin, L. (2010) "Thirty years of spatial econometrics." *Papers in Regional Science*, Vol. 89(1), pp. 3–25.

Bayer, P.; Ferreira, F.; and McMillan, R. (2007) "A Unified Framework for Measuring Preferences for Schools and Neighborhoods." *Journal of Political Economy*, Vol. 115(4), pp. 588–638.

Black, S. E. (1999) "Do Better Schools Matter? Parental Valuation of Elementary Schools." *Quarterly Journal of Economics*, Vol. 114(2), pp. 577–599.

Bourassa, S. C.; Cantoni, E.; and Hoesli, M. (2007) "Spatial Dependence, Housing Submarkets, and House Price Prediction." *Journal of Real Estate Finance and Economics*, Vol. 35, pp. 143–160.

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

Bourassa, S. C.; Hoesli, M.; and Peng, V. S. (2003) "Do Housing Submarkets Really Matter?." *Journal of Housing Economics*, Vol. 12, pp. 12–28.

Brasington, D. M. and Haurin, D. R. (2006) "Educational Outcomes and House Values: A Test of the Value-Added Approach." *Journal of Regional Science*, Vol. 46(2), pp. 245–268.

Brasington, D. M. and Haurin, D. R. (2009) "Parents, Peers, or School Inputs: Which Components of School Outcomes are Capitalized into Home Value?." *Regional Science and Urban Economics*, Vol. 39(5), pp. 523–529.

Case, B.; Clapp, J.; Dubin, R.; and Rodriguez, M. (2004) "Modeling Spatial and Temporal House Price Patterns: A Comparison of Four Models." *Journal of Real Estate Finance and Economics*, Vol. 29(2), pp. 167–191.

Dorsey, R. E.; Hu, H.; Mayer, W. J.; and Wang, H. (2010) "Hedonic Versus Repeat-Sales Housing Price Indexex for Measuring the Recent Boom-Bust Cycle." *Journal of Housing Economics*, Vol. 19, pp. 75–93.

Dorsey, R. E. and Mayer, W. J. (1995) "Genetic Algorithms for Estimation Problems with Multiple Optima, Nondifferentiability, and Other Irregular Features." *Journal of Business and Economic Statistics*, Vol. 13(1), pp. 53–66.

Downes, T. A. and Zabel, J. E. (2002) "The Impact of School Characteristics on House Prices: Chicago 1987-1991." *Journal of Urban Economics*, Vol. 52(1), pp. 1–25.

Drukker, D. M.; Egger, P.; and Prucha, I. R. (2013) "On two-step estimation of a spatial autoregressive model with autoregressive disturbances and endogenous regressors." *Econometric Reviews*, Vol. 36(5–6).

Elhorst, J. P. (2010) "Applied spatial econometrics: raising the bar." *Spatial Economic Analysis*, Vol. 5(1), pp. 9–28.

Fingleton, B. and Le Gallo, J. (2008) "Estimating spatial models with endogenous variables, a spatial lag and spatially dependent disturbances: Finite sample properties*." *Papers in Regional Science*, Vol. 87(3), pp. 319–339.

Getis, A. (2009) "Spatial weights matrices." *Geographical Analysis*, Vol. 41(4), pp. 404–410.

Getis, A. and Aldstadt, J. (2004) "Constructing the Spatial Weights Matrix Using a Local Statistic." *Geographical Analysis*, Vol. 36(2), pp. 90–104.

Gibbons, S. and Machin, S. (2003) "Valuing English Primary Schools." *Journal of Urban Economics*, Vol. 53(2), pp. 197–219.

Gibbons, S. and Machin, S. (2006) "Paying for primary schools: admission constraints, school popularity or congestion?*." *The Economic Journal*, Vol. 116(510), pp. C77–C92.

Goodman, A. C. (1981) "Housing Submarkets Within Urban Areas." *Journal of Regional Science*, Vol. 21(2), pp. 175–185.

Goodman, A. C. and Dubin, R. A. (1990) "Sample Stratification with Non-Nested Alternatives: Theory and a Hedonic Example." *Review of Economics and Statistics*, Vol. 72, pp. 168–173.

Goodman, A. C. and Thibodeau, T. G. (1998) "Housing Market Segmentation." *Journal of Housing Economics*, Vol. 7, pp. 121–143.

Goodman, A. C. and Thibodeau, T. G. (2003) "Housing Market Segmentation and Hedonic Prediction Accuracy." *Journal of Housing Economics*, Vol. 12, pp. 181–201.

Goodman, A. C. and Thibodeau, T. G. (2007) "The Spatial Proximity of Metropolitan Area Housing Submarkets." *Real Estate Economics*, Vol. 35(2), pp. 209–232.

Hill, R. J. (2012) "HEDONIC PRICE INDEXES FOR RESIDENTIAL HOUSING: A SURVEY, EVALUATION AND TAXONOMY*." *Journal of Economic Surveys*, Vol. .

Kapetanios, G. (2006) "Cluster Analysis of Panel Data Sets using Non-Standard Optimization of Information Criteria." *Journal of Economic Dynamics and Control*, Vol. 30(8), pp. 1389–1408.

Kauko, T. (2004) "A Comparative Perspective on Urban Spatial Housing Market Structure: Some More Evidence of Local Sub-markets Based on a Neural Network Classification of Amsterdam." *Urban Studies*, Vol. 41(13), pp. 2555–2579.

Kelejian, H. 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 Finance and Economics*, Vol. 17(1), pp. 99–121.

Kelejian, H. H. and Prucha, I. R. (2004) "Estimation of simultaneous systems of spatially interrelated cross sectional equations." *Journal of Econometrics*, Vol. 118(1), pp. 27–50.

Kelejian, H. H. and Prucha, I. R. (2007) "HAC estimation in a spatial framework." *Journal of Econometrics*, Vol. 140(1), pp. 131–154.

Kelejian, H. H. and Robinson, D. P. (1993) "A Suggested Method of Estimation for Spatial Interdependent Models with Autocorrelated Errors and an Application to a County Expenditure Model." *Papers in Regional Science*, Vol. 72(3), pp. 297–312.

L., A. and N., L.-G. (2008) "Errors in variables and spatial effects in hedonic house price models of ambient air quality." *Empirical Economics*, Vol. 34, pp. 5–34.

Lee, L. (2003) "Best Spatial Two Stage Least Squares Estimators for a Spatial Autoregressive Model with Autoregressive Disturbances." *Economic Reviews*, Vol. 22(4), pp. 99–121.

Lee, L.-f. (2007) "GMM and 2SLS estimation of mixed regressive, spatial autoregressive models." *Journal of Econometrics*, Vol. 137(2), pp. 489–514.

Lee, L.-f. and Liu, X. (2010) "Efficient GMM estimation of high order spatial autoregressive models with autoregressive disturbances." *Econometric Theory*, Vol. 26(01), pp. 187–230.

Leishman, C. (2009) "Spatial Change and the Structure of Urban Housing Submarkets." *Housing Studies*, Vol. 24(5), pp. 563–585.

Liu, X. and Lee, L.-f. (2012) "Two stage least squares estimation of spatial autoregressive models with endogenous regressors and many instruments." *Econometric Reviews*, Vol. (just-accepted).

Maclennan, D. and Tu, Y. (1996) "Economic Perspectives on the Structure of Local Housing Systems." *Housing Studies*, Vol. 11(3), pp. 387–407.

Mayer, W. and Smith, J. T. (2013) "Estimating the Capitalization of School Quality into Housing Prices: a Spatial Approach." Working Paper.

Newey, W. K. (1990) "Efficient Instrumental Variables Estimation of Nonlinear Models." *Econometrica*, Vol. 58(4), pp. 809–837.

Newey, W. K. (2007) "NONPARAMETRIC CONTINUOUS/DISCRETE CHOICE MODELS*." *International Economic Review*, Vol. 48(4), pp. 1429–1439.

Nguyen-Hoang, P. and Yinger, J. (2011) "The capitalization of school quality into house values: A review." *Journal of Housing Economics*, Vol. 20(1), pp. 30–48.

Oates, W. E. (1969) "The Effects of Property Taxes and Local Public Spending on Property Values: an Emperical Study of Tax Capitalization and the Tiebout Heypothesis." *Journal of Political Economy*, Vol. 77(6), pp. 957–971.

Ord, K. (1975) "Estimation methods for models of spatial interaction." *Journal of the American Statistical Association*, Vol. 70(349), pp. 120–126.

Ries, J. and Somerville, T. (2010) "School Quality and Residential Values: Evidence from Vancouver Zoning." *Review of Economics and Statistics*, Vol. 92(4), pp. 928–944.

Rosen, S. (1974) "Hedonic Prices and Implict Markets: Product Differentiation in Pure Competition." *Journal of Political Economy*, Vol. 82, pp. 35–55.

Schnare, A. B. and Struyk, R. J. (1976) "Segmentation in Urban Housing Markets." *Journal of Urban Economics*, Vol. 3, pp. 144–166.

Sedgley, N. H.; Williams, N. A.; and Derrick, F. W. (2008) "The Effect of Educational Test Scores on House Prices in a Model with Spatial Dependence." *Journal of Housing Economics*, Vol. 17(2), pp. 191–200.

Sin, C. and White, H. (1996) "Information Criteria for Selected Possibly Misspecified Parametric Models." *Journal of Econometrics*, Vol. 71, pp. 207–225.

Tiebout, C. M. (1956) "A pure theory of local expenditures." *Journal of Political Economy*, Vol. 64(5), pp. 416–424.

Tu, Y.; Sun, H.; and Yu, S. (2007) "Spatial Autocoreelations and Urban Housing Market Segmentation." *Journal of Real Estate Finance and Economics*, Vol. 34, pp. 385–406.

Watkins, C. (2001) "The Definition and identification of Housing Submarkets." *Environment and Planning A*, Vol. 33, pp. 2235–2253.

Wooldridge, J. M. (2010) *Econometric Analysis of Cross Section and Panel Data, Vol. 1 of MIT Press Books*. The MIT Press.

List of Appendices

APPENDIX A: FULL RESULT TABLES

Appendix 1

FULL RESULT TABLES

A.1 MN Data

A.1.1 KP Model with Exogenous SQ

Table 23: 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Exogenous School Quality Variables in MN Dataset when $k = 5$

Variable	**Coefficient**	**t-Stat**	**P-Value**
Constant	6.003930	60.384786	0.000000
Age	-0.003596	-23.775328	0.000000
Age2	0.000011	8.305050	0.000000
Beds	0.057587	14.333409	0.000000
Beds2	-0.007191	-13.602929	0.000000
Baths	0.089688	22.614954	0.000000
Baths2	-0.000418	-0.627404	0.530394
Rooms	0.028000	21.623713	0.000000
Rooms2	-0.000265	-12.038522	0.000000
GLA	0.015958	23.181490	0.000000
LotSize	0.000000	0.388584	0.697584
ElemScore	0.054689	6.476469	0.000000
MiddScore	0.066899	7.865194	0.000000
HighScore	0.024159	2.692979	0.007082

R-squared = 0.6626, N = 101993

Table 24: 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Exogenous School Quality Variables in MN Dataset when $k = 10$

Variable	**Coefficient**	**t-Stat**	**P-Value**
Constant	6.580606	60.887298	0.000000
Age	-0.003810	-24.822698	0.000000
Age2	0.000011	8.510994	0.000000
Beds	0.061682	15.067511	0.000000
Beds2	-0.007829	-14.517002	0.000000
Baths	0.097403	24.102237	0.000000
Baths2	-0.000517	-0.762900	0.445523
Rooms	0.030095	22.834639	0.000000
Rooms2	-0.000283	-12.640641	0.000000
GLA	0.017238	24.620857	0.000000
LotSize	0.000000	0.309919	0.756623
ElemScore	0.056820	6.609719	0.000000
MiddScore	0.067807	7.831481	0.000000
HighScore	0.023237	2.544667	0.010938

R-squared = 0.6526, N = 101993

Table 25: 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Exogenous School Quality Variables in MN Dataset when $k = 15$

Variable	**Coefficient**	**t-Stat**	**P-Value**
Constant	4.736033	49.219925	0.000000
Age	-0.003298	-21.990285	0.000000
Age2	0.000009	7.323155	0.000000
Beds	0.056666	14.099629	0.000000
Beds2	-0.006908	-13.056498	0.000000
Baths	0.083201	21.078054	0.000000
Baths2	-0.000366	-0.549072	0.582956
Rooms	0.027094	20.991128	0.000000
Rooms2	-0.000251	-11.456926	0.000000
GLA	0.014977	21.910568	0.000000
LotSize	-0.000000	-0.059696	0.952398
ElemScore	0.054658	6.472312	0.000000
MiddScore	0.066219	7.784834	0.000000
HighScore	0.022794	2.541036	0.011052

R-squared = 0.6659, N = 101993

Table 26: 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Exogenous School Quality Variables in MN Dataset when $k = 20$

Variable	**Coefficient**	**t-Stat**	**P-Value**
Constant	4.454618	46.419853	0.000000
Age	-0.003217	-21.522286	0.000000
Age2	0.000009	7.016144	0.000000
Beds	0.056444	14.060487	0.000000
Beds2	-0.006832	-12.924450	0.000000
Baths	0.081797	20.760742	0.000000
Baths2	-0.000310	-0.465901	0.641286
Rooms	0.026874	20.852499	0.000000
Rooms2	-0.000249	-11.347703	0.000000
GLA	0.014912	21.851510	0.000000
LotSize	-0.000000	-0.143897	0.885582
ElemScore	0.054055	6.405570	0.000000
MiddScore	0.065860	7.748941	0.000000
HighScore	0.022542	2.514873	0.011908

R-squared = 0.6665, N = 101993

Table 27: 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Exogenous School Quality Variables in MN Dataset when $k = 25$

Variable	**Coefficient**	**t-Stat**	**P-Value**
Constant	4.264289	44.429123	0.000000
Age	-0.003175	-21.263059	0.000000
Age2	0.000009	6.845933	0.000000
Beds	0.056347	14.034764	0.000000
Beds2	-0.006789	-12.839675	0.000000
Baths	0.081189	20.615645	0.000000
Baths2	-0.000291	-0.437053	0.662073
Rooms	0.026813	20.808653	0.000000
Rooms2	-0.000248	-11.296216	0.000000
GLA	0.014815	21.716990	0.000000
LotSize	-0.000000	-0.204226	0.838177
ElemScore	0.054081	6.407514	0.000000
MiddScore	0.065907	7.752513	0.000000
HighScore	0.022431	2.501993	0.012350

R-squared = 0.6665, N = 101993

Table 28: 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Exogenous School Quality Variables in MN Dataset when $k = 30$

Variable	**Coefficient**	**t-Stat**	**P-Value**
Constant	4.096314	42.669924	0.000000
Age	-0.003138	-21.051462	0.000000
Age2	0.000008	6.699476	0.000000
Beds	0.056099	13.981545	0.000000
Beds2	-0.006725	-12.725541	0.000000
Baths	0.080609	20.488146	0.000000
Baths2	-0.000266	-0.400621	0.688699
Rooms	0.026669	20.710445	0.000000
Rooms2	-0.000246	-11.218678	0.000000
GLA	0.014700	21.561185	0.000000
LotSize	-0.000000	-0.230754	0.817506
ElemScore	0.053665	6.360635	0.000000
MiddScore	0.065476	7.704958	0.000000
HighScore	0.021964	2.450991	0.014246

R-squared = 0.6668, N = 101993

A.1.2 Method A

Table 29: 2SLS Regression of ln(*SalesPrice*) on Housing Characteristics, Spatially-Lagged ln(*SalesPrice*), and Endogenous School Quality Variables in MN Dataset when $k = 5$ with $E(WP|X)$ as Method A

Variable	**Coefficient**	**t-Stat**	**P-Value**
ElemScore	0.129382	0.407447	0.683679
MiddScore	0.137941	3.930911	0.000085
HighScore	0.064516	1.920786	0.054759
Constant	4.338250	19.954314	0.000000
Age	-0.005701	-25.234922	0.000000
Age2	0.000024	11.654841	0.000000
Beds	0.063093	8.186050	0.000000
Beds2	-0.007088	-6.131324	0.000000
Baths	0.107426	22.537011	0.000000
Baths2	-0.002992	-3.170286	0.001523
Rooms	0.030544	12.165022	0.000000
Rooms2	-0.000284	-3.714112	0.000204
GLA	0.014189	1.681072	0.092749
LotSize	0.000000	0.016830	0.986572
lambda	0.525280	52.458954	0.000000
rho	-0.382380	-31.471123	0.000000

R-squared = 0.5823, N = 101993

Table 30: 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in MN Dataset when $k = 10$ with $E(WP|X)$ as Method A

Variable	**Coefficient**	**t-Stat**	**P-Value**
ElemScore	0.411813	4.566361	0.000005
MiddScore	0.201482	7.113272	0.000000
HighScore	0.046896	5.706410	0.000000
Constant	3.902148	33.381678	0.000000
Age	-0.005371	-27.003200	0.000000
Age2	0.000022	12.568003	0.000000
Beds	0.063573	8.185992	0.000000
Beds2	-0.007027	-6.116712	0.000000
Baths	0.102516	21.648206	0.000000
Baths2	-0.002497	-2.686822	0.007214
Rooms	0.029066	11.682884	0.000000
Rooms2	-0.000270	-3.586670	0.000335
GLA	0.014083	1.680382	0.092883
LotSize	-0.000000	-0.079841	0.936364
lambda	0.582584	62.528849	0.000000
rho	-0.336640	-27.821706	0.000000

R-squared = 0.5848, N = 101993

Table 31: 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in MN Dataset when $k = 15$ with $E(WP|X)$ as Method A

Variable	**Coefficient**	**t-Stat**	**P-Value**
ElemScore	-0.165125	-1.258460	0.208225
MiddScore	-0.241726	-3.501513	0.000463
HighScore	0.041515	3.790732	0.000150
Constant	3.330179	27.797502	0.000000
Age	-0.005658	-27.747540	0.000000
Age2	0.000025	13.775145	0.000000
Beds	0.062067	8.228499	0.000000
Beds2	-0.007006	-6.342269	0.000000
Baths	0.106302	21.833200	0.000000
Baths2	-0.003500	-3.654702	0.000257
Rooms	0.030793	12.109862	0.000000
Rooms2	-0.000280	-3.581377	0.000342
GLA	0.013763	1.672558	0.094414
LotSize	0.000000	0.190236	0.849124
lambda	0.627224	64.676595	0.000000
rho	-0.369808	-33.463240	0.000000

R-squared = 0.5759, N = 101993

Table 32: 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in MN Dataset when $k = 20$ with $E(WP|X)$ as Method A

Variable	**Coefficient**	**t-Stat**	**P-Value**
ElemScore	0.786382	3.844643	0.000121
MiddScore	-0.025843	-1.068587	0.285256
HighScore	0.051604	3.114287	0.001844
Constant	2.875069	21.273782	0.000000
Age	-0.006050	-28.502485	0.000000
Age2	0.000027	14.709939	0.000000
Beds	0.065441	8.269217	0.000000
Beds2	-0.006999	-6.001361	0.000000
Baths	0.105398	21.425372	0.000000
Baths2	-0.003124	-3.245954	0.001171
Rooms	0.029094	11.538625	0.000000
Rooms2	-0.000269	-3.577314	0.000347
GLA	0.013867	1.684112	0.092160
LotSize	-0.000000	-0.099435	0.920793
lambda	0.608908	62.924216	0.000000
rho	-0.423462	-25.361306	0.000000

R-squared = 0.5556, N = 101993

Table 33: 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in MN Dataset when $k = 25$ with $E(WP|X)$ as Method A

Variable	**Coefficient**	**t-Stat**	**P-Value**
ElemScore	0.394005	0.990505	0.321928
MiddScore	0.114856	0.233067	0.815710
HighScore	0.009840	0.375551	0.707251
Constant	2.881592	16.234047	0.000000
Age	-0.005914	-24.353977	0.000000
Age2	0.000026	10.224497	0.000000
Beds	0.063705	7.977793	0.000000
Beds2	-0.006843	-5.979176	0.000000
Baths	0.106673	16.984392	0.000000
Baths2	-0.003482	-2.851301	0.004354
Rooms	0.029454	10.484591	0.000000
Rooms2	-0.000271	-3.566831	0.000361
GLA	0.013738	1.680939	0.092775
LotSize	-0.000000	-0.162728	0.870732
lambda	0.621146	33.293568	0.000000
rho	-0.463734	-18.013241	0.000000

R-squared = 0.5817, N = 101993

Table 34: 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in MN Dataset when $k = 30$ with $E(WP|X)$ as Method A

Variable	**Coefficient**	**t-Stat**	**P-Value**
ElemScore	0.347297	3.563294	0.000366
MiddScore	0.123718	2.115317	0.034403
HighScore	0.128239	1.011050	0.311992
Constant	2.939498	15.852520	0.000000
Age	-0.005934	-28.388461	0.000000
Age2	0.000026	12.079954	0.000000
Beds	0.066458	7.284133	0.000000
Beds2	-0.007245	-5.398444	0.000000
Baths	0.106141	21.647361	0.000000
Baths2	-0.003075	-2.673619	0.007504
Rooms	0.029213	11.718800	0.000000
Rooms2	-0.000270	-3.664558	0.000248
GLA	0.013800	1.672953	0.094337
LotSize	-0.000000	-0.184159	0.853889
lambda	0.605993	29.385294	0.000000
rho	-0.442715	-38.149282	0.000000

R-squared = 0.5754, N = 101993

A.1.3 Method B

Table 35: 2SLS Regression of ln(*SalesPrice*) on Housing Characteristics, Spatially-Lagged ln(*SalesPrice*), and Endogenous School Quality Variables in MN Dataset when $k = 5$ with $E(WP|X)$ as Method B

Variable	**Coefficient**	**t-Stat**	**P-Value**
ElemScore	0.321855	0.309606	0.756861
MiddScore	0.001336	0.005239	0.995820
HighScore	-0.029575	-0.129033	0.897332
Constant	-0.987561	-0.114801	0.908603
Age	-0.004048	-1.611629	0.107043
Age2	0.000015	2.887476	0.003883
Beds	0.051794	1.369747	0.170766
Beds2	-0.004779	-0.682071	0.495194
Baths	0.072271	0.860085	0.389742
Baths2	-0.002625	-2.552891	0.010683
Rooms	0.024040	1.288166	0.197688
Rooms2	-0.000216	-1.122743	0.261547
GLA	0.009258	0.672040	0.501558
LotSize	-0.000000	-0.296670	0.766718
lambda	1.062090	1.575053	0.115244
rho	-0.673354	-35.205528	0.000000

R-squared = 0.4549, N = 101993

Table 36: 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in MN Dataset when $k = 10$ with $E(WP|X)$ as Method B

Variable	**Coefficient**	**t-Stat**	**P-Value**
ElemScore	0.290975	1.624083	0.104358
MiddScore	0.075395	0.489360	0.624587
HighScore	0.011343	0.248019	0.804120
Constant	2.272349	0.937482	0.348510
Age	-0.004873	-5.735393	0.000000
Age2	0.000020	7.472412	0.000000
Beds	0.058998	4.499322	0.000007
Beds2	-0.006254	-2.968200	0.002995
Baths	0.092071	3.979887	0.000069
Baths2	-0.002591	-2.901067	0.003719
Rooms	0.027347	5.305334	0.000000
Rooms2	-0.000251	-2.911465	0.003597
GLA	0.012498	1.516954	0.129278
LotSize	-0.000000	-0.162744	0.870720
lambda	0.743311	3.601855	0.000316
rho	-0.605686	-40.749803	0.000000

R-squared = 0.5846, N = 101993

Table 37: 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in MN Dataset when $k = 15$ with $E(WP|X)$ as Method B

Variable	**Coefficient**	**t-Stat**	**P-Value**
ElemScore	-0.220752	-1.804976	0.071078
MiddScore	-0.183914	-3.016731	0.002555
HighScore	0.013338	0.921624	0.356725
Constant	1.227197	1.662719	0.096369
Age	-0.004898	-14.810882	0.000000
Age2	0.000021	10.606399	0.000000
Beds	0.057727	8.058546	0.000000
Beds2	-0.006132	-5.676161	0.000000
Baths	0.091908	10.829806	0.000000
Baths2	-0.003310	-3.625054	0.000289
Rooms	0.027950	10.147806	0.000000
Rooms2	-0.000251	-3.450579	0.000559
GLA	0.011774	1.654154	0.098096
LotSize	-0.000000	-0.054390	0.956624
lambda	0.849356	13.605095	0.000000
rho	-0.578540	-23.848680	0.000000

R-squared = 0.5550, N = 101993

Table 38: 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in MN Dataset when $k = 20$ with $E(WP|X)$ as Method B

Variable	**Coefficient**	**t-Stat**	**P-Value**
ElemScore	0.528592	3.483383	0.000495
MiddScore	-0.045755	-2.057315	0.039656
HighScore	0.016122	0.862225	0.388564
Constant	0.697534	0.894061	0.371289
Age	-0.005298	-16.966056	0.000000
Age2	0.000023	12.309106	0.000000
Beds	0.061299	8.318760	0.000000
Beds2	-0.006223	-5.541805	0.000000
Baths	0.092706	11.126737	0.000000
Baths2	-0.003144	-3.431011	0.000601
Rooms	0.026966	9.739555	0.000000
Rooms2	-0.000246	-3.429391	0.000605
GLA	0.012019	1.672728	0.094381
LotSize	-0.000000	-0.251810	0.801188
lambda	0.868738	13.771685	0.000000
rho	-0.662756	-21.613679	0.000000

R-squared = 0.5384, N = 101993

Table 39: 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in MN Dataset when $k = 25$ with $E(WP|X)$ as Method B

Variable	**Coefficient**	**t-Stat**	**P-Value**
ElemScore	0.418277	1.652117	0.098511
MiddScore	0.169755	0.671993	0.501588
HighScore	0.003389	0.163636	0.870018
Constant	2.522029	3.072571	0.002122
Age	-0.005768	-17.988850	0.000000
Age2	0.000026	12.683119	0.000000
Beds	0.063184	7.829576	0.000000
Beds2	-0.006701	-5.663135	0.000000
Baths	0.104233	14.267453	0.000000
Baths2	-0.003417	-3.499316	0.000466
Rooms	0.028874	10.798023	0.000000
Rooms2	-0.000265	-3.534898	0.000408
GLA	0.013426	1.680438	0.092872
LotSize	-0.000000	-0.241010	0.809548
lambda	0.663587	9.228215	0.000000
rho	-0.514725	-27.046450	0.000000

R-squared = 0.5712, N = 101993

Table 40: 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in MN Dataset when $k = 30$ with $E(WP|X)$ as Method B

Variable	**Coefficient**	**t-Stat**	**P-Value**
ElemScore	0.460921	1.611670	0.107034
MiddScore	0.027439	0.125366	0.900234
HighScore	0.346349	0.693113	0.488239
Constant	9.522841	0.824351	0.409740
Age	-0.008039	-3.248090	0.001162
Age2	0.000038	2.768935	0.005624
Beds	0.078196	3.783176	0.000155
Beds2	-0.009753	-2.550872	0.010745
Baths	0.142321	3.717932	0.000201
Baths2	-0.002944	-1.864340	0.062274
Rooms	0.036116	4.672589	0.000003
Rooms2	-0.000344	-2.805088	0.005030
GLA	0.019194	1.449286	0.147258
LotSize	0.000000	0.438376	0.661114
lambda	-0.190093	-0.195976	0.844629
rho	0.353759	2.958208	0.003094

R-squared = 0.4850, N = 101993

A.1.4 Method C

Table 41: 2SLS Regression of ln(*SalesPrice*) on Housing Characteristics, Spatially-Lagged ln(*SalesPrice*), and Endogenous School Quality Variables in MN Dataset when $k = 5$ with $E(WP|X)$ as Method C

Variable	**Coefficient**	**t-Stat**	**P-Value**
ElemScore	0.131097	0.412003	0.680337
MiddScore	0.136724	3.888646	0.000101
HighScore	0.063677	1.885580	0.059352
Constant	4.290806	17.424606	0.000000
Age	-0.005686	-25.232184	0.000000
Age2	0.000024	11.661765	0.000000
Beds	0.062992	8.184804	0.000000
Beds2	-0.007068	-6.113888	0.000000
Baths	0.107113	22.081411	0.000000
Baths2	-0.002989	-3.170669	0.001521
Rooms	0.030486	12.143887	0.000000
Rooms2	-0.000283	-3.714245	0.000204
GLA	0.014145	1.680384	0.092883
LotSize	0.000000	0.012249	0.990227
lambda	0.530062	41.214416	0.000000
rho	-0.386870	-29.998490	0.000000

R-squared = 0.5820, N = 101993

Table 42: 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in MN Dataset when $k = 10$ with $E(WP|X)$ as Method C

Variable	**Coefficient**	**t-Stat**	**P-Value**
ElemScore	0.405547	4.524997	0.000006
MiddScore	0.194944	6.848614	0.000000
HighScore	0.045053	5.445381	0.000000
Constant	3.817639	27.151949	0.000000
Age	-0.005345	-26.907190	0.000000
Age2	0.000022	12.573861	0.000000
Beds	0.063336	8.188978	0.000000
Beds2	-0.006987	-6.104498	0.000000
Baths	0.101975	21.397905	0.000000
Baths2	-0.002502	-2.700254	0.006929
Rooms	0.028977	11.684021	0.000000
Rooms2	-0.000269	-3.584749	0.000337
GLA	0.014001	1.680589	0.092843
LotSize	-0.000000	-0.084428	0.932716
lambda	0.590918	51.960678	0.000000
rho	-0.348688	-27.241525	0.000000

R-squared = 0.5854, N = 101993

Table 43: 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in MN Dataset when $k = 15$ with $E(WP|X)$ as Method C

Variable	**Coefficient**	**t-Stat**	**P-Value**
ElemScore	-0.164914	-1.260583	0.207459
MiddScore	-0.241945	-3.474475	0.000512
HighScore	0.041622	3.802056	0.000144
Constant	3.338144	23.078804	0.000000
Age	-0.005661	-27.464772	0.000000
Age2	0.000025	13.731221	0.000000
Beds	0.062083	8.228383	0.000000
Beds2	-0.007009	-6.337614	0.000000
Baths	0.106357	21.461037	0.000000
Baths2	-0.003501	-3.653630	0.000259
Rooms	0.030804	12.083730	0.000000
Rooms2	-0.000280	-3.580781	0.000343
GLA	0.013770	1.672574	0.094411
LotSize	0.000000	0.191347	0.848254
lambda	0.626383	54.368303	0.000000
rho	-0.372416	-31.533242	0.000000

R-squared = 0.5759, N = 101993

Table 44: 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in MN Dataset when $k = 20$ with $E(WP|X)$ as Method C

Variable	**Coefficient**	**t-Stat**	**P-Value**
ElemScore	0.776323	3.821749	0.000133
MiddScore	-0.026620	-1.107715	0.267985
HighScore	0.050220	3.041080	0.002357
Constant	2.790101	17.004269	0.000000
Age	-0.006021	-28.410228	0.000000
Age2	0.000027	14.705165	0.000000
Beds	0.065280	8.283956	0.000000
Beds2	-0.006969	-5.996275	0.000000
Baths	0.104902	21.154450	0.000000
Baths2	-0.003125	-3.255626	0.001131
Rooms	0.029011	11.532453	0.000000
Rooms2	-0.000268	-3.574244	0.000351
GLA	0.013795	1.683636	0.092252
LotSize	-0.000000	-0.106000	0.915583
lambda	0.619046	50.921899	0.000000
rho	-0.437732	-25.989334	0.000000

R-squared = 0.5556, N = 101993

Table 45: 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in MN Dataset when $k = 25$ with $E(WP|X)$ as Method C

Variable	**Coefficient**	**t-Stat**	**P-Value**
ElemScore	0.401510	0.870480	0.384038
MiddScore	0.131830	0.219376	0.826357
HighScore	0.007846	0.263738	0.791982
Constant	2.770422	18.220664	0.000000
Age	-0.005869	-18.527775	0.000000
Age2	0.000026	8.187620	0.000000
Beds	0.063544	7.985434	0.000000
Beds2	-0.006799	-5.920706	0.000000
Baths	0.105918	12.658746	0.000000
Baths2	-0.003462	-2.539779	0.011092
Rooms	0.029274	9.294987	0.000000
Rooms2	-0.000269	-3.528573	0.000418
GLA	0.013641	1.677442	0.093456
LotSize	-0.000000	-0.182414	0.855258
lambda	0.634268	54.670248	0.000000
rho	-0.477799	-15.720907	0.000000

R-squared = 0.5788, N = 101993

Table 46: 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in MN Dataset when $k = 30$ with $E(WP|X)$ as Method C

Variable	**Coefficient**	**t-Stat**	**P-Value**
ElemScore	0.344779	3.580628	0.000343
MiddScore	0.125852	2.185216	0.028873
HighScore	0.123404	0.996558	0.318979
Constant	2.793562	15.147295	0.000000
Age	-0.005888	-27.826356	0.000000
Age2	0.000026	11.887879	0.000000
Beds	0.066198	7.346887	0.000000
Beds2	-0.007189	-5.429084	0.000000
Baths	0.105339	21.142877	0.000000
Baths2	-0.003078	-2.691904	0.007105
Rooms	0.029060	11.684495	0.000000
Rooms2	-0.000268	-3.659974	0.000252
GLA	0.013680	1.671395	0.094644
LotSize	-0.000000	-0.199292	0.842034
lambda	0.623640	31.906987	0.000000
rho	-0.464136	-39.647508	0.000000

R-squared = 0.5743, N = 101993

A.2 PA Data

A.2.1 KP Model with Exogenous SQ

Table 47: 2SLS Regression of ln(*SalesPrice*) on Housing Characteristics, Spatially-Lagged ln(*SalesPrice*), and Exogenous School Quality Variables in PA Dataset when $k = 5$

Variable	**Coefficient**	**t-Stat**	**P-Value**
Variable	Coefficient	t-Stat	P-Value
Constant	8.318267	33.534162	0.000000
Age	-0.002949	-8.851886	0.000000
Age2	0.000008	4.258116	0.000021
Beds	0.111230	3.433116	0.000597
Beds2	-0.014797	-3.239066	0.001199
Baths	0.066162	3.658040	0.000254
Baths2	-0.004490	-1.292608	0.196147
Rooms	0.017952	0.924901	0.355018
Rooms2	-0.000305	-0.240204	0.810172
GLA	0.178807	19.485036	0.000000
LotSize	0.002243	19.535930	0.000000
ElemScore	-0.006673	-0.298652	0.765206
Lambda	0.207717	10.715958	0.000000
Rho	0.046427	3.505683	0.000455

R-squared = 0.6904, N = 3728

Table 48: 2SLS Regression of ln(*SalesPrice*) on Housing Characteristics, Spatially-Lagged ln(*SalesPrice*), and Exogenous School Quality Variables in PA Dataset when $k = 10$

Variable	**Coefficient**	**t-Stat**	**P-Value**
Variable	Coefficient	t-Stat	P-Value
Constant	8.004858	31.828256	0.000000
Age	-0.002890	-8.703876	0.000000
Age2	0.000008	4.181412	0.000029
Beds	0.111624	3.459227	0.000542
Beds2	-0.014881	-3.270558	0.001073
Baths	0.068241	3.785540	0.000153
Baths2	-0.004754	-1.373732	0.169525
Rooms	0.018696	0.966887	0.333600
Rooms2	-0.000386	-0.305564	0.759937
GLA	0.175724	19.165053	0.000000
LotSize	0.002225	19.452296	0.000000
ElemScore	-0.008157	-0.366460	0.714022
Lambda	0.229231	11.533437	0.000000
Rho	0.053568	2.637580	0.008350

R-squared = 0.6929, N = 3728

Table 49: 2SLS Regression of ln(*SalesPrice*) on Housing Characteristics, Spatially-Lagged ln(*SalesPrice*), and Exogenous School Quality Variables in PA Dataset when $k = 15$

Variable	**Coefficient**	**t-Stat**	**P-Value**
Variable	Coefficient	t-Stat	P-Value
Constant	7.877641	31.175136	0.000000
Age	-0.002846	-8.565440	0.000000
Age2	0.000008	4.077764	0.000045
Beds	0.112011	3.474282	0.000512
Beds2	-0.014955	-3.289523	0.001004
Baths	0.069908	3.880252	0.000104
Baths2	-0.004954	-1.432828	0.151907
Rooms	0.018840	0.975113	0.329504
Rooms2	-0.000415	-0.328307	0.742680
GLA	0.174876	19.081914	0.000000
LotSize	0.002217	19.392923	0.000000
ElemScore	-0.008719	-0.392011	0.695050
Lambda	0.237694	11.834701	0.000000
Rho	0.057167	2.419374	0.015547

R-squared = 0.6935, N = 3728

Table 50: 2SLS Regression of ln(*SalesPrice*) on Housing Characteristics, Spatially-Lagged ln(*SalesPrice*), and Exogenous School Quality Variables in PA Dataset when $k = 20$

Variable	**Coefficient**	**t-Stat**	**P-Value**
Variable	Coefficient	t-Stat	P-Value
Constant	7.808525	30.720087	0.000000
Age	-0.002827	-8.511875	0.000000
Age2	0.000008	4.050370	0.000051
Beds	0.112873	3.502401	0.000461
Beds2	-0.015094	-3.321516	0.000895
Baths	0.070798	3.930553	0.000085
Baths2	-0.005065	-1.465528	0.142777
Rooms	0.018526	0.959176	0.337470
Rooms2	-0.000408	-0.323035	0.746669
GLA	0.174413	19.039024	0.000000
LotSize	0.002216	19.408332	0.000000
ElemScore	-0.009244	-0.415781	0.677571
Lambda	0.243364	12.029181	0.000000
Rho	0.057598	2.355569	0.018494

R-squared = 0.6937, N = 3728

Table 51: 2SLS Regression of ln(*SalesPrice*) on Housing Characteristics, Spatially-Lagged ln(*SalesPrice*), and Exogenous School Quality Variables in PA Dataset when $k = 25$

Variable	**Coefficient**	**t-Stat**	**P-Value**
Variable	Coefficient	t-Stat	P-Value
Constant	7.745925	30.338706	0.000000
Age	-0.002807	-8.458205	0.000000
Age2	0.000008	4.018789	0.000058
Beds	0.112723	3.499676	0.000466
Beds2	-0.015081	-3.320455	0.000899
Baths	0.071575	3.975603	0.000070
Baths2	-0.005182	-1.500057	0.133600
Rooms	0.018440	0.955275	0.339438
Rooms2	-0.000412	-0.326673	0.743915
GLA	0.174059	19.013051	0.000000
LotSize	0.002210	19.365246	0.000000
ElemScore	-0.009505	-0.427724	0.668852
Lambda	0.248369	12.197118	0.000000
Rho	0.058308	2.281146	0.022540

R-squared = 0.6941, N = 3728

Table 52: 2SLS Regression of ln(*SalesPrice*) on Housing Characteristics, Spatially-Lagged ln(*SalesPrice*), and Exogenous School Quality Variables in PA Dataset when $k = 30$

Variable	**Coefficient**	**t-Stat**	**P-Value**
Variable	Coefficient	t-Stat	P-Value
Constant	7.703794	29.963859	0.000000
Age	-0.002787	-8.398483	0.000000
Age2	0.000008	3.978763	0.000069
Beds	0.113235	3.516830	0.000437
Beds2	-0.015153	-3.337486	0.000845
Baths	0.072109	4.007016	0.000061
Baths2	-0.005250	-1.520384	0.128414
Rooms	0.018207	0.943614	0.345367
Rooms2	-0.000402	-0.318789	0.749886
GLA	0.173522	18.958390	0.000000
LotSize	0.002205	19.348927	0.000000
ElemScore	-0.009767	-0.439685	0.660165
Lambda	0.253129	12.349892	0.000000
Rho	0.056823	2.193187	0.028294

R-squared = 0.6943, N = 3728

A.2.2 Method A

Table 53: 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in PA Dataset when $k = 5$ with $E(WP|X)$ as Method A

Variable	**Coefficient**	**t-Stat**	**P-Value**
ElemScore	-0.037752	-1.244584	0.213285
Constant	19.479338	2.169307	0.030059
Age	-0.002212	-2.155488	0.031124
Age2	0.000001	0.115270	0.908231
Beds	-0.019938	-0.156555	0.875596
Beds2	0.005288	0.279453	0.779898
Baths	0.062909	1.437651	0.150533
Baths2	-0.003986	-0.466018	0.641203
Rooms	0.069261	1.194915	0.232120
Rooms2	-0.003391	-0.941329	0.346536
GLA	0.154724	4.864018	0.000001
LotSize	0.001932	5.342184	0.000000
lambda	0.249046	4.498784	0.000007
rho	0.019742	0.566929	0.570762

R-squared = -0.4367, N = 3728

Table 54: 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in PA Dataset when $k = 10$ with $E(WP|X)$ as Method A

Variable	**Coefficient**	**t-Stat**	**P-Value**
ElemScore	-0.035823	-1.249324	0.211547
Constant	18.324837	2.148919	0.031641
Age	-0.002202	-2.228465	0.025849
Age2	0.000001	0.141562	0.887426
Beds	-0.012560	-0.104325	0.916911
Beds2	0.004139	0.231336	0.817054
Baths	0.065885	1.566949	0.117127
Baths2	-0.004475	-0.547154	0.584273
Rooms	0.066280	1.219776	0.222550
Rooms2	-0.003262	-0.966841	0.333624
GLA	0.154289	5.110129	0.000000
LotSize	0.001943	5.702731	0.000000
lambda	0.262710	5.016780	0.000001
rho	0.040881	0.919308	0.357934

R-squared = -0.3239, N = 3728

Table 55: 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in PA Dataset when $k = 15$ with $E(WP|X)$ as Method A

Variable	**Coefficient**	**t-Stat**	**P-Value**
ElemScore	-0.029776	-1.407245	0.159355
Constant	16.321046	2.597774	0.009383
Age	-0.002263	-2.674532	0.007483
Age2	0.000002	0.318305	0.750254
Beds	0.008788	0.093950	0.925149
Beds2	0.000862	0.062341	0.950291
Baths	0.068137	1.873675	0.060975
Baths2	-0.004764	-0.679954	0.496534
Rooms	0.058112	1.330054	0.183501
Rooms2	-0.002800	-1.028316	0.303801
GLA	0.157050	6.101492	0.000000
LotSize	0.001981	6.963266	0.000000
lambda	0.267353	5.880460	0.000000
rho	0.047217	0.995115	0.319680

R-squared = -0.0102, N = 3728

Table 56: 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in PA Dataset when $k = 20$ with $E(WP|X)$ as Method A

Variable	**Coefficient**	**t-Stat**	**P-Value**
ElemScore	-0.034232	-1.309994	0.190198
Constant	17.571319	2.256275	0.024053
Age	-0.002180	-2.317918	0.020454
Age2	0.000001	0.164808	0.869095
Beds	-0.006124	-0.054691	0.956384
Beds2	0.003140	0.189013	0.850082
Baths	0.068580	1.691729	0.090698
Baths2	-0.004808	-0.611564	0.540826
Rooms	0.063923	1.253349	0.210079
Rooms2	-0.003158	-0.995904	0.319297
GLA	0.154346	5.388272	0.000000
LotSize	0.001951	6.065094	0.000000
lambda	0.273620	5.453113	0.000000
rho	0.045370	0.946298	0.343997

R-squared = -0.2350, N = 3728

Table 57: 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in PA Dataset when $k = 25$ with $E(WP|X)$ as Method A

Variable	**Coefficient**	**t-Stat**	**P-Value**
ElemScore	-0.034287	-1.315637	0.188296
Constant	17.491416	2.252059	0.024319
Age	-0.002163	-2.303069	0.021275
Age2	0.000001	0.157956	0.874492
Beds	-0.006507	-0.058141	0.953637
Beds2	0.003191	0.192185	0.847598
Baths	0.069460	1.712023	0.086893
Baths2	-0.004951	-0.629453	0.529052
Rooms	0.063736	1.252189	0.210501
Rooms2	-0.003160	-0.998012	0.318273
GLA	0.154065	5.388028	0.000000
LotSize	0.001944	6.039416	0.000000
lambda	0.278506	5.510880	0.000000
rho	0.044913	0.913486	0.360987

R-squared = -0.2375, N = 3728

Table 58: 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in PA Dataset when $k = 30$ with $E(WP|X)$ as Method A

Variable	**Coefficient**	**t-Stat**	**P-Value**
ElemScore	-0.038999	-1.219165	0.222781
Constant	18.816319	1.972734	0.048526
Age	-0.002065	-1.961375	0.049835
Age2	0.000000	0.025675	0.979517
Beds	-0.022415	-0.167779	0.866757
Beds2	0.005637	0.283436	0.776842
Baths	0.069724	1.546763	0.121920
Baths2	-0.004979	-0.566883	0.570794
Rooms	0.069895	1.175425	0.239825
Rooms2	-0.003536	-0.957654	0.338237
GLA	0.150888	4.692155	0.000003
LotSize	0.001905	5.184174	0.000000
lambda	0.286075	5.050015	0.000000
rho	0.041010	0.844346	0.398476

R-squared = -0.5094, N = 3728

A.2.3 Method B

Table 59: 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in PA Dataset when $k = 5$ with $E(WP|X)$ as Method B

Variable	**Coefficient**	**t-Stat**	**P-Value**
ElemScore	-0.039087	-1.302721	0.192670
Constant	21.465114	2.519203	0.011762
Age	-0.002820	-2.279886	0.022614
Age2	0.000003	0.326858	0.743776
Beds	-0.029473	-0.235544	0.813786
Beds2	0.006871	0.369398	0.711831
Baths	0.058457	1.282244	0.199757
Baths2	-0.003491	-0.391956	0.695090
Rooms	0.074788	1.304177	0.192173
Rooms2	-0.003507	-0.970292	0.331901
GLA	0.171127	4.731577	0.000002
LotSize	0.002145	4.914251	0.000001
lambda	0.113921	0.906087	0.364890
rho	0.058079	0.898930	0.368690

R-squared = -0.5281, N = 3728

Table 60: 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in PA Dataset when $k = 10$ with $E(WP|X)$ as Method B

Variable	**Coefficient**	**t-Stat**	**P-Value**
ElemScore	-0.036803	-1.298584	0.194087
Constant	20.134877	2.512092	0.012002
Age	-0.002776	-2.331794	0.019712
Age2	0.000003	0.351365	0.725315
Beds	-0.020667	-0.175267	0.860870
Beds2	0.005493	0.313669	0.753772
Baths	0.061131	1.406673	0.159524
Baths2	-0.003964	-0.469202	0.638925
Rooms	0.070372	1.312094	0.189488
Rooms2	-0.003304	-0.976524	0.328805
GLA	0.170302	4.931243	0.000001
LotSize	0.002140	5.219956	0.000000
lambda	0.130329	1.059264	0.289479
rho	0.073811	1.047104	0.295051

R-squared = -0.3888, N = 3728

Table 61: 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in PA Dataset when $k = 15$ with $E(WP|X)$ as Method B

Variable	**Coefficient**	**t-Stat**	**P-Value**
ElemScore	-0.030536	-1.459371	0.144463
Constant	18.047934	3.067185	0.002161
Age	-0.002837	-2.804471	0.005040
Age2	0.000004	0.569436	0.569060
Beds	0.001327	0.014495	0.988435
Beds2	0.002116	0.156395	0.875722
Baths	0.062883	1.672480	0.094430
Baths2	-0.004193	-0.579463	0.562276
Rooms	0.061686	1.424342	0.154347
Rooms2	-0.002813	-1.027052	0.304396
GLA	0.172828	5.939787	0.000000
LotSize	0.002173	6.387668	0.000000
lambda	0.135242	1.267819	0.204862
rho	0.088676	1.139155	0.254638

R-squared = -0.0547, N = 3728

Table 62: 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in PA Dataset when $k = 20$ with $E(WP|X)$ as Method B

Variable	**Coefficient**	**t-Stat**	**P-Value**
ElemScore	-0.035058	-1.357388	0.174658
Constant	19.184476	2.659524	0.007825
Age	-0.002697	-2.343570	0.019100
Age2	0.000003	0.362652	0.716865
Beds	-0.013658	-0.124808	0.900676
Beds2	0.004397	0.271139	0.786284
Baths	0.063472	1.518369	0.128921
Baths2	-0.004242	-0.524098	0.600211
Rooms	0.067437	1.340514	0.180078
Rooms2	-0.003185	-1.000244	0.317193
GLA	0.168580	5.060094	0.000000
LotSize	0.002121	5.428603	0.000000
lambda	0.153383	1.242490	0.214056
rho	0.077466	0.996169	0.319168

R-squared = -0.2870, N = 3728

Table 63: 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in PA Dataset when $k = 25$ with $E(WP|X)$ as Method B

Variable	**Coefficient**	**t-Stat**	**P-Value**
ElemScore	-0.035109	-1.364281	0.172479
Constant	19.046835	2.664225	0.007717
Age	-0.002657	-2.296189	0.021665
Age2	0.000003	0.345450	0.729756
Beds	-0.013762	-0.125881	0.899826
Beds2	0.004400	0.271697	0.785855
Baths	0.064332	1.537657	0.124132
Baths2	-0.004369	-0.539752	0.589368
Rooms	0.067165	1.339252	0.180489
Rooms2	-0.003188	-1.003162	0.315783
GLA	0.167591	5.000887	0.000001
LotSize	0.002107	5.354768	0.000000
lambda	0.162940	1.291995	0.196359
rho	0.075467	0.939921	0.347258

R-squared = -0.2888, N = 3728

Table 64: 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in PA Dataset when $k = 30$ with $E(WP|X)$ as Method B

Variable	**Coefficient**	**t-Stat**	**P-Value**
ElemScore	-0.039845	-1.263898	0.206267
Constant	20.178275	2.319098	0.020390
Age	-0.002480	-1.868994	0.061624
Age2	0.000002	0.167340	0.867102
Beds	-0.029168	-0.224102	0.822678
Beds2	0.006751	0.349332	0.726840
Baths	0.065209	1.407535	0.159269
Baths2	-0.004460	-0.495284	0.620400
Rooms	0.073047	1.253400	0.210060
Rooms2	-0.003575	-0.966926	0.333581
GLA	0.162253	4.192668	0.000028
LotSize	0.002041	4.461975	0.000008
lambda	0.188367	1.279597	0.200687
rho	0.062284	0.811300	0.417193

R-squared = -0.5666, N = 3728

A.2.4 Method C

Table 65: 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in PA Dataset when $k = 5$ with $E(WP|X)$ as Method C

Variable	**Coefficient**	**t-Stat**	**P-Value**
ElemScore	-0.037780	-1.244526	0.213306
Constant	19.524439	2.167911	0.030165
Age	-0.002226	-2.176184	0.029542
Age2	0.000001	0.121030	0.903667
Beds	-0.020150	-0.158059	0.874410
Beds2	0.005324	0.280996	0.778714
Baths	0.062806	1.434325	0.151480
Baths2	-0.003975	-0.464325	0.642415
Rooms	0.069385	1.195564	0.231867
Rooms2	-0.003394	-0.941346	0.346527
GLA	0.155104	4.890840	0.000001
LotSize	0.001936	5.374330	0.000000
lambda	0.245929	4.494217	0.000007
rho	0.020160	0.578494	0.562931

R-squared = -0.4385, N = 3728

Table 66: 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in PA Dataset when $k = 10$ with $E(WP|X)$ as Method C

Variable	**Coefficient**	**t-Stat**	**P-Value**
ElemScore	-0.035845	-1.249472	0.211492
Constant	18.368495	2.146544	0.031830
Age	-0.002216	-2.253195	0.024247
Age2	0.000001	0.147823	0.882482
Beds	-0.012751	-0.105824	0.915722
Beds2	0.004171	0.232892	0.815845
Baths	0.065769	1.563317	0.117978
Baths2	-0.004463	-0.545269	0.585569
Rooms	0.066377	1.220465	0.222289
Rooms2	-0.003263	-0.966883	0.333603
GLA	0.154682	5.144201	0.000000
LotSize	0.001948	5.744401	0.000000
lambda	0.259471	5.033977	0.000000
rho	0.041586	0.934568	0.350011

R-squared = -0.3252, N = 3728

Table 67: 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in PA Dataset when $k = 15$ with $E(WP|X)$ as Method C

Variable	**Coefficient**	**t-Stat**	**P-Value**
ElemScore	-0.029801	-1.407601	0.159249
Constant	16.364930	2.592863	0.009518
Age	-0.002277	-2.703530	0.006861
Age2	0.000002	0.326200	0.744273
Beds	0.008583	0.091674	0.926957
Beds2	0.000896	0.064724	0.948394
Baths	0.068009	1.868766	0.061655
Baths2	-0.004750	-0.677375	0.498168
Rooms	0.058207	1.330911	0.183218
Rooms2	-0.002800	-1.028361	0.303780
GLA	0.157430	6.138830	0.000000
LotSize	0.001986	7.011384	0.000000
lambda	0.264144	5.895790	0.000000
rho	0.048234	1.016400	0.309439

R-squared = -0.0114, N = 3728

Table 68: 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in PA Dataset when $k = 20$ with $E(WP|X)$ as Method C

Variable	**Coefficient**	**t-Stat**	**P-Value**
ElemScore	-0.034254	-1.309847	0.190248
Constant	17.614202	2.250399	0.024424
Age	-0.002194	-2.346887	0.018931
Age2	0.000001	0.171562	0.863782
Beds	-0.006324	-0.056406	0.955018
Beds2	0.003174	0.190733	0.848735
Baths	0.068444	1.687390	0.091528
Baths2	-0.004793	-0.609225	0.542376
Rooms	0.064016	1.253676	0.209960
Rooms2	-0.003159	-0.995850	0.319323
GLA	0.154725	5.428831	0.000000
LotSize	0.001955	6.115249	0.000000
lambda	0.270422	5.481298	0.000000
rho	0.046123	0.963229	0.335433

R-squared = -0.2362, N = 3728

Table 69: 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in PA Dataset when $k = 25$ with $E(WP|X)$ as Method C

Variable	**Coefficient**	**t-Stat**	**P-Value**
ElemScore	-0.034312	-1.315497	0.188343
Constant	17.538830	2.245208	0.024755
Age	-0.002178	-2.334963	0.019545
Age2	0.000001	0.165285	0.868719
Beds	-0.006730	-0.060044	0.952120
Beds2	0.003228	0.194086	0.846109
Baths	0.069305	1.706972	0.087827
Baths2	-0.004933	-0.626690	0.530862
Rooms	0.063841	1.252567	0.210363
Rooms2	-0.003161	-0.997998	0.318281
GLA	0.154475	5.433161	0.000000
LotSize	0.001949	6.095147	0.000000
lambda	0.274999	5.542202	0.000000
rho	0.045826	0.933141	0.350747

R-squared = -0.2390, N = 3728

Table 70: 2SLS Regression of ln($SalesPrice$) on Housing Characteristics, Spatially-Lagged ln($SalesPrice$), and Endogenous School Quality Variables in PA Dataset when $k = 30$ with $E(WP|X)$ as Method C

Variable	**Coefficient**	**t-Stat**	**P-Value**
ElemScore	-0.039032	-1.218761	0.222935
Constant	18.873419	1.966178	0.049278
Age	-0.002083	-1.994877	0.046056
Age2	0.000000	0.032662	0.973944
Beds	-0.022693	-0.169545	0.865368
Beds2	0.005683	0.285136	0.775540
Baths	0.069532	1.541176	0.123274
Baths2	-0.004957	-0.563832	0.572869
Rooms	0.070025	1.175578	0.239764
Rooms2	-0.003538	-0.957558	0.338286
GLA	0.151372	4.741710	0.000002
LotSize	0.001911	5.243243	0.000000
lambda	0.281926	5.094336	0.000000
rho	0.041890	0.863301	0.387972

R-squared = -0.5115, N = 3728

APPENDIX B: PSEUDO-REPEAT SALES APPROACH

Appendix 2

PSEUDO-REPEAT SALES APPROACH

B.1 A Basic Repeat-Sales Model

Let

$$P_{iht} = \theta SQ_{ht} + \alpha R_{it} + \gamma U_{it} + \epsilon_{iht} \tag{B.1}$$

$$SQ_{ht} = \delta \bar{P}_{ht} + \nu_{ht} \tag{B.2}$$

where P_{iht} is the price of home i in school attendance zone h at time t, SQ_{ht} is the school quality of the school in attendance zone h at time t, R_{it} is some exogenous, observed characteristic (say, Rooms) of home i at time t, and U_{it} is some unobserved characteristic of home i at period t like utility gained from a local park.

Because U is unobserved, a hedonic regression of the above model would yield biased estimates of θ due to school quality being correlated with the omitted variable. A repeat sales approach tries to remedy this bias by differencing the hedonic equation to remove any time invariant characteristics. The repeat sales approach makes two major assumptions:

Assumption B.1. $U_{it} = U_{is} = U_i$

Assumption B.2. $R_{it} = R_{is} = R_i$

That is, both the observed and unobserved housing characteristics are assumed to be constant over time. Therefore the differencing described above produces

$$P_{iht} - P_{ihs} = \theta(SQ_{ht} - SQ_{hs}) + (\epsilon_{iht} - \epsilon_{ihs}) \tag{B.3}$$

So then θ can be estimated by regressing the change in the price of the home from period t to period s on the change in school quality from period t to period s.

B.1.1 A Numerical Example

Let

$$P = SQ + 3R + 2U + \epsilon \tag{B.4}$$

$$SQ = \frac{1}{2}\bar{P} + \nu = \frac{1}{2}AP + \nu \tag{B.5}$$

where ϵ and ν are errors that are always zero, except as described below and

$$A = \begin{pmatrix} 1/4 & 1/4 & 1/4 & 1/4 & 0 & 0 & 0 & 0 \\ 1/4 & 1/4 & 1/4 & 1/4 & 0 & 0 & 0 & 0 \\ 1/4 & 1/4 & 1/4 & 1/4 & 0 & 0 & 0 & 0 \\ 1/4 & 1/4 & 1/4 & 1/4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/4 & 1/4 & 1/4 & 1/4 \\ 0 & 0 & 0 & 0 & 1/4 & 1/4 & 1/4 & 1/4 \\ 0 & 0 & 0 & 0 & 1/4 & 1/4 & 1/4 & 1/4 \\ 0 & 0 & 0 & 0 & 1/4 & 1/4 & 1/4 & 1/4 \end{pmatrix}$$

So then (B.4) and (B.5) imply

$$P = (I - \frac{1}{2}A)^{-1}(3R + 2U + \nu + \epsilon) \tag{B.6}$$

Assume that in period 2 both schools get a new principal that causes school quality to increase by 50 points ($\nu_{t=2} = 50$). Data for R and U as well as solutions to the above equation for P can be seen in Figure 6. Assuming all homes are observed both periods one could construct a repeat sales model to find the effect of SQ on P. For all homes,

$$\Delta SQ = 100 \text{ and } \Delta P = 100 \Rightarrow \frac{\Delta P}{\Delta SQ} = 1$$

If we try to estimate the same effect using a hedonic regression, we find:

$$E(P|SQ) = 1.58SQ + 5.98R$$

So then an exogenous shock to SQ allows us to estimate the true effect of SQ on housing prices using a repeated sales methodology whereas we get a biased estimate from the hedonic regression due to the omitted variables.

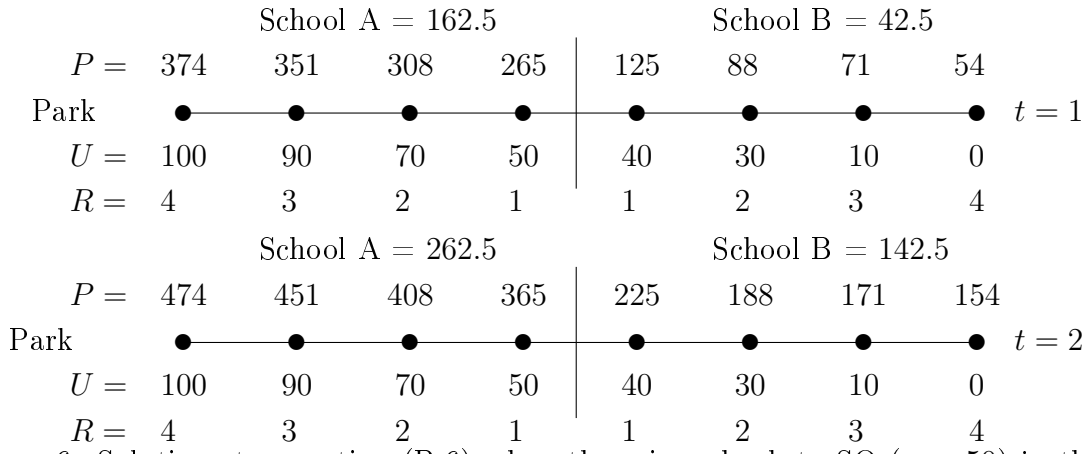

Figure 6: Solutions to equation (B.6) when there is a shock to SQ ($\nu = 50$) in the second period.

B.2 The Problem with the Repeat-Sales Model

Unfortunately assumptions B.1 and B.2 don't tend to hold in the real world and this can have disasterous consequences on repeat sales estimates. For example, assume that because the a pool was build next to the park in the above example so that $U_2 = 1.1U_1$. So the unobserved utility gained by each home as increased, but this increase will not be accounted for in the difference equation above and will bias the estimates. This can be seen in Figure 7.

If we regress the change in price on the change in school quality we find

$$\Delta P = 2\Delta S$$

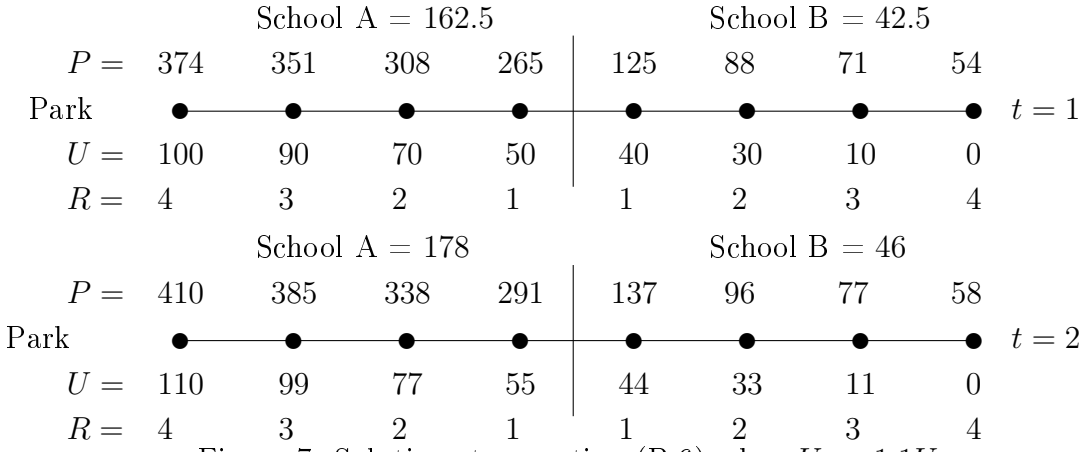

Figure 7: Solutions to equation (B.6) when $U_2 = 1.1U_1$.

which is an incorrect estimate of the effect of school quality on housing prices. Using a hedonic regression, we get

$$E(P|SQ) = 1.95SQ + 3.14R$$

B.3 How Rezoning Helps

Ries and Somerville (2010) use a method similar to the above methods to try to estimate the effect of school quality on housing prices. However, one major difference in their dataset is that during the timeframe of their study homes were rezoned so that students living in those homes attended a different school. This rezoning provides an exogenous change in school quality of the homes that can be used to better estimate the effect of quality on prices. When there is no rezoning, all of the change in the school quality (and housing prices) is being driven by the change in the unobserved variable. By introducing a zoning change, part of the change in the quality variable is an exogenous change. If the exogenous change in school quality is large relative to the change in school quality created by the change in the unobserved variable, then it is possible to get a good estimate of the effect. It is interesting to note that there is an exogenous shock to school quality for *all* the homes due to the rezoning, no matter if the individual homes were rezoned or not. However, the exogenous change is largest in the homes that are rezoned, giving rise to a possible estimation method

of using only rezoned homes to estimate the effect of school quality on housing prices. By restricting the sample to only homes that are rezoned, it should be possible to obtain the least biased estimate of the effect.

B.3.1 A Numerical Example

In order to have redistricting in these simple examples, we define

$$A_1 = \begin{pmatrix} 1/3 & 1/3 & 1/3 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 0 & 0 & 0 & 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 0 & 0 & 0 & 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 0 & 0 & 0 & 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 0 & 0 & 0 & 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \end{pmatrix}$$

and

$$A_2 = \begin{pmatrix} 1/5 & 1/5 & 1/5 & 1/5 & 1/5 & 0 & 0 & 0 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 & 0 & 0 & 0 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 & 0 & 0 & 0 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 & 0 & 0 & 0 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 0 & 0 & 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 0 & 0 & 0 & 1/3 & 1/3 & 1/3 \end{pmatrix}$$

so that the solutions to (B.6) are given in Figure 8.

If we use the above data in Figure 8 to calculate a repeat sales estimate, we find

$$E(\Delta P|\Delta S) = 1.025\Delta S$$

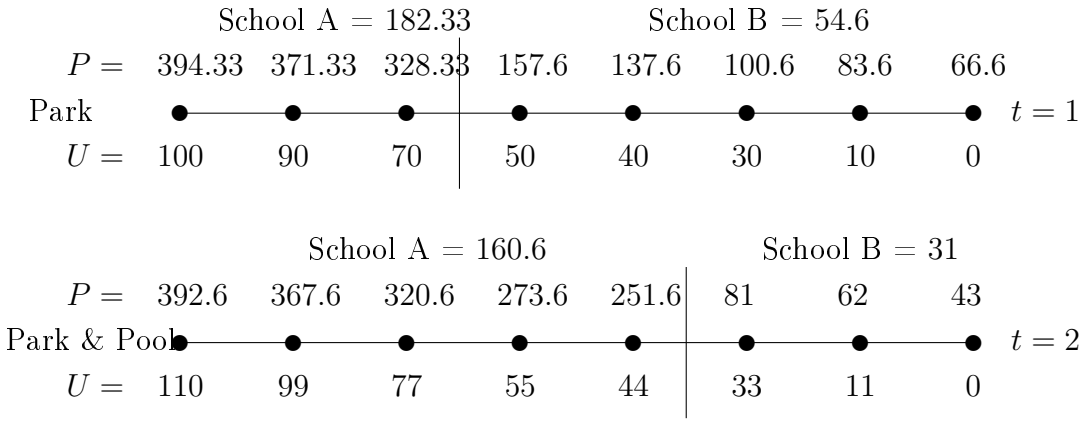

Figure 8: Solutions to equation (B.6) when U, the utility the home draws from the park, increases by a factor of 1.1 in the second period and there is a rezoning of school attendance zones.

which is a much better estimate than the one given in the previous section. In fact, this is an even better estimate than if we used only the rezoned homes which gives an estimate of about 1.08. Again, this is being driven by the exogenous change in school quality in all of the houses. In fact, looking at Table 71 we see that while the change in school quality is the largest for the two rezoned homes, in all of the homes the change in quality is much larger than the change in U. We can in fact calculate what part of the change in school quality is exogenous and what part is due to the change in U, and that result can be found in the last column in Table 71. We see that the exogenous change is quite larger than the change in U in all the cases. This is likely due to the fact that in a population of 8 homes, redistricting 2 of them can dramatically change the school quality in a zone. We are almost doubling the number of homes in School Zone A in this example while almost halving the homes in zone B.

If however, the population was larger then rezoning a small number of homes would have little effect on the school quality of homes that were not rezoned but would have a large effect on the homes that were rezoned. For example, if there were 100 homes and we only rezoned two of them, letting U_1 decrease from 100 to 1 linearly, then using all the homes to estimate the effect of school quality we find:

Table 71: Changes in Price, Quality, and Unobserved Variables in Figure 8

House	ΔP	ΔS	ΔU	$\Delta_{ex} S$
1	-1.733	-21.733	10	-35.733
2	-3.733	-21.733	9	-35.733
3	-7.733	-21.733	7	-35.733
4	116	106	5	92
5	114	106	4	92
6	-19.6	-23.6	2	-25.6
7	-21.6	-23.6	1	-25.6
8	-23.6	-23.6	0	-25.6

$$E(\Delta P|\Delta S) = 1.368\Delta S$$

whereas, if only the two redistricted homes are used to estimate the effect we find

$$E(\Delta P|\Delta S) = 1.089\Delta S$$

In this case the rezoning caused a much smaller exogenous change in the school quality of the non-rezoned homes and so including those homes into the sample introduced more bias than using the rezoned homes only.

B.4 Pseudo-Repeat Sales

Another issue that tends to plague the repeat sales literature is that of small sample sizes. In order to run a repeat sales estimate, it is necessary to have two sales of the home within the study period. This can dramatically reduce sample size as most homes are not resold in short amounts of time. As discussed above, the purpose of the repeat sales method is to remove any time-invariant unobserved variables from regression equation and therefore produce unbiased results. However, as unobserved neighborhood amenities are likely almost constant between homes that are very close, it could then be possible to match homes that sold in later periods to homes that sold in previous periods if they are in very close proximity. We are then estimating the difference equation

$$P_{iht} - P_{jhs} = \theta(SQ_{ht} - SQ_{hs}) + \alpha(R_i - R_j) + \gamma(U_i - U_j) + (\epsilon_{iht} - \epsilon_{jhs}). \tag{B.7}$$

The second term on the right hand side takes care of controlling for the differences in the observed, exogenous housing characteristics while the third term is likely close to zero. By matching homes, we should see an increase in efficiency of the estimator because of the increase in observations.

By combining a pseudo-repeat sales approach with a the rezoning effects discussed above, it should be possible to estimate the effect of school quality with less bias than traditional methods while still maintaining efficiency due to the increased sample size. We test this hypothesis using a Monte Carlo experiment below.

B.5 Monte-Carlo Results

In order to study the different estimators presented above, I designed a Monte Carlo experiment that randomized most of the important aspects of the model. Each iteration consisted of 4000 homes that were randomly assigned an (x, y) coordinate between 0 and 1. An example of one iteration can be seen in Figure **??**. Then a park is randomly placed on the plane and the distance is calculated between it and each home is calculated. In each period a school boundary is assigned as some x value so that the plane is vertically partitioned into two school zones. Also, 20 percent of the homes are randomly chosen to be "observed" and used as the sample in each period. The number of rooms is randomly assigned to each home as an integer between 1 and 10. Prices are calculated using the equations

$$P_1 = (I - 0.5A_1)^{-1}(50R - 100dist + \epsilon)$$

$$P_2 = (I - 0.5A_2)^{-1}(50R - 100(0.9dist) + \epsilon)$$

where ϵ ranges from -50 to 50.

Four estimators were tested in this Monte Carlo simulation:

1. RS-All – A repeat sales estimator where rezoning is present and all repeated sales are used in the regression analysis.

2. RS-Rezone – A repeat sales estimator where rezoning is present but only rezoned homes are used in the regression analysis.

3. PRS-All – A pseudo-repeat sales estimator where rezoning is present and all homes observed in the second period are matched with the closet homes in the first period for the regression analysis.

4. PRS-Rezone – A pseudo-repeat sales estimator where rezoning is present but only rezoned homes in the second period are matched and used in the regression analysis.

The simulation was run through 1000 iterations and statistics on the results can be found in Table 72. What we find is that the estimators that used only rezoned homes seem to exhibit less bias than those that use all homes in the sample. The mean of RS-All and PRS-All are both very close to 1.43 while the other two estimators are closer to the true parameter of 1. It is also worth noting that only using rezoned homes increases the MSE of the estimator which is to be expected as it is reducing the number of homes used in each estimation. Moving to a pseudo-repeat sales method however lowers MSE in both cases (RS-All to PRS-all and RS-Rezone to PRS-Rezone) as it allows for more observations to be used in the regression analysis.

Table 72: Results of Monte Carlo Experiment on 2 Repeat Sales Estimators and 2 Pseudo-Repeat Sales Estimators

Estimator	SQ Mean	SQ MSE	R Mean	R MSE
RS-All	1.438	0.423	-	-
RS-Rezone	0.854	96.175	-	-
PRS-All	1.429	0.271	19.995	0.200
PRS-Rezone	0.988	0.933	19.986	2.276

APPENDIX C: HOUSING SUBMARKET IDENTIFICATION

Appendix 3

HOUSING SUBMARKET IDENTIFICATION

C.1 Literature

When studying housing submarkets, a researcher is faced with three basic questions: "How should we define a submarket?", "What forces create and maintain submarkets?", and "How can we identify housing submarkets?" The answer to this first question will obviously affect the answer to the last two. Unfortunately, there does not seem to be a consensus in the submarket literature on what constitutes a submarket. Watkins (2001) concisely states the problem of multiple definitions by writing that it "reflects the absence of a coherent explanation of the different processes considered important in determining submarket dimensions." Once the definition question is answered, a theoretical justification of the submarkets is necessary. However, like the definition of a submarket, there is no unifying theory subscribed to by researchers in the housing field. The problem of identification then looms and this issue has more varied applications than either of the other two problems.

A key concept in these studies is that a home is not a single consumable good, but a bundle of consumable characteristics (bedrooms, baths, fireplaces, etc.). This concept lends itself to the hedonic pricing techniques described by Rosen (1974). By letting the price of the house, P, be a function of housing characteristics, we are able to estimate market shadow prices for the given attributes.

In an early work, Schnare and Struyk (1976) look at housing in the Boston suburban market to try to determine if housing submarkets should be incorporated into models and how they affect predictive capacity. They argue that the housing market cannot be thought of as perfectly functioning market. In a perfectly clearing market, an attribute's shadow price

would equalize throughout the market due to buyer mobility and substitutable attributes. They point out, however, that "the degree of substitution requisite to the intrametropolitan equality of attribute prices is unlikely to occur in any give market at any point in time" (pg. 148). This could occur because of inelastic demand or supply of housing services and implies a segmented market with each market having different attribute prices and perhaps an altogether different hedonic equation. They point out that this segmentation could occur due to neighborhood characteristics like school quality, or be based on structural characteristics of the homes. Schnare and Struyk then experiment with different techniques to identify and test for this segmentation. They make a point to note that in testing for segmentation, it is not sufficient to only test for different attribute prices between the submarkets. One must also evaluate that difference against the overall variation in the home price. If the variation is small relative to total price variation, then this would not constitute enough evidence to suggest a segmented market. They stratified the sample into submarkets using the number of rooms, average census tract income, and distance to central business districts. In a separate test, they also segmented by political boundaries (city limits of three towns). They found that while attribute prices did differ across submarkets, they had a small overall effect on the total cost of the home. Furthermore, they found the predictive power of the unstratified (full-market) hedonic equation was higher than that of the stratified model.

Goodman (1981) was an early attack at the problem of submarket identification. Like Schnare and Struyk (1976), he theorizes that submarkets form due to inefficacies in market equalization like search costs or racial discrimination. These inefficiencies lead to short-run differences in the hedonic parameters. In order to start his submarket identification, Goodman assumes that all houses within a given municipality should be grouped together due to public service offerings of the municipality. Submarkets are then either contiguous or non-contiguous clusters of municipalities. He also lays out three criteria for submarkets: a grouping with fewer submarkets is superior to one with many submarkets, houses within a submarket should be as similar as possible, and optionally that contiguous municipalities

should be grouped together. Two indexes are then given as a way to measure the simplicity of the submarket structure and the homogeneity of the submarkets. The summation of these two indexes is then maximized through an algorithm to obtain the optimum submarket configuration. By constructing these submarkets, Goodman finds evidence for geographic market segmentation in his data.

In an effort to more fully express the theory that submarkets arise from disequilibrium in the housing market, Maclennan and Tu (1996) work through many mechanisms that might lead to inefficiencies. The three largest factors that contribute to this disequilibrium are the variety of home characteristics, space, and time. If attributes of homes are indivisible (design style for example) or not easily replicated, substitutes may be difficult to obtain and lead to disequilibrium. Space factors into the discussion because buyers face search costs when looking for a new home. Due to the large amount of listings in a metropolitan area, searches may be limited to certain areas or become hierarchical in nature. Also because of a sizable supply-side lag in offering new housing stock, the market will tend to be in disequilibrium most, if not all, of the time. In an effort to identify submarkets Maclennan and Tu perform factor analysis on key housing characteristics in order to cluster homes into "product groups". These product groups are then clustered into submarkets through hedonic analysis.

Goodman and Thibodeau (1998) seek to contribute to the submarket literature by introducing hierarchical submarket modeling. They criticize previous papers for imposing a submarket structure onto the data based on prior assumptions. They suggest there may be neighborhood, school district, and municipality effects, each which could be nested under the next to create a submarket. A level-one model is given such that the submarket home prices are a function of structural attributes. A second level model of the coefficients is then given based on fixed effects. These two models are combined and the parameters estimated through Expectation-Maximum likelihood. Applying this model to their dataset, Goodman and Thibodeau estimate a level-one model using only the home's square footage (in log form) and its age. The level-two model uses the school quality of home's assigned school as

a predictor for the coefficient on square footage under the assumption that school quality is capitalized into property size. Submarkets are then generated by estimating the model for two school zones and determining if the school quality has an effect on the beta coefficient of square footage. If it does, then the two zones are seen as different submarkets. New zones are added individually and this is repeated until all zones are assigned to a submarket. Using this method, they segment Dallas County into 5 distinct submarkets.

In 2003, Goodman and Thibodeau returned to this topic to test the predictive accuracy of their hierarchical models. They point out that pooling housing data may not be detrimental if a researcher wants to assess an entire metropolitan area. However, if the purpose is to assess individual homes for, say tax assessment or appraisal, then not taking submarkets into account could have a critical impact. In order to test predictive capabilities of their procedure, Goodman and Thibodeau also estimate models using the following submarket specifications: no submarkets (pooled), zip codes, and census tracts. The hedonic equations are generated using 90 percent of the dataset while the remaining 10 percent are withheld for out-of-sample testing. Their results showed that while all three submarket definitions easily dominated the pooled data, none of the three submarket definitions definitively dominated the others.

Goodman and Thibodeau (2007) look to determine if housing submarkets must be contiguous. At first glance it seems intuitive that homes near each other would be similar due to shared neighborhood services. Forcing submarkets to be contiguous also reduces the possible number submarkets, making optimization routines converge more quickly. However, Goodman and Thibodeau point out that similar buyers may locate in different areas of a metropolitan area yet those neighborhoods would be very similar. They define a submarket as “an area where the per-unit price of housing is constant." (pg. 210) So if homes in these two neighborhoods are priced similarly (the hedonic coefficients are the same) then they should be grouped into the same submarket. The geographically contiguous submarkets were constructed from adjacent census block groups within a school district until each had about

120 transactions. This resulted in 372 submarkets. The non-contiguous submarkets were generated by stratifying the dataset by per-square-foot transaction price and then further segmenting each percentile by living area. This resulted in 325 submarkets. They find that the best method depends on the criteria used to score. The spatially adjacent submarkets had a higher amount of estimated values within 10 percent of the actual transaction price. However, it also had a significantly higher mean-squared prediction error.

Goodman and Dubin (1990) propose a way to test non-nested stratifications. Goodman (1981) tried to get around this problem by generating two indexes to balance simplicity and similarity. However, these indexes and their weights were very arbitrary and therefore unreliable. In an effort to alleviate the problem the Cox, J, JA, and non-nested F tests are discussed. They then apply these tests to submarkets in a hedonic model. They give examples of how to use each test to determine whether the stratification is significant and suggest the J test as an easy to calculate solution.

Bourassa *et al.* (1999) strives to optimally identify submarkets using a combination of factor analysis and clustering algorithms. They begin by emphasizing the role substitutability plays in the generation of housing submarkets. Optimum submarkets should have "a maximum degree of internal homogeneity and external heterogeneity." (pg. 161) Like others, they criticize the use of predetermined submarket boundaries on the idea they have very little chance of meeting this criterion. In order to generate a set of submarkets to meet their requirements, Bourassa *et al.* first performed factor analysis on their dataset. This reduced set of factors is then used to cluster the homes using the k-means algorithm. These clusters are considered as submarkets and hedonic equations are estimated for each. The weighted mean squared errors are then compared to other estimated equations which used different submarket definitions. They found their identification strategy yielded much better results in some cases than other submarket definitions and that the segmented models always performed better than pooled models.

Like Goodman and Thibodeau (2003), Bourassa *et al.* (2003) seeks to identify the submarket definition that offers the best predictive value. They test the submarket identifi-

cation method put forth in Bourassa *et al.* (1999) against a set of "sales groups" which are defined by local government appraisers in New Zealand. Like their previous paper (Bourassa *et al.*, 1999), they focus on substitutability as being a major contributor to the formation of a submarket. They acknowledge, however, that statistically generated submarkets may not always exhibit internal substitutability: "It is possible that quite different dwellings could have similar hedonic functions yet not be substitutes. That is because the hedonic method focuses on the prices of characteristics rather than the existence or quantity of those characteristics." (pg. 14) They point out, however, that if prediction is the goal of the researcher, than this is not an issue. To test the two definitions, 80 percent of their dataset was used to generate hedonic equations and the remaining 20 percent were used as out-of-sample observations. They use two methods of controlling for the submarkets: submarket dummy variables and separate equations for each submarket. The measure of how well each method-definition pair performed is the percentage of estimated values that fall within 10 and 20 percent of the actual sales price. Their results show that sales group dummy variables dominated all other method-definition pairs in all but one sample, and in that one sample separate sales group equations dominated. This implies that small geographical areas may tend to have a higher predictive power than statistically defined submarkets.

As a follow-up to the two previous papers, Bourassa *et al.* (2007) expanded the submarket definitions tested and investigated which offered the highest predictive power. In this paper, a total of 16 models are tested against each other: four geostatistical models, two lattice models, and two OLS models (then a set of submarket dummy variables are added to each model specification, doubling the total number). Lattice models (SAR and CAR) use a weighting matrix to specify spatial dependence in the error terms of each home. Geostatistical models use variograms to model the covariance matrix. The four geostatistical models each use a different definition for the semivariogram. In order to test predictive power, 100 random samples of 80 percent of the dataset were drawn and used to estimate each model. The remaining 20 percent of each sample was used for out-of-sample predictions. Unfortu-

nately, lattice models do not easily lend themselves to out-of-sample predictions due to being unable to modify the weighting matrix. This caused both lattice methods to perform poorly in relation to the other submarket definitions. Bourassa *et al.* found that incorporating simple submarket dummy variables into all models improved their performance. The best model was found to be a geostatistical model with submarket dummy variables. However, an OLS model with dummy variables performed better than all models without the dummy variables. They conclude that the small predictive benefit gained with complex statistical techniques might be overcome by the benefit of simplicity in OLS regressions with dummy variables.

Tu *et al.* (2007) also tries to statistically generate submarkets using clustering techniques. They begin with a delineation of the difference between a neighborhood and a submarket. A neighborhood is "a bundle of spatially based attributes associated with clusters of residences." A submarket, however, is a collection of these neighborhoods. This speaks to Bourassa *et al.*'s claim that hedonic methods ignore substitutability and focus solely on prices of characteristics. Reworded in Tu *et al.*'s nomenclature, the hedonic method identifies submarkets, but not neighborhoods. In order to generate their submarkets, they use a geostatistical technique (like Bourassa *et al.*, 2007) to estimate a semivariogram and construct a covariance matrix. This matrix is then used to cluster the homes into neighborhoods. These neighborhoods are then aggregated using Chow tests to define submarkets with constant hedonic prices.

Leishman (2009) uses a multi-level hedonic model to identify submarkets within a dataset. He first uses factor analysis to reduce multicollinearity within his explanatory variables. He then estimates a multi-level random intercept hedonic model. The data zones are then clustered using the random intercept to form a submarket. This method was applied to the Glasgow metropolitan area in two separate time periods (2002 and 2006) and resulted in very similar though not identical submarkets mappings.

Case *et al.* (2004) reports the results from a competition which initially was individual but eventually became a cooperative effort between four researchers. Each was given a

data set from which 20 percent of the observations had been withheld for out-of-sample predictions. The researchers then submitted their predicted transaction price which was measured against the actual price. Case's model estimated a hedonic regression for each census tract in the dataset. The parameters were then used as inputs to a k-means clustering algorithm. The clustering was repeated 10,000 times (as k-means is sensitive to initial cluster allocation) and a frequency matrix was generated based on how often tract i was clustered with tract j. This frequency is then regressed on other variables like distance between tracts, tract demographics, etc. The estimated value from this regression is then used to actually cluster the tracts into submarkets. Case also included as a second stage estimation the residuals from the nearest five neighbor homes. Case's round one results had the lowest standard deviation of prediction error by far. He was able to further increase the effectiveness of his model by increasing to nine nearest neighbors and correcting for transformation bias.

In order to identify submarkets in the most general way possible, Kauko (2004) turns to neural network techniques. These neural network techniques are in contrast to other clustering techniques based on hedonic coefficients. A hedonic model requires an assumed functional form while a neural network is non-parametric. A neural network is "a sophisticated statistical method that captures non-linear, but regular associations (patterns) within a data-set without specifying much of the model beforehand." (pg. 2561) The network maps inputs (housing characteristics) to an output grid. By examining how the inputs map to the grid, a visual representation of the clusters can be seen.

C.2 Methodology

In this paper, we define a housing submarket as a group of homes who share a common hedonic regression equation. That is, the hedonic coefficients are constant within a submarket but vary between submarkets. Why might housing submarkets arise? Some researchers (Schnare and Struyk, 1976; Goodman, 1981) argue that submarkets could form due to inelastic demand and supply for housing. For instance, Schnare and Struyk (1976) gives an example where some consumers have relatively inelastic demand for good schools.

This leads to a two-tiered search process: first finding a quality school and then finding a suitable home within its attendance zone. This would lead to different hedonic prices between zones if housing supply was also inelastic within the zones due to previous development and high costs of home modification. Watkins (2001) extends this argument by referring to 'consumer groups', consumers with similar tastes, lifestyle, socioeconomic status, etc., and 'product groups', groups of homes that are viewed as close substitutes by consumers due to housing and locational characteristics. The matching of these consumer and product groups could then give rise to housing submarkets.

If submarkets do exist, then the hedonic model for housing prices in an area is then

$$P_i = \beta_0^j + \beta_1^j x_{i1} + \ldots + \beta_p^j x_{ip} + \epsilon_i \text{ where } i = 1, \ldots, N; j = 1, \ldots, k \tag{C.1}$$

Here k is the number of submarkets. So house i has the coefficients $\beta^j = (\beta_0^j, \beta_1^j, \ldots, \beta_p^j)$ if house i is in submarket j. If we ignore the possibility of housing submarkets, we are imposing a restriction on (C.1) so that it becomes

$$P_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i \text{ where } i = 1, \ldots, N;$$

The j equations of the form in (C.1) can be written in one single equation as

$$P_i = \sum_{h=0}^{p} \beta_{hi}^1 x_{hi} D_1 + \sum_{h=0}^{p} \beta_{hi}^2 x_{hi} D_2 + \ldots + \sum_{h=0}^{p} \beta_{hi}^k x_{hi} D_k + \epsilon_i \tag{C.2}$$

where D_j is equal to one if house i is in submarket j and zero otherwise. It is the model in (C.2) we wish to estimate. Unfortunately, the researcher usually has no prior knowledge of k nor does he know to which submarket each house belongs. Therefore the researcher will likely be estimating misspecifications of (C.2). How can one compare the different estimated models in a search for the true model? Sin and White (1996) show that under some general assumptions, an information criterion like the one proposed by Akaike (1974) can be used to choose between two possibly misspecified models. Comparing the Akaike information

criterion (AIC) of all models would allow the researcher to find the model with the highest AIC which Sin and White show asymptotically corresponds to the true model. The effort then is placed on maximizing the AIC across all possible model specifications.

This is no easy task. If there are n homes and k submarkets, the number of possible submarket partitions is given by the Sterling number of the second kind:

$$\left\{ {n \atop k} \right\} = \frac{1}{k!} \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j}$$

For even moderately large values of n, the solution becomes prohibitively large for modern desktop computers[1] . It would therefore be impossible to programmatically try each possible partition due to time constraints. A more efficient method must be used.

Researchers have used many different algorithms in order to cluster homes into submarkets. Bourassa *et al.* (2003) and Case *et al.* (2004) both use the k-means clustering algorithm to identify submarkets. The k-means algorithm seeks to minimize the Euclidean distance between each data point and k centroids. The data are initially assigned to randomly generated centroids. Subsequent iterations continually redefine the centroid as the center of the data points assigned to it and reassign the data to the closest centroid. This repeats until no changes are made in assignments and location of the centroids. K-means has several benefits. First, it is relatively quick, especially when combined with factor analysis to reduce the dimensionality of the data being studied. It is also easily implemented in modern software packages and programming languages. K-means does however have drawbacks. Its results are sensitive to initial cluster assignment and therefore the algorithm is usually run multiple times in an effort to find the optimum clustering. Also, the optimum clustering is determined by minimizing the sum of the Euclidean distances between data points and their assigned centroid (referred to as the distortion). Yet, homes could have very different characteristics yet still be within the same submarket given our definition above. There is then no theoretical justification for using k-means to identify housing submarkets. Finally,

[1] $\left\{ {50 \atop 5} \right\} \approx 7.4\text{x}10^{32}$, and $\left\{ {4000 \atop 5} \right\} \approx 6.3\text{x}10^{2793}$

k-means requires the number of submarkets to be determined before running the algorithm which begs the question of how to determine k. Initially one might think to run the k-means algorithm for several values of k and choose the one with the smallest distortion. However, the algorithm produces results which are monotonically decreasing in k. That is, using the above criteria, we would find the optimum number of clusters to be equal to n because the distortion would be zero.

C.2.1 Genetic Algorithm

In an effort to solve this pooling problem within a panel dataset, Kapetanios (2006) utilizes a genetic algorithm to find the optimal partition. Other economic applications of genetic algorithms can be found in Dorsey and Mayer (1995). A genetic algorithm uses a process akin to evolutionary biology to maximize an objective function without resorting to a point-to-point search of the domain. This is done through a four step process: fitness calculation, drawing, crossover, and mutation.

Assume we have an objective function of p parameters. The algorithm initially randomly generates m strings (where m must be even) of length p. These m strings are referred to as the population. The fitness of each string is the then calculated. The fitness of string m_i is simply the objective function that is being maximized evaluated at the parameters contained in string m_i. These fitness values are then normalized into the range [0,1] and such that the sum of all the fitness values sum to one. These normalized fitness values are then used as weights for drawing a new generation of m strings from the initial generation's m strings. This weighted drawing process gives the fittest strings (ones with the highest fitness value) a higher probability of moving to the next generation. The strings then go through the crossover process. Each string is randomly paired with one of the other newly drawn strings. Then a cut index, i, is randomly chosen for each pair where the index is between 2 and p. The ends of the strings are then switched from index i forward. For example, if our two strings were [xxxxx] and [yyyyy] (so that $p = 5$) and we randomly selected $i = 4$, then the two new strings generated from the crossover step would be [xxxyy] and [yyyxx].

Finally, these new strings go through a process of mutation. Using a predefined probability of mutation, each character of each string is evaluated and, if determined to be mutated, randomly assigned a new value. These new strings are then evaluated for fitness and the process repeats. The algorithm finally stops when there has been no or little improvement in the highest fitness level for some predetermined number of generations.

In order to use a genetic algorithm, one must have an objective function to maximize. We use the AIC evaluated for the partition Γ where Γ has length n and $\Gamma_i = j$ if $\beta_i = \beta_j$ and $j = 1, ..., k$. The information criterion has the general form

$$IC(\Gamma) = 2L(\Gamma) - C(\Gamma)$$

where $L(\Gamma)$ is the log-likelihood of the model with partition Γ and $C(\Gamma)$ is a penalty term. Specifically, the AIC has the form

$$L(\Gamma) = -\ln\left(\frac{\sum_{i=1}^{k} \acute{e}_i e_i}{n}\right)$$

and

$$C(\Gamma) = \frac{2kp}{n}$$

where $\acute{e}_i e_i$ is the sum of the squared residuals for the ith cluster, p is the number of regressors in the model, and k is the number of clusters. For a given k and p, this is essentially choosing the k clusters to minimize the sum of the squared residuals for the model. While this technique requires us to specify k prior to running the algorithm, we also note that it is possible to rerun the algorithm for different values of k and use the overall partition that returns the highest AIC.

C.2.2 Initial Results

Because the housing data used in the school quality section of this paper has 3728 observations, I generated a dataset of 4000 test observations in order to evaluate k-means

Table 73: K-Means Algorithm Results

		K-Means Cluster Assignments					
		A	B	C	D	E	Total
	1	181	167	79	194	179	800
Actual	2	193	173	254	0	180	800
Cluster	3	190	171	256	0	183	800
Assignments	4	181	154	203	78	184	800
	5	104	146	0	445	105	800
	Total	849	811	792	717	831	

and genetic algorithms for finding submarkets. I assigned all observations to one of five clusters (800 in each). I then randomly generated $x_1, x_2, \text{and} x_3$ values for each observation. $\beta_1, \beta_2, \text{and} \beta_3$ were randomly generated for each cluster. I then calculated the y-values for each observation using the cluster beta parameters, the independent x-values and an error term. I first used the k-means algorithm in an effort to cluster the data. Table 73 gives the results from that experiment.

Ideally, a perfect algorithm would generate a table with 5 non-zero cells, each having a value of 800. In this case however, k-means did a fairly poor job of identifying the correct clusters. Cluster D was the best of the 4 with most of the assignments coming from original cluster 5. However, all the other assignments seem to be evenly spread amongst all the original clusters. This is expected however because k-means is not minimizing the AIC as put forth in the model section. It is minimizing the within cluster sum of squares (WCSS) which is defined as the sum of the squared Euclidean distance from each cluster's centroid.

This is now contrasted with the genetic algorithm which minimizes the AIC. Its results are in Table 74.

The difference is striking. The genetic algorithm correctly clustered 97.7 percent of all the observations. It is necessary, however, to discuss the major downside of this approach: computing time. The code producing the above results took approximately 3 days to run whereas the k-means algorithm takes about 8 seconds to run. The computer was processing one generation every 0.13 seconds on average which means the algorithm took roughly 2

Table 74: Genetic Algorithm Results

		GA Cluster Assignments					
		A	B	C	D	E	Total
	1	1	1	0	0	798	800
Actual	2	3	767	29	0	1	800
Cluster	3	3	36	761	0	0	800
Assignments	4	783	8	5	0	4	800
	5	0	0	0	799	1	800
	Total	790	812	795	799	804	

Table 75: GA with K-Means Pre-cluster Results

		Algorithm Cluster Assignments					
		A	B	C	D	E	Total
	1	536	74	5	30	155	800
Actual	2	14	11	0	759	16	800
Cluster	3	11	21	0	760	8	800
Assignments	4	144	254	5	321	76	800
	5	34	18	708	0	40	800
	Total	739	378	718	1870	295	

million generations to reach the optimum. While code optimization might yield somewhat faster results, I do not believe this would result in significant time savings.

However, there must be a middle ground between the fast yet poor performing k-means algorithm and the slow yet highly effective genetic algorithm. I have experimented with "pre-clustering" the data into small 20 observation groups and then using the GA to cluster these groups into larger submarkets. This effectively decreases the string size in the GA which should result in faster runtimes. In order to create the pre-clusters, I used the k-means algorithm to generate 200 clusters from the 4000 observations. Four of these clusters returned as empty so the result was 196 pre-clusters. I then ran the GA on these 196 pre-clusters, grouping them into one of five clusters. The results are in Table 75.

These are admittedly poor results. However, I feel this stems from a poor generation of the pre-clusters. If those initial assignments could be generated more accurately, then this method could show some promise. As it is, this GA took about 7 hours to run with an average time of 1.7 seconds per generation. This increase in per generation runtime

can be attributed to non-optimized code and with some work should be able to be lowered significantly. In order to improve these results, I must experiment with sample stratification techniques other than k-means to generate these pre-cluster groups.

One method that may be theoretically superior to the above technique would be to cluster observations within a set Euclidean distance from each other. This distance would be inversely proportional to the number of observations. So the clustering rule might then be to cluster two observations if

$$\|(p,x)_i - (p,x)_j\| \leq \delta$$

where $\delta = \frac{c}{n}$ and c is some constant.

I have attached the code used for the GA as an appendix. It was written in Python 2.7 and makes heavy use of Numpy n-dimensional arrays. The first several lines of code initialize global parameters which control much of how the GA runs. I've included comments within the code to help where the programming syntax might obfuscate the logic. The algorithm was adapted from a C program written by Dr. Robert Dorsey of FNC, Inc. in Oxford, MS Dorsey and Mayer (1995).

VITA

J. Taylor Smith

Fields

Econometrics and Applied Microeconomics

Education

Louisiana College - Pineville, LA

Bachelor of Science - Mathematics, December 2006

Magna Cum Laude

University of Mississippi - Oxford, MS

Master of Arts - Economics, August 2011

Working Papers

"Measuring the Capitalization of School Quality into Housing Prices: A Spatial Approach". (With Walter J. Mayer, University of Mississippi)

Abstract: Researchers have long studied how school quality capitalizes into local housing prices. Using housing data from FNC, Inc. and standardized test results collected from the Boyertown School District in Pennsylvania for 2004 to 2010, we estimate the impact of school quality on local housing prices while controlling for the endogeneity of school quality measures. Most previous studies have focused on the endogeneity issue. An instrumental variable approach is often used to correct for omitted variable bias, but the instruments are usually fairly weak. Other studies have applied spatial hedonic models of housing prices to the school quality problem. These studies allow endogenous prices but fail to address the

endogeneity of school quality. This paper reconciles these two streams of the literature by specifying a hedonic spatial model that controls for both sources of endogeneity. The spatial model is augmented with a reduced-form equation for school quality that is exploited to improve the asymptotic efficiency of the instrumental variable estimates.

"Do We Know Whether Prices Are Forward Looking?" (With Qiang Zhang, University of Leicester and John Conlon, University of Mississippi)

Abstract: This paper reexamines the debate on the relative importance of forward versus backward looking price-setting behavior in a hybrid Phillips curve model. We first discuss the challenges in identifying the separate effects of expected future inflation as opposed to lagged inflation. We then develop a model where short run fluctuations in flexible-price-sector inflation resemble measurement errors relative to the long-term-contract/core component of inflation modeled in standard New Keynesian Phillips curves. The measurement error perspective suggests that one should be careful to lag instruments, in order to avoid correlations between the measurement errors in lagged, current, and lead overall inflation and the measurement errors in the instruments. When we do this, the estimated hybrid Phillips curve tends to become significantly more backwards looking. To offset the decline in instrument strength caused by lagging instruments for additional periods, we find that expanding the instrument set by including alternative inflation measures, such as the producer price index for finished goods, improves identification of model parameters. The resulting estimates still point to significantly more backward-looking behavior, implying that the share of backward-looking firms may be close to one.