The Pennsylvania State University

The Graduate School

College of Agricultural Sciences

SPATIAL ECONOMETRIC ISSUES IN HEDONIC PROPERTY VALUE MODELS: MODEL CHOICE AND ENDOGENOUS LAND USE

A Thesis in

Agricultural, Environmental and Regional Economics

by

Li Wang

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

December, 2006

The thesis of Li Wang was reviewed and approved* by the following:

Richard C. Ready
Associate Professor of Agricultural and Environmental Economics
Thesis Advisor
Chair of Committee

James S. Shortle
Distinguished Professor of Agricultural and Environmental Economics

Edward N. Coulson
Professor of Economics

Kathryn Brasier
Assistant Professor of Rural Sociology

Stephen M. Smith
Professor of Agricultural and Regional Economics
Head of the Department of Agricultural Economics and Rural Sociology

* Signatures are on file in the Graduate School

ABSTRACT

The dissertation consists of two major components on the topic of environmental valuation using spatial hedonic pricing models.

The first component deals with the specification of a spatial hedonic model. The focus is on how to include the spatial effects of observed house prices into a hedonic model and selecting a suitable spatial model. Several popular spatial models are considered. I argue that, based on theoretical grounds, the spatial error components (SEC) specification provides a better model for house price than the spatial lag model and the SAR error model. I also question the "convention" of row-standardizing the spatial weights matrix in practice and discuss its implications within each of these spatial models. An empirical application is conducted using a large dataset on house sales near three landfills. To illustrate and justify the arguments, I estimate the impacts that these different landfills have on nearby house prices using several spatial models. My claims are supported by the large-sample empirical evidence.

The second component investigates issues related to estimation of the impact of privately-owned and developable open space on nearby house prices. Because privately-owned and developable open space is considered as a part of the residential market, its level responds to area house prices. As a result, an endogeneity problem arises in the hedonic regression model of house price as the house price and surrounding open space are simultaneously determined. The usual approach to deal with this endogeneity is to use IV/2SLS estimation. I propose a new approach that improves on IV/2SLS in two ways. First, in addition to the hedonic modeling of house price, a dynamic process of open space conversion is modeled to include more information, resulting in a simultaneous-equation model. Second, based on the arguments made in the first component, the SEC specification is applied to both equations of the model to incorporate the spatial effects embedded in neighboring house prices and neighboring open spaces. As a result, a nonlinear spatial simultaneous-equation model is suggested to estimating the marginal implicit price of developable open spaces around properties. The estimation results show that the house buyers of the two areas have distinct preference of surrounding land uses of properties. In addition, people only appreciate the amenity effect of open space that is protected from future development.

These two components are related with each other in the sense that the first component establishes a methodology basis for estimating spatial hedonic property value model; the second one comes with an empirical application to modeling spatial effects of neighboring open space (and neighboring house price). Both are to contribute to the application of spatial econometrics in estimating hedonic pricing model.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I’d like to thank my advisor, Dr. Richard C. Ready, for his guidance during the past 2 years and invaluable assistance and comments during the development of my dissertation. I greatly appreciate his patient assistance in revising and correcting the intermediate drafts. Throughout him I have learned how to conduct a thorough research.

I also deeply appreciate other committee members, Dr. James S. Shortle**,** Dr. Edward N. Coulson, and Dr. Kathryn Brasier, who provided valuable suggestions and insights into the thesis. I would also thank Dr. Stephan J. Goetz for his past service in my committee.

My parents and my wife, Xiaochao Deng, deserve special thanks for their encouragement, support and love. It was their love and support that helped me get through the process.

Chapter 1 Introduction

The idea that location is important in determination of property value is not new. However, the spatial dimension of real estate data is not always fully accounted for in traditional hedonic property value models although some location characteristics are included in the model. It is widely observed in real estate data that the sale price of a house at one location is similar to that of a house located nearby for reasons other than those explicitly incorporated in the hedonic model. This is important because the marginal implicit price (MIP) estimates of some environmental variables of interest, such as proximity to undesirable land use, from the hedonic analysis may be biased and/or inefficient. Spatial econometrics is a tool that would remedy this problem. However, different spatial specifications exist. The first part of the thesis explores which spatial specification may be theoretically more appropriate in the context of house price and how sensitive the estimation result is to the choice of spatial model.

One environmental variable of particular interest is open space. The effects on property values of protection of open space have important policy implication. For example, tax assessors have raised concerns regarding the fiscal impact of open space on local government revenue generation. This issue is important since ad valorem property tax is the primary source of revenue for counties, cities and school districts. To explore further this research question, the second part of the thesis employs a simultaneous-equation approach to estimate an individual property owner's willingness-to-pay for a marginal change in the open space surrounding the property based on the methodology of spatial hedonic analysis, in which the endogeneity of developable open space and spatial correlation of neighboring house price are taken into account. In addition to estimate the amenity effect of open space, this study also compares different land uses around house.

This first chapter gives an introduction to the specification of a spatial hedonic regression model and the hedonic analysis of the effect of privately-owned and developable open space on nearby house prices. In section 1.1, the spatial correlation of house prices is examined to motivate the spatial hedonic model, followed by a general discussion of spatial model specification and a brief review on studies of house price using spatial hedonic analysis. Section 1.2 starts with a literature review on hedonic evaluation of open space amenity, and then discusses the endogeneity of developable open space and its implication on model estimation to introduce a spatial simultaneous-equation modeling method adopted in the thesis. Section 1.3 is about the organization of the thesis.

1.1 Spatial hedonic model

1.1.1 Spatial correlation of house price

In real estate economics, hedonic pricing models are commonly used to estimate how certain housing characteristics marginally contribute to house price, for example, the effects of negative externalities generated by undesirable land uses such as landfills, toxic waste sites, incinerators, etc, on house price. These models rely on spatial variation in house price, and typically measure the gradient in house price as distance to the undesirable land use increases. However, spatial correlation of house prices is widely observed in real estate data (high-valued houses tend to cluster together as do low-valued houses). An issue that needs to be explored is how reliable hedonic pricing models are when house prices are spatially correlated.

The sale prices of houses near each other are similar for reasons not explicitly incorporated in a hedonic model. This occurs for several reasons. First, a seller of one house located very close to another similar house that was recently sold may set the selling price according to that of the neighboring house. This would result in a “herd effect”, where price expectations are formed based on neighboring values. Similarly, in cases where the selling prices are set by real estate professionals, recent local house sales, or “comparables”, are given strong weight when the asking price is set. Second,

the conventional “wisdom” among real estate professionals is that the cheapest house on the block appreciates faster than more expensive neighbors so that nearby prices will tend to compress over time. Third, people may receive positive utility from living close to nicer (high-valued) houses, with the result that surrounding high-priced houses push your house price up. The herd effect and price externality lead to the spatial lag model where a spatially lagged dependent variable (spatially weighted neighboring prices) helps to explain the determination of house prices.

In addition to the intuitive motivations for spatial correlation, there are two “technical” sources of spatially correlated house prices that are commonly discussed: common omitted explanatory variable(s) and measurement error(s). These can happen when the externalities in housing market are imperfectly measured or immeasurable. If no attempt is made to include these effects in the model, their entire impact is relegated to the error term, since proximate houses share neighborhood externalities. Spatial correlation in the error term is the result. If an attempt is made to measure and include these effects in the model, the error in the measurement of the externalities, which is similar for proximate houses and manifests in the error term, creates error term correlation. Both omitted variables and measurement error lead to a specification of spatial error model.

The traditional OLS hedonic model does not explicitly take the spatial dimension of housing price data into account, even though the model is “spatial” in the sense that it measures characteristics that vary spatially, e.g., the distance to central business district or distance to an undesirable land use. If house prices are spatially correlated, either in their levels or in the errors, then simple OLS regression can give spurious results. This is important because it means that significant marginal implicit price (MIP) estimate of, say, proximity to a landfill, could be due to spatial correlation of house price, rather than to a real impact of the landfill on house prices. Houses with correlated selling price could happen to locate near a landfill that has no significant externality. Without accounting for the spatial correlation of nearby house prices, its

effect would be manifest through an erroneously significant MIP estimate of the landfill. Even if the landfill does depress nearby house prices, its negative effect spills across houses and may get entangled with the spatial correlation of nearby house prices, resulting in an imprecise or biased MIP estimate. Spatial econometrics explicitly accounts for the influence of space in real estate, urban and regional models.

1.1.2 Specification of a spatial hedonic regression model

An important aspect of the specification of a spatial econometric model is how to incorporate the spatial effect in the model. The two most commonly used spatial econometric models are the spatial lag model and spatial error model (Anselin, 1988). The spatial lag model specifies a spatially lagged dependent variable as an additional explanatory variable for the regression model. In other words, the price of given house is assumed to be determined in part by the price of nearby houses. The stories of herd effect, conventional "wisdom" among real estate professional and price externality fit to the spatial lag specification. Spatial error autocorrelation is either modeled directly from the general principle of spatial statistics or by assuming a particular spatial process for the error term, e.g., spatial autoregressive process or spatial moving average process (Anselin and Bera 1998, Anselin 2001). In spatial error models, the error term in the model for a given house is assumed to be correlated with other error terms for nearby houses. Omitted explanatory variable(s) and measurement error lead to the specification of a spatial error model.

In early spatial hedonic house price modeling literature, Dubin (1988, 1992) refers to geostatistical approach (kriging) to estimating the covariance structure of the model. Can (1990, 1992) applies the spatial lag model with varying coefficients to capture neighborhood effects. Recent years have seen many applications of spatial econometric approaches in hedonic pricing studies in the context of housing markets. Their findings demonstrate the consequences of aspatial specification and suggest that an explicit spatial hedonic specification is beneficial (although not always required for every hedonic analysis).

The two most frequently used models in the spatial hedonic pricing literature are the spatial lag model and the spatial autoregressive (SAR) error model [1]. Using a house dataset in Boston, Pace and Gilley (1997) show that the SAR error model can improve the overall prediction of house price (a 44% reduction of the errors relative to the OLS) by modeling the spatial dependence of the errors. Can and Megbolugbe (1997) investigate spatial dependence and house price index construction. They start their discussion with the often-used hedonic price index model and compare it with a spatial hedonic model. They conclude that incorporating a spatially dependent variable to consider the house price index not only increases the explanatory power of the model, reflected in a higher R^2, but also addresses to some extent the problem of omitted house structure variables. Using a dataset including all arm-length sales (1377 transactions) of single-family houses between January 2000 and May 2001 in the municipality of Stockholm, Sweden, Wilhelmsson (2002) find that there exists some spatial effects in the data and the spatial lag model and the SAR error model explain more of the variation in price than does the OLS. Bowen et al (2001) consider the housing market of Cuyahoga County, Ohio. They find that diagnostic tests call for the explicit modeling of space effect and some drastic differences are found between the space-neglected model and the specified spatial lag model. In their air quality study, Kim et al (2003) report that the spatial lag model is favored over the SAR model and the OLS overestimates the effect of air quality on house price in Seoul, Korea in the presence of spatial lag dependence.

One reason for the popularity of the spatial lag model and the SAR error model is that computer packages, e.g., the Spatial Econometrics Toolbox for Matlab by LeSage and GeoDa by Anselin, contain programming to estimate these two models. While some authors have explored which of these two models might provide a better fit to a particular dataset, there has been less exploration of the sensitivity of externality estimates to the choice of spatial model, and even less reflection on which models might be theoretically more appropriate for hedonic analysis of house price. Another often ignored point in practice is the “conventional” use of a row-standardized spatial

weights matrix. Row-standardization is appealing in part because the associated spatial parameter has a clear interpretation as a measure of spatial dependence (autocorrelation). This makes the spatial parameter comparable between models. However, as Bell and Bockstael (2000) argue, row-standardization changes the assumed spatial structure of the sample data and so the intended "economic" relationship among observations.

The first two chapters of this thesis simultaneously explore both spatial model selection and the issue of row-standardization of the spatial weights matrix, as these two concerns are interdependent. I argue that, based on theoretical grounds, the spatial error components (SEC) model suggested by Kelijian and Robinson provides a better model for house price data than the spatial lag model and the SAR error model. The SEC model is infrequently used in hedonic pricing analyses. I also claim that, for house price, a row-standardized spatial weights matrix is more appropriate for the spatial lag model and may better fit the SEC model than a non-standardized weights matrix; by contrast, there is no strong a priori reason to favor using a row-standardized spatial weights matrix in the SAR error model. To illustrate, the hedonic price model is used to estimate the impacts that three different landfills have on nearby house prices in Berks County, PA. I estimate the implicit price function using a spatial lag model, the SAR error model and the SEC model with both row-standardized and non-standardized weights matrices. My claims are supported by the large-sample empirical evidence.

1.2 Land use and property value

1.2.1 Literature review on hedonic price estimates of open space amenity

The revealed preference approach is widely used to estimate the economic value of nonmarket goods and services. Relying on market transactions and the hedonic pricing methodology, this approach has also been used to estimate the value of open space amenity under the context of housing market. The basic concept in this modeling framework is that a residential property is a heterogeneous good made up of

a bundle of characteristics, each of which contributes to the sale price of the property. These characteristics include environmental attributes of the residential parcel, for example, the amount of open space in the neighborhood of a property.

Open spaces, such as public parks, golf courses and natural areas, can provide numerous amenities for nearby residents including attractive view and recreation opportunities. However, nearby residents may also experience disamenities such as noise and traffic congestion. Many hedonic pricing studies try to examine the net effect of open space on nearby property values.

Open space amenities are often cited as an attractor for neighboring residents and so contribute positively to property value. Several contingent valuation studies have provided evidence of a positive willingness-to-pay for open space amenities. However, the results from hedonic studies are inclusive and mixed. Some studies find a negative effect of surrounding open space on property value, (e.g., Tyrvainen and Miettinan 2000, Smith et al 2002) and the open space effects depend on the size of the neighborhood considered (Geoghegan et al, 1997).

Using housing sale data from the Portland metropolitan area, Oregon, Lutzenhiser and Netusil (2001) consider the effect of specific open space types. They find a positive and statistically significant influence for all the open space types considered except cemeteries. However, some open space types, e.g., urban parks, golf courses, while exerting a positive impact, do not exhibit a stable extent and magnitude of influence up to 1500 feet. For the same area, Bolitzer and Netusil (2000) examine the impact of proximity to open space using one aggregate open space variable and find that, at distances greater than 100 feet and up to 1500 feet from open space, houses are sold for a statistically higher price than houses more than 1500 feet from open space. Nevertheless, within 100 feet of open space the effect is positive but not statistically significant. The authors reason that this anomaly might be the result of both amenity effects and negative externalities from open space present for these houses.

Geoghegan et al (1997) consider the amount of agricultural and forested lands surrounding residential properties in an exurban setting in central Maryland. They conclude that the percentage of open space land in the vicinity of a residential property has a statistical influence on its value and this influence depends on the scale at which the open space variable is measured. Specifically, open space within a tenth of a kilometer radius positively impacts property values, but within a one-kilometer radius this variable negatively impact house prices. The authors interpret this result to suggest that people value open space as a view scene from their house, but at a larger scale, people prefer more diverse land uses, for example, commercial land use. In Berks County, Pennsylvania, Ready and Abdalla (2005) estimate the marginal implicit price for different land uses including open space. They conclude that surrounding land uses have the potential to affect residential property values and find that, within 400 meters of a residential property, open space is the most desirable land use. At greater distances between 400m and 1600m from the house, though, the land use with most positive impact on house price is commercial. From these estimates, the effect of land use conversion, e.g., conversion of farmland (agricultural open space) to commercial use, on property value can be calculated, which has important implications for land use policy.

It is reasonable to assume that the amenity value associated with open space may vary depending on the development potential of the open space. In other words, the buyer of property close to open space might value permanently-protected open space more than open space that can be developed sometime in the future. As a result, open spaces will be distinguished into developable and protected. This distinction introduces an endogeneity problem of open space that can be developed into residential use (developable open space), which will be discussed in detail in the next section.

Focusing on a rapidly developing county in Maryland, Geoghegan (2002) studies the amenity effects of developable versus permanent open space. Using sale data of

residential properties between 1993 and 1996, the author shows that “permanent” open space increases nearby residential property values over three time more than an equivalent amount of developable open space. Similarly, based on the house sale dataset from suburban and exurban counties in Maryland, Irwin (2002) distinguishes open space by land ownership (preserved or developable) and by the land use type. The result shows that the spillover effects from preserved open space are significantly greater than those associated with developable farmland and forest.

Smith et al (2002) analyze the difference between the effects of permanently protected open space such as parks, greenways, golf courses and developable open space such as privately owned vacant land, agricultural and forest land in a suburban area north of Raleigh, North Carolina. They find that private vacant land acts as an open space amenity and properties adjacent to it command a statistically significant higher price than properties further away. Proximity to privately-owned agricultural and forest land, however, negatively impact property value. The authors suggest that the difference between the two results may come from the difference in parcel size. Compared with smaller parcels of vacant land, larger agricultural and forest lands are expected to convert in a more dramatic way. The buyers may fear future changes to land use that would make living nearby less desirable. Contrary to expectation, they find a negative impact of permanently-protected open space on property values. To explain the unexpected result, they searched for possible land uses that would bring negative externalities to property values and thus override the amenity benefits from open space, such as landfills and airport. But no such evidence was found.

1.2.2 Endogenous land use – developable open space

The inconclusive results of open space amenity effects from hedonic price studiess could be attributed to some extent to the specification of open space variables, and/or to the relative scarcity of open space and differences across study areas. However, Irwin and Bockstael (2001) argue that, where privately-owned and developable open space is concerned, a potential identification problem associated with the endogenous

open space arises in estimating the hedonic pricing model. This argument could explain why the positive amenity of open space, when it exists, may not always be detected in empirical hedonic studies.

When the open space considered is privately-owned and can be developed into residential use sometime in the future, the land is part of the residential land market. The economic factors determining the value of residential land play an important role in determining whether open space is converted to residential land. Owners of private open space will base their land use decision (convert or not, convert to which land use and how much) on comparing the payoffs from different land uses, which depends on the expected residential price. As a result, the amount of open space that can be developed into residential use (developable open space) is not just an exogenous variable on which the house price is based, it also responds to area house prices. This simultaneous interaction between house price and surrounding developable open space introduces an endogeneity problem when an open space variable appears in the hedonic property value model. If the open space variable is indeed endogenous to the residential land market, the estimated marginal implicit price for open space will be biased and inconsistent.

Irwin and Bockstael (2001) consider a simplified two-neighbor model to illustrate the interaction effects between the value of residential land and the amount of surrounding developable open space. Given two neighboring parcels A and B, whether parcel A is a open space or residential land depends on its value in residential use, which is a function of the land use of neighboring parcel B because of the spatial spillover effect. In return, parcel B's land use is influenced by its value as a residential land, which is a function of the land use of neighboring parcel A. As a result, the residential value of parcel A is determined by the residential value of parcel B, and the measure of open space around parcel A, which is a function of the residential value of parcel B, is endogenous. In spite of the endogeneity issue, the authors also show that the open space variable can also be correlated with the regression error if the two

hedonic regression errors of neighboring parcels are spatially correlated. Therefore, there are two potential sources of correlation between the open space variable and the error term in hedonic regression analysis. If the spatial correlation of regression errors of neighboring parcels is detected, which could be a result of omitted spatially correlated explanatory variables, it will yield not only biased and inconsistent coefficient estimates on the developable open space variable but also inefficient parameter estimates for other variables.

To address the correlation between open space variable and regression error, an instrumental variable (IV) approach is often adopted. The appropriate instruments for endogenous open space measure need to correlate with the spatial pattern of land use but be exogenous to the residential housing market (i.e. be uncorrelated with the regression error in the hedonic pricing equation), and minimize potential multicollinearity problem. The instruments often employed are (exogenous) features of the land, i.e. Irwin and Bockstael (2001) use the parcel's slope, the soil's drainage ability, whether the parcel has high quality soil – as proxy for the physical cost and opportunity cost of development.

1.2.3 Spatial simultaneous-equation modeling

There are generally three important concerns in using a simple instrumental variable technique to correct for the endogeneity problem of the open space variable in hedonic pricing regression.

The first concern is the availability of suitable instruments. In practice, it's not always easy to find good instruments. Even if the theoretically suitable instruments can be identified, the data are generally not available for at least some of them. Incomplete and/or "insufficient" instruments may not address the endogeneity problem in a satisfactory way. From the estimate result, Irwin (2002) suggests that the simple instrumental variable strategy does not fully resolve the problem. The second concern lies in the spatial correlation of the regression errors. It is necessary to correct for the

spatial correlation to achieve efficient estimates by selecting an appropriate spatial error structure. The last concern comes from the fact that, as house price and surrounding open space are jointly determined, the single hedonic regression is limited in information it utilizes because it does not take into account the information on the dynamic process of the evolution of residential land use pattern. In reality, the level of open space in a specific area can only adjust downward; in other words, the conversion process of open space to other land uses are irreversible. It is therefore important to model such a dynamic adjustment process of open space.

In order to further explore the research question of what is the impact of developable open space on nearby house prices, a simultaneous-equation approach is proposed to model their joint determination, where the set of instruments can be reasonably expanded and the information on dynamics of open space evolution is accounted for. Moreover, a combined spatial error components (SEC) structure for both house price and open space is assumed to include the spatial correlation of regression errors. This spatial simultaneous-equation model is expected to provide a better understanding of the mechanism under which house price and developable open space are interacted.

1.3 Organization of the dissertation

The remainder of the thesis is organized as follows. Chapter 2 discusses the theoretical framework of spatial hedonic pricing model. I first review the three popular spatial model specifications focusing on their theoretical implications for house price. Next I analyze the consequences of row-standardizing spatial weights matrix for different spatial specifications. Finally, I review of protocols for spatial regression model selection, including the LM test, a pseudo-R^2 goodness-of-fit criterion and Bayes factor method. In chapter 3, I estimate several spatial hedonic models using data on house sales in Berks County, PA and explore empirically the issue of model selection.

Based on the conclusions made in the chapter 2 and 3, chapter 4 presents a spatial simultaneous-equation model of joint determination of house price and surrounding developable open space, including model setup, spatial spillover modeling and model identification. Then, a three-stage least square (3SLS) procedure with strict 2SLS is introduced as the estimation methodology to dealing with the endogeneity, nonlinearity and spatial effects in the model. In chapter 5, this model is estimated using a second dataset from Berks County. The data development with GIS is first discussed in detail. Then a two-market model is specified based on the data and knowledge of the study area. The estimation procedure and results are discussed next, including the choice of spatial weights matrix and the analysis and comparison of estimation results. Chapter 6 concludes the dissertation.

Footnote of Chapter 1:

1. There are some differences in how these models are labeled, e.g., Lesage (1999) uses spatial autoregressive (SAR) model to label the spatial lag model and spatial error model (SEM) to label the SAR error model I employ here.

Chapter 2 Theoretical framework of spatial hedonic pricing model

This chapter examines some important theoretical considerations in the specification of a spatial hedonic model. Section 2.1 discusses three popular spatial specifications, the spatial lag model, the SAR error model and the SEC model, in the context of house price. Section 2.2 briefly reviews the spatial weights matrix and discusses the effects of row-standardizing a spatial weights matrix as well as its implication within each spatial model. Section 2.3 describes three empirical methods of spatial model selection, the LM tests, a pseudo-R^2 criterion and the Bayes factor method.

2.1 Review of spatial regression models: theoretical implications

In essence, each spatial model specifies a spatial process under which the observations are generated. Different beliefs on the appropriate spatial relationship (spatial dependence and spatial heterogeneity) lead to different spatial models. The applications of the spatial lag model, the SAR error model, and the SEC model to house price represent three distinct spatial data generating processes (DGP) of house price. De Graaff et al (2001) list three reasons for handling spatial dependence (correlation) and spatial heterogeneity jointly. First, there may be no difference between spatial heterogeneity and spatial dependence in an observational sense. Second, spatial dependence induces a particular form of heteroscedasticity. Third, it may be difficult empirically to separate the two effects. As a result, I will just focus on spatial dependence of house price in subsequent analysis.

2.1.1 Spatial lag model

Anselin (2002) discusses two main motivations for including spatial effects in regression models: one from a theory-driven and the other from a data-driven perspective. A theory-driven framework follows from the formal specification of spatial interaction (e.g., interacting agents, social interaction) in an economic model, where an explicit interest in the spatial interaction of a particular variable prevails and a theoretical ground generates the model specification. In other words, dependence of

spatial process of the particular variable is *substantive*, compared with the so-called "*nuisance*" correlation of spatial error processes.

If we believe that the selling price of a house at a location acts as a signal that guides the selling prices of its neighboring houses, or a herd behavior exists, spatial dependence of house prices can be modeled directly as including a spatially lagged dependent variable into the hedonic pricing model. This model captures the intuitive idea that a house surrounded by expensive houses is worth more than the same house surrounded by inexpensive houses. Then, the neighboring house prices act as an explanatory variable of the house price at a particular location.

As a result, the spatial lag model takes the form

$$\begin{aligned} y &= \rho W y + X\beta + \varepsilon \\ \varepsilon &\sim N(0, \sigma^2 I_n) \end{aligned} \tag{2.1}$$

where y is a $n \times 1$ vector of house prices, W is the pre-specified $n \times n$ matrix of spatial weights which relates the sale price of one house to the sale prices of other houses in the sample by specifying a neighborhood for each house (see section 2.2 for detail), ρ is a spatial autocorrelation coefficient (when W is row-standardized), X is a $n \times k$ matrix of explanatory variables including house structural characteristics, location characteristics and environmental attributes of interest, β is a $k \times 1$ vector of parameters to be estimated, and ε is a $n \times 1$ vector of errors. The assumption of normality for ε is necessary for ML estimation; however, the IV/2SLS estimation does not depend on it (Anselin, 1988, Chapter 7).

Bell and Bockstael (2000) argue that there is less reason to expect a spatial lag structure for house price. It would occur only when the selling price reacts directly to selling prices of neighboring houses and not just to characteristics of neighboring houses. Furthermore, this relationship would only be one-way, namely, the first sold house would affect subsequent sales but not vice versa. This is an important

observation as the observed spatial dependence (correlation) of house prices may not be the result of this "price-bidding" process; its existence could be purely statistical. Nevertheless, the one-way relationship is valid only if it is due to a herd effect. If the price externality is the main reason of spatial correlation, the price of sold house will also be affected by the (expected) prices of incoming houses in the neighborhood, resulting in interacted neighboring house prices. Therefore, the relationship between neighboring house prices needs not to be one-way.

We can rewrite the lag model as $y = (I - \rho W)^{-1} X\beta + (I - \rho W)^{-1}\varepsilon$. Conditioning on X , the covariance structure of y takes the form of $\sigma^2 (I - \rho W)^{-1}(I - \rho W')^{-1}$. Since $(I - \rho W)^{-1}$ is a full matrix with $\rho \neq 0$, a first implication is that the house prices are globally correlated since each location is correlated with every other location in the system, although in a fashion that decays with order of neighbors. More importantly, it says that house price at one location depends on the characteristics (X) of all other houses in the sample and the errors of all other observations. The resulting global correlation implies that, for the same house, as the sample gets larger, its sale price is more affected just because more houses are included in the sample, which does not make much sense. This outcome can be easily seen when W is row-standardized. Footnote 8 of Kelejian and Prucha (1998) allows for the "Leontief expansion" of

$(I - \rho W)^{-1} = I + \rho W + \rho^2 W^2 + \rho^3 W^3 + \ldots$, then we have

$$y = (I + \rho W + \rho^2 W^2 + \rho^3 W^3 + \ldots) X\beta + (I + \rho W + \rho^2 W^2 + \rho^3 W^3 + \ldots)\varepsilon$$

$$= X\beta + \rho W X\beta + \rho^2 W^2 X\beta + \ldots + \varepsilon + \rho W\varepsilon + \rho^2 W^2 \varepsilon + \ldots \qquad (2.2)$$

The spatial lag specification of house price is justified under the real-estate appraisal process using comparable sale prices or by the herd effect of house price externality. It is also the reason why more empirical studies favor the SAR error process over the lag structure.

As another possibility, spatial weights could also be applied to some or all of the regressors in the model, e.g., $y = X\beta + \gamma WX + \varepsilon$. This model implies that, in addition to own house characteristics (X), the sale price of a house also depends on the characteristics (WX) of its neighbors. The parameter γ represents the contextual effect. Although this type of model presents few estimation problem, the argument is not strong enough because if it is the case, it is more reasonable to direct relate the price of a house to its neighboring prices, just as the spatial lag model suggests.

2.1.2 SAR error model

The theory-driven perspective motivating spatial correlation in house prices is appealing. However, in practice, the motivation for applying a spatial econometric model is usually not driven by formal theoretical concerns, but a result of "peculiarities" of the available spatial data. This framework considers outcomes in neighboring places to be related to one another because of unobserved components that are spatially correlated. The interest is to obtain efficient coefficient estimates in hedonic model and make correct inference. Most spatial hedonic pricing models fall into this data-driven framework where the spatial model is used to accommodate a spatially correlated error structure in hedonic regression.

Spatial correlation among regression errors can result from two sources: (erroneously) omitted spatially-correlated variable(s) and measurement error or misspecification of the functional form. The location of a house influences its selling price, and nearby houses will be affected by the same location factors. Since the inclusion of all relevant location attributes is seldom fulfilled and the effects of all omitted variables are subsumed into the error term; if omitted variables are spatially correlated, so are the regression errors. For example, consider the case that there are houses located near an airport but this fact is not included in the hedonic analysis. Since all of these houses are affected by this location factor, when it is omitted, the errors associated with the observations containing the omitted information will be also correlated.

Measurement error may occur when the spatial unit of observation does not coincide with the spatial extent of the economic behavior, and then systematically relates to location. Durbin (1998) provides two main reasons to suspect that neighbor effects are measured with errors: use of proxy variable to account for unobservable neighborhood, e.g., crime rate and socioeconomic characteristics of residents, and mismatch between neighborhood boundaries and the data gathering boundaries. To the extent that neighborhood boundaries differ from the data gathering boundaries, neighborhood variables will be measured with error, with the result that the regression errors will be correlated. An example might be census data that are averaged over a larger area in rural block groups than in urban ones.

The error term can take different forms of spatial structure. The spatial autoregressive (SAR) process is the most popular one [1]. It is similar to the counterpart in a time-series context, although more complex. The SAR error model takes the form

$$\begin{aligned} y &= X\beta + \varepsilon \\ \varepsilon &= \lambda W\varepsilon + u \\ u &\sim N(0, \sigma^2 I_n) \end{aligned} \tag{2.3}$$

where λ is a spatial parameter similar to ρ in (2.1) and all other notations are as previously defined.

This specification says that the error for house i depends on the average of the errors in neighboring observations and its idiosyncratic component u_i, implying that the unobserved errors ε are entirely predictable from neighboring error $W\varepsilon$. Solving for ε as $\varepsilon = (I - \lambda W)^{-1}u$, we have $E(\varepsilon\varepsilon') = \sigma^2(I - \lambda W)^{-1}(I - \lambda W')^{-1}$, which is a full matrix with $\lambda \neq 0$. Thus, every house's error is correlated with other house's error in the system, showing the global correlation of error ε. Each error depends on the idiosyncratic errors of all other places including higher-order neighbors through a distance decay effect. This can also be seen by applying "Leontief expansion" to $(I - \lambda W)^{-1}$ if W is row-standardized. As a result, a shock in the idiosyncratic error u at

any location will be transmitted to all other locations following a spatial multiplier $(I-\lambda W)^{-1}$. As in the spatial lag model, conditional on house characteristics, the house prices are globally correlated.

2.1.3 Spatial error components (SEC) model

In both the spatial lag model and the SAR error model, any error that affects one house's price must also affect neighboring house prices. This is particularly problematic if we consider errors in the explanatory variables. Suppose, for example, that square footage is mis-measured for one house, and that recorded square footage is less than actual square footage (due, possibly, to an addition that was not reported to the assessor's office). That house will sell for more than predicted by the hedonic model, and will have a positive estimated error. Obviously, the measurement error of square footage for one particular house is house-specific, the resulting positive estimated error will not spillover to its neighbors. However, both the spatial lag model and the SAR error model force neighboring house prices to adjust because of that positive estimated error. The SEC model, in contrast, allows for idiosyncratic errors that do not spill over to neighboring houses.

The spatial error components model was originally proposed by Kelejian and Robinson (1993, 1995) to avoid the singularity problem [2] associated with the SAR process of the dependent variable/error term. The SEC model combines both a local error term and a spillover error term in the covariance structure for the error term in regression, taking the form

$$\begin{aligned} y &= X\beta + \varepsilon \\ \varepsilon &= W\phi + u \end{aligned} \tag{2.4}$$

where ϕ is a $n \times 1$ vector of errors that spill over across neighbors, u is a $n \times 1$ vector of house-specific errors that do not spill over. Each element in both ϕ and u is assumed to be iid (or weakly, uncorrelated and id) with mean zero; all other notations are as previously defined.

Note carefully that the difference between this and the SAR error model. Now only one component of the composite error terms has a spatial property. This is the spillover error term ϕ, which will capture spatially correlated omitted variable(s). The location-specific errors *u* capture idiosyncratic unobserved components that are specific to each house, whereas ϕ capture unobserved components that are correlated across nearby houses. This idea is highlighted by the complete error variance-covariance matrix of ε,

$$E(\varepsilon\varepsilon') = \sigma_\phi^2 WW' + \sigma_u^2 I$$

$$\text{or } E(\varepsilon\varepsilon') = \sigma_u^2 (I + \theta WW'),\ \theta = \sigma_\phi^2 / \sigma_u^2, \qquad (2.5)$$

where σ_ϕ^2 is the variance component associated with the spillover error and σ_u^2 is the usual local variance term. As can be easily seen, the only nonzero covariance entries in the variance-covariance matrix are those corresponding nonzero entries in WW'. For first-order W, WW' consists of first and second order neighbors, no higher-order neighbors exist. As a result, as opposed to the spatial lag model and the SAR error model, which induce global spatial correlation, the SEC model yields zero covariance beyond the second-order neighbor and can be considered as a model for local spatial correlation [3].

Of the models considered here, the SEC model best captures our intuition about the process that drives spatial correlation in house prices. A spatial moving average component captures the common unmeasured factors that affect neighboring house prices, while each house has a second error component that captures house-specific unobserved errors. The SEC model explicitly considers the two sources of variation in house price given observed house characteristics. Some errors are “contagious” such that neighboring house prices are observed to be correlated. These errors spill over to neighboring houses via the spatial structure denoted by weights matrix. Other errors are house-specific in the sense that they will not be transmitted and so have no influence on neighboring houses.

2.2 Spatial weights matrix and row-standardization

2.2.1 Spatial weights matrix

As an indirect representation of the covariance structure among spatial objects, the spatial weights matrix (W) is used to relate an observation at one location to the observations in other spatial units in the system by specifying a neighborhood for each observation. Each element in W represents whether the two observations are spatially correlated and how strong the relationship is. The purpose of including a spatial weights matrix is to correct for potential problems due to spatial effect, e.g., inefficient parameter estimates. When considering a weighted sum of neighboring observations on dependent variable (y), we create a spatial lag term Wy weighted by neighbors' proximities to each observation; when considering a weighted sum of neighboring errors (ε), we create a proximity-weighted error term $W\varepsilon$. If we model Wy as an explanatory variable of y in addition to other explanatory variables (X), a spatial lag model is resulted. If we try to model a spatial error process by including the weighted error $W\varepsilon$, it ends up with a spatial error model. Then, the spatial weights matrix represents prior knowledge or beliefs about the underlying spatial structure of the variable of interest (y) or associated error term (ε). If the specification is a good approximation to the reality, it would "correctly" describe the dependence among all neighbors of any given observation.

There are two basic types of spatial weights matrices. The first type is called contiguity-based, which establishes a contiguity relationship based on shared borders or vertices of a lattice. The second type is called distance-based, which establishes a spatial relationship based on the distance between observations and is more pertinent to house sale dataset, where locations of houses are specified as points in space. For both types of spatial weights matrices, the analyst must specify two general parameters before their construction. The first is the spatial extent of the influence or the definition of the neighborhood. For a contiguity-based matrix, if two polygons are contiguous they are considered neighbors. For a distance-based matrix, a critical value

of distance must be specified within which two points are thought to be neighbors. The second parameter is the "power" of influence of two neighbors, which answers the question, does every member of a neighborhood exert an equal influence or do neighbors influence each others to different degrees, depending on the distance between them.

The availability of polygon or lattice data permits constructing contiguity-based spatial weights matrix. There are many applications of such matrix in public finance where spatial relationship is based on jurisdictions that share borders, such as census tracts and blocks, counties and states. Two basic types of contiguity exist: rook contiguity (e.g., two polygon share a common border) and bishop contiguity (e.g., two polygons share a common vertex). Queen contiguity is a combination of these two. Specifically, a contiguity-based spatial weights matrix (W) is typically specified as

$$w_{ij} = \begin{cases} 1 & \text{if i and j are contiguous} \\ 0 & \text{otherwise} \end{cases}$$

where w_{ij} is the ijth element of W.

Distance-based spatial weights matrix is widely used in applications where the data is best represented by points, for example, houses at different locations: the prices/error terms associated with close neighbors are more highly correlated than those of distant neighbors. The relationship in the distance-based spatial weights matrix is typically represented as an inverse function of distance, within the assumed critical value

$$w_{ij} = \begin{cases} 1/(d_{ij})^{\theta} & \text{if } i \neq j \text{ and } d_{ij} \leq m \\ 0 & \text{if } i \neq j \text{ and } d_{ij} > m \text{ or if } i = j \end{cases}$$

The term d_{ij} is the distance between points i and j and usually calculated according to their latitude and longitude (or X,Y planar coordinates). The parameter m is the extent of influence or critical distance value. The choice of its value is an empirical problem that depends on, say, the scale of data and the extent of the perceived neighborhoods.

The θ parameter measures the "power" of influence, whose value represents the distance decay effect within neighborhoods. As θ increases, the influence of nearby observations becomes greater than those further away. An alternative distance-based weights matrix uses linear decay. The weight corresponding to points i and j is assumed to be linearly inverse to the distance between them (d_{ij}) and equal zero at a specified distance. It is the type of weights matrix used in the landfill example in chapter 3 (see equation 3.1 for an exemplification).

The selection of a spatial weights matrix is made prior to running the model, in other words, it is not estimated as part of the model. We need to specify the two general parameters beforehand, e.g., based on past experience or intuition for a given problem. One harsh criticism of spatial econometric models is the use of ad hoc spatial weights matrix. The criticism comes from the lack of empirical justification for any specific type of weights matrix and the observation that small changes in the spatial weights matrix can result in significant changes to the model results.

2.2.2 Row-standardization of the spatial weights matrix

In most empirical studies, the spatial weights matrix is row-standardized to have row-sums of unity producing a spatially weighted average term Wy of dependent variable in spatial lag model or $W\varepsilon$ in spatial error model. The associated spatial parameter then has an intuitive interpretation as a spatial autocorrelation coefficient [4]; it also facilitates the maximum likelihood (ML) estimation of spatial models. As a result, row-standardization has become a convention in practice without further exploration. However, row-standardization of spatial weights matrix assumes a specific spatial structure of the dependent variable or the error term, which may have important implications on estimation results.

2.2.2.1 Number effect and distance effect

A common feature of property sales dataset is that the number of neighbors varies across observations, as does the distance to each neighbors. Some houses have many near neighbors while others have few. For some houses, neighbors are located close by, while other houses have closest neighbors located farther away.

As the number and density of neighbors for each spatial observation are generally not the same, row-standardization rescales each row of the weights matrix by different factors. For a given observation, although not changing the relative dependence among all neighbors, row-standardization does change the total impact of neighbors across observations. It is equivalent to assuming that the total effect of neighbors is constant across all observations no matter how many neighbors an observation has. This is what I label the "number effect" of row-standardization. Consider two houses. For one house, data is available for 20 near neighbors. For the other house, data is available for only 2 neighbors. If the spatial weight is row-standardized, the total effects of neighboring houses are forced to be the same for the two houses. However, common sense tells us that, in this case, the house with 20 near neighbors has more neighboring effects on it than the house with only 2 neighbors.

The second issue to be considered is what I call the "distance effect". For a contiguity-based spatial weights matrix, where each neighbor is assigned the same weight of unity, the influence of any one neighbor for a given observation is inversely proportional to the number of total neighbors after row-standardization. While for a distance-based spatial weights matrix, where the interdependence of spatial observations is assumed to decline with distance, this practice rescales the weights based on the absolute distance to neighbors for each row. Again consider two houses. To exclude the number effect, this time suppose that both houses have the same number of neighbors, say 3. For house i, the distances from the three neighbors are 1, 2 and 4. For house j, the distances from the three neighbors are 2, 4 and 8. Assume an inverse-distance weight as in section 2.2.1 and set $\theta = 1$ and $m > 8$, then the weights

for house i are 1, 1/2 and 1/4 and the weights for house j are 1/2, 1/4 and 1/8, respectively. After row-standardization, the weights for both houses are the same: 4/7, 2/7 and 1/7. As a result, remote neighbors are given the same weights as close neighbors across observations (even though the close neighbors still get larger weights than remote ones for the same observation). In this case, Tobler's first law of geography is violated: remote neighbors are correlated in the same way as closer neighbors and the effect of distance disappears.

For both contiguity-based and distance-based weights matrices, row-standardization tends to alter the assumed spatial structure; nevertheless, the "distortion" of distance effect is much worse for a distanced-based weights matrix than for a contiguity-based one, since distance matters in the former case. A by-product of this practice is that the resulting spatial weights matrix is no longer symmetric, which may complicate the test procedure of spatial models.

Among others, Bell and Bockstael (2000) find far greater sensitivity of parameter estimates to the specification of the spatial weights matrix than the choice of estimation technique. They also point out that this sensitivity, and the concern over the change in spatial structure by row-standardization are related each other. Consequently, caution is needed when row-standardizing the spatial weights matrix.

2.2.2.2 Row-standardization within spatial models

In this subsection, I address the question whether the spatial weights matrix should be row-standardized for the different spatial models listed in section 2.1. I analyze the number effect and distance effect of row-standardization for each model. It can be seen that, although the distance effect argues against row-standardization, the number effect is important in some cases.

Given the spatial lag model for house price, non-standardization keeps the assumed spatial structure where house prices are positively correlated with declining distance (in most cases); however, it also implies that a house with more neighbors in the dataset will attract more of a price premium (if $\rho > 0$) than one with fewer neighbors, even if the two houses share the same structural and location characteristics. So the number of neighboring houses in the dataset matters in determining house price. Further, adding or dropping an observation from the dataset will influence all other prices, because the number of neighbors is changed. This does not make sense in housing markets. The positive impact of high-priced neighbors may increase with more neighbors as the number of signals increases, but there is some limit to this process. Further, the number of neighbors in a dataset is as much a result of when sales occur as it is a result of spatial distribution of houses. In contrast, a spatial lag model with row-standardized spatial weights matrix is not plagued by this problem, since the total effect of neighbors is normalized to be the same for each house (the number effect), although at the price of changing the assumed spatial structure. There exists a tradeoff between the two "bads", 1) pseudo-importance of number of neighboring houses (for non-standardized weights matrix) and 2) over-importance of distant houses when few near neighbors exist (for row-standardized weights matrix). The negative impact of number of neighbors is likely to be a more serious problem and row-standardization is preferred on balance. Therefore, in the empirical work in chapter 3, the "conventional" use of row-standardized weights matrix is adopted in the spatial lag model for house price.

Unlike the lag model in which a house attracts more of a price premium when surrounded by additional (expensive) neighboring houses, in the SAR error model, each error could be positive or negative; more neighbors do not necessarily add to the magnitude of the error at a particular location. Furthermore, a large number of neighboring errors may provide more information about what the error of a particular house price might be. There is less reason to limit the total effect of neighboring

errors to be the same for each observation. Because the number effect is not expected to be as critical in the SAR model, we do not have a strong a priori reason to favor using a row-standardized weights matrix in the SAR error model. In the empirical application in chapter 3, both a row-standardized and a non-standardized weights matrix will be considered for the SAR error model.

Compared with the SAR error model, where the error ε for each observation is a linear combination of all idiosyncratic errors u, including its own, the spillover error component of the SEC model does not go beyond the second-order neighbors for first-order W, resulting in a much more abrupt distance decay effect. The potential "bad" of changing the assumed spatial structure by row-standardization may not be as prominent for the SEC model. On the other hand, it is reasonable to assume that the total spillover effect on price is stable across all houses, favoring the row-standardization. As a result, the row-standardized spatial weights matrix is more suitable for the SEC modeling of house price.

2.3 Spatial regression model selection

I have argued that, on theoretical grounds, the SEC error model may better fit hedonic house price data than the spatial lag or the SAR error model. However, there exist empirical methods for choosing among these competing models. Three spatial regression model selection methods, the LM tests, a pseudo-R^2 criterion and the Bayes factor method, are described below. None of the model selection methods is decisive. Each may incur some sources of misspecifications, e.g., functional form, non-normality. It is best to consider all three methods to see if they provide consistent guidance. LM tests and Bayes factor have been used previously, while the pseudo-R^2 method is a new approach proposed here. Although the predictive goodness-of-fit idea is not new, the proposed pseudo-R^2 criterion expands this idea with consideration of different spatial structures. As a result, detailed descriptions of this criterion for the three spatial regression models are given.

2.3.1 LM tests

Test statistics based on the Lagrange Multiplier (LM) principle have been developed for detecting the existence of spatial correlation in the regression error and of a spatially lagged dependent variable (Anselin 1988, Anselin et al 1996, Anselin and Bera 1998, Anselin 2001). There are two main types of LM tests, one-directional tests and robust tests. One-directional tests are designed to test a single specification by assuming that the rest of the model is correctly specified. The test statistics are developed for the null hypothesis of $H_0 : \rho = 0$ assuming $\lambda = 0$ for the spatial lag model or $H_0 : \lambda = 0$ assuming $\rho = 0$ for the SAR error model. These tests are not valid even asymptotically when $\lambda \neq 0$ or $\rho \neq 0$ because λ or ρ will be involved in the information matrix if neither is zero. Although it is not possible to develop robust tests in the presence of global misspecification (λ and ρ take values far from zero), Anselin et al (1996) propose LM tests for the spatial lag model and SAR error model which are robust to local misspecification. These robust LM tests are also helpful for spatial model selection.

Anselin (2005, p198-200) provides a decision rule using the results of LM tests and their robust forms that can guide the choice between the spatial lag model and the SAR error model. The basic idea is as follows, if only one of the LM-Lag and LM-Error test statistics is significant, choose the model rejecting the null hypothesis of no spatial correlation; if both are significant, consider their robust forms. If only one robust LM statistic is significant, that model should be chosen; if both robust LM tests are significant, the model with the larger test statistic value is favored. This routine is followed by many empirical studies (e.g., Kim et al, 2003).

Strictly speaking, this decision rule is not theoretically justified, although it may have some power against the worst spatial regression specification. For LM tests where the null hypothesis is no spatial correlation of the OLS residuals, the model selection is

made between the OLS and the spatial lag model/SAR error model. In their robust forms, the null hypotheses are $H_0 : \rho = 0$ in the local presence of $\lambda \neq 0$ for the spatial lag model and $H_0 : \lambda = 0$ in the local presence of $\rho \neq 0$ for the SAR error model. Again, neither test is derived for the direct comparison between the lag model and the SAR error model, although they are robust to possible local misspecification of the alternative, and so more powerful than the non-robust counterparts. If we want to differentiate these two spatial models, the null hypothesis should be the lag model (SAR error model) and the corresponding alternative hypothesis is the SAR error model (lag model). Unfortunately, no such test has been developed [5].

Anselin and Moreno (2003) develop a LM test for the SEC model along with a GMM-based test and two variants of the Kelejian-Robinson test. The LM-SEC statistic performs well in terms of power against the null hypothesis of no spatial correlation in Monte Carlo experiments, even in the non-normality situations. There is no test to distinguish the SEC model from the lag model or the SAR error model, and no robust test for the SEC model.

As the LM test also depends on the specification of spatial weights matrix (since the spatial weights matrix is involved in the test statistic), it may provide us a way to select between row-standardized and non-standardized weights matrices for a given spatial regression specification. The weights matrix with the (most) significant LM statistic is preferred (i.e., the decision rule guiding the choice between the spatial lag model and the SAR error model; see section 3.2.2.1 for detail).

2.3.2 Pseudo-R^2 criterion

The second model selection criterion considered is called here the pseudo-R^2 approach, and is based on prediction goodness-of-fit. Given sample information, the goal is to predict the selling price of a house with known house characteristics. Each spatial model is compared by its performance on prediction. For each spatial model, a

pseudo-R^2 value is calculated as the 1 minus the ratio of the variance of predicted errors over the variance of the dependent variable. First, I calculate the predicted price of each house in the sample as if the observation does not exist, and then pair the "predicted" house prices with actual house prices to calculate the pseudo-R^2 value. A model with higher pseudo-R^2 value is preferred. For the OLS model, this is the standard R^2 statistic. For the spatial models, this statistics is called "pseudo" R^2 because the sum of the predicted errors is not guaranteed to be zero or the average of predicted prices does not need to equal the average of actual prices. However, since the average predicted price is numerically close enough to the average actual price in our case for all three spatial models, I'll still expect this criterion to work well.

Ideally, to calculate the predicted price of the *ith* house, we need to exclude the *ith* observation from the sample and estimate the parameters for the *ith* house. For different house in the sample, I drop corresponding observation and derive corresponding parameter estimate. This procedure has to be repeated n times for a size-n sample with one observation dropped each time. When n is large, as in our case, this procedure is computationally intensive but the improvement upon prediction is expected to be negligible since the sample information for estimating model parameters are almost the same each time (only one observation excluded from a large dataset). As a way out, I will just utilize the parameter estimate from the whole dataset (including the observation on the house whose price is to be "predicted") and calculate the predicted prices of all in-sample houses [6].

Once the "predictions" are available from each spatial model, the calculation of pseudo-R^2 is straightforward. However, it is necessary to be careful about how the "predicted" price is derived under different spatial setups, and which information is incorporated into the prediction, such that we can compare the three spatial models using this criterion.

Given each spatial model, there are two information sets available for spatial prediction. The first set, called the basic information set, in general contains the parameter estimates (i.e., $\widehat{\beta}$, $\widehat{\rho}$ of the spatial lag model, $\widehat{\lambda}$ of the SAR error model or $\widehat{\sigma}_u^2$ and $\widehat{\sigma}_\phi^2$ of the SEC model), observations on house characteristics (i.e., X), and the spatial relationship among houses (i.e., the spatial weights matrix W). Different spatial models differ in the contents of the basic information set, i.e., the basic information set for one spatial model may contain more or less information than another one. This distinction is important as we'll prefer one spatial model with smaller (basic) information set over the other when these two models have close pseudo-R^2 values. The second information set generally consists of all information of the basic set and the observations on all house prices. So the first information set is a subset of the second (expanded) information set. However, we'll see that it is not the case for the spatial lag model because the two information sets of the spatial lag model are actually the same.

Inclusion of observations on all other house prices (y) in the sample should improve prediction precision. Note that, given the basic information set, knowing y is equivalent to knowing the residual estimate from the spatial model, a point that will be clear later. It will be seen that we can apply both information sets to get predictions for the lag model and the SAR error model; however, we cannot use additional information on house price to improve prediction for the SEC model, only the basic information set applies. For easy notation, I label the basic information set as S^1 and the second information set as S^2. I'll specify the contents of each of the information set for all three spatial models when presenting the prediction formulae below.

SEC model

$$\hat{y}_i = X_i \hat{\beta}_{SEC} \tag{2.6}$$

where $\hat{y}_i$ is predicted price of the *ith* house, X_i is a row vector of observations on *ith* house's characteristics, $\hat{\beta}_{SEC}$ is parameter estimate using the SEC model. Therefore, for price prediction of the *ith* house, $S^1_{SEC} = (X_i, \hat{\beta}_{SEC}), i = 1,...,n.$ To see why additional information on y is useless, first note that, the error estimate, $\hat{\varepsilon} = y - X\hat{\beta}_{SEC}$ is known if we know y. From the SEC structure $\varepsilon = W\phi + u$, a simple OLS estimate of ϕ can be obtained as $\hat{\phi} = (W'W)^{-1} X'\hat{\varepsilon}$. Nevertheless, for high-dimensional W (i.e., 7493×7493 in the landfill example of Chapter 3), $(W'W)^{-1}$ is computationally impossible, so we cannot solve for $\hat{\phi}$ (and $\hat{u}$). As a result, we have to rely on the basic information set only. Moreover, as the spatial weights matrix is not included in (2.6), S^1_{SEC} does not contain the information on spatial relationship among houses.

SAR error model

Using the basic information set, we have

$$\hat{y}_i(S^1_{SAR}) = X_i \hat{\beta}_{SAR} \tag{2.7}$$

where $\hat{\beta}_{SAR}$ is the parameter estimate using SAR error model.

Similar to S^1_{SEC}, $S^1_{SAR} = (X_i, \hat{\beta}_{SAR})$. As the SEC model and the SAR error model share the same explanatory variables (they only differ in the spatial error structure specification), we may compare their pseudo-R^2 values to find a model with better goodness-of-fit.

Using additional information on house prices, we end up with

$$\hat{y}_i(S^2_{SAR}) = X_i\hat{\beta}_{SAR} + \hat{\lambda}W_i\hat{\varepsilon} \tag{2.8}$$

where $\hat{\lambda}$ is an estimate of spatial parameter, W_i is the *ith* row of W, $\hat{\varepsilon} = y - X\hat{\beta}_{SAR}$ is the error estimate, which implies that, given S^1_{SAR}, the information on y is equivalent to the information on ε, as just mentioned before. Therefore, we have $S^2_{SAR} = (X_i, \hat{\beta}_{SAR}, \hat{\lambda}, W_i, \hat{\varepsilon})$.

By including the term $\hat{\lambda}W_i\hat{\varepsilon}$, this formula takes into account the effect of neighboring errors on the price prediction of the "target" house. To see why $\hat{\lambda}W_i\hat{\varepsilon}$ is relevant, following the SAR error structure, we have $\hat{\varepsilon}_i = \hat{\lambda}W_i\hat{\varepsilon} + \hat{u}_i$. Since the price of the *ith* house (y_i) is assumed to be unknown, we don't know $\hat{\varepsilon}_i$ and $\hat{u}_i$ (the estimate of the idiosyncratic error of the *ith* house). However, we know the error estimate of all other houses in the sample $\hat{\varepsilon}_{-i} = (\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_{i-1}, \hat{\varepsilon}_{i+1}, \ldots, \hat{\varepsilon}_n)$. As the *ith* element of W_i, $W_{ii} = 0$, $\hat{\lambda}W_i\hat{\varepsilon}$ is a weighted average of neighboring error estimates for the *ith* house (the $\hat{\varepsilon}_i$ in $\hat{\varepsilon}$ is multiplied with $W_{ii} = 0$), which may act as a good error estimate of the *ith* house and should be included into the prediction. So $\hat{y}_i(S^2_{SAR})$ should give better pseudo-R^2 value than $\hat{y}_i(S^1_{SAR})$.

Spatial lag model

Using the basic information set, we have

$$\hat{y}_i(S^1_{LAG}) = (I - \hat{\rho}W)^{-1}_i X\hat{\beta}_{LAG}$$

$$= X_i\hat{\beta}_{LAG} + \hat{\rho}W_i X\hat{\beta}_{LAG} + \hat{\rho}^2(W^2)_i X\hat{\beta}_{LAG} + \ldots\ldots \tag{2.9}$$

where $\hat{\rho}$ is an estimate of spatial parameter, $(W^2)_i$ is the *ith* row of W^2 and $\hat{\beta}_{LAG}$ is the parameter estimate using spatial lag model. Therefore, we

have $S^1_{LAG} = (X, \hat{\beta}_{LAG}, \hat{\rho}, W)$. In equation (2.9), I omit the term $(I - \hat{\rho}W)^{-1}_i \hat{\varepsilon}$ as $\hat{\varepsilon}$ is not available since we don't know y.

Note that, since the "Leontief expansion" does not apply for $(I - \rho W_n)^{-1}$ where W_n is non-standardized, it is not feasible to invert this matrix for large dataset, making the predicted price and pseudo-R^2 value unavailable for non-standardized spatial weights matrix. Furthermore, since the observations (X) on characteristics of all houses (not just the observations on characteristics of the "target" house (X_i)) and the whole weights matrix W (not just the *ith* row of W, W_i) are required to constitute the basic information set for the lag model, the information contained in S^1_{LAG} is much more than those contained in S^1_{SEC} and S^1_{SAR}.

The effect of more information is revealed through the increasing number of explanatory variables ($X, WX, W^2X, ...$) in the lag model. Because only X is involved in (2.6) and (2.7), the information set for the lag model is larger than for the SEC or SAR error model, so that the lag model is expected to yield a higher pseudo-R^2 value. Therefore, given the basic information set for each spatial model, if the calculated pseudo-R^2 of the lag model is higher than those from the other two spatial models, it is not necessarily true that the lag model is superior.

Using additional information on house prices, we end up with

$$\hat{y}_i(S^2_{LAG}) = \hat{\rho} W_i y + X_i \hat{\beta}_{LAG} \qquad (2.10)$$

Now we have $S^2_{LAG} = (X_i, \hat{\beta}_{LAG}, \hat{\rho}, W_i, y)$. As a result, the basic information set of the lag model, S^1_{LAG}, is not a subset of S^2_{LAG}. S^1_{LAG} includes the observations on characteristics of all houses (X) and the whole weights matrix W; S^2_{LAG} only requires

the observations on characteristics of the "target" house (X_i) and the *ith* row of W, (W_i). On the other hand, S^2_{LAG} needs additional information on y. Now that we know y, we may omit the error estimate $\hat{\varepsilon}$ and write $y = \hat{\rho}Wy + X\hat{\beta}_{LAG}$. Substituting $\hat{\rho}Wy + X\hat{\beta}_{LAG}$ into the y in (2.10) and iterating this procedure, it can be easily shown that (2.9) is equivalent to (2.10) since $W_i W = (W^2)_i$, which also implies that $S^1_{LAG} = S^2_{LAG}$. This result is not surprising as (2.9) is just a transformation of $y = \hat{\rho}Wy + X\hat{\beta}_{LAG}$ where $\hat{\varepsilon}$ is omitted. Thus, the resulted pseudo-R^2 values should be numerically close.

An advantage of using the second information set is that, compared with (2.9), (2.10) works for both row-standardized and non-standardized weights matrices given the IV/2SLS estimate. It also provides a way to choose among different spatial weights matrices.

To summarize, there are two information sets available to calculate the predicted house price for the spatial lag model and the SAR error model, only the first (basic) information set can feasibly be used for the SEC model. Thus, we cannot compare the SEC model with the other two models using the pseudo-R^2 criterion based on the second information set. However, as the basic information sets required for the SEC model and the SAR error model are equivalent in content, we can distinguish these two spatial models by using the pseudo-R^2 criterion based on the basic information set. A quick look at S^2_{LAG} and S^2_{SAR} shows that the second information set of the SAR error model is equivalent in content to that of the lag model, suggesting the comparability of their pseudo-R^2 values using the second information set. As the basic information set of the lag model contains much more information than the SAR error model and the SEC model, we cannot differentiate the lag model from the other two models if the lag model yields a higher pseudo-R^2 value based on the basic information set. If

the lag model yields a smaller pseudo-R^2 value, we may safely assert that the SEC model/the SAR error model provides a better predictive goodness-of-fit than the lag model.

2.3.3 Bayes factor method

Though used much less frequently, Bayesian methodology also provides statistics useful in spatial regression model selection when models are not nested. Hepple (2004) develops the formulae for calculating Bayes factors for a family of spatial regression models including the spatial lag model, the SAR error model, the SMA error model, etc. By assuming equal prior probabilities to each model, we can compute the magnitude of the Bayes factor for any two spatial models to select the more probable model. Hepple provides a decision rule for choosing among models, which is illustrated in two small-sample examples ($n = 25$ and $n = 49$) . However, in large sample cases, the Bayes factor method is subject to numerical problems, which will be discussed in section 3.2.2.3.

Given the spatial models to be selected, we want to find the posterior probability of each model being the "true" model conditional on the data, and choose the model with the largest posterior probability. The extent to which the data support one model over another can be measured by the posterior odds of the one model against the other, i.e., the ratio of their posterior probabilities. Since posterior odds = Bayes factor$\times$prior odds, if we assume equal prior probability for all candidate models, we only need to find the Bayes factor for any two candidate models, which is the ratio of their marginal likelihoods (of the data given model). Under the assumptions of normal regression error, non-informative and uniform priors for β and σ^2 , and a uniform distribution with the range $D = 1/k_{\max} - 1/k_{\min}$ for the spatial parameter (where $k_{\max}$ and $k_{\min}$ are the largest and smallest eigenvalues of the spatial weights matrix), Hepple (2004) derives the marginal likelihoods for each of a family of spatial

regression models, from which the Bayes factor can be computed for any pair of candidate models. It should be noted that in each spatial model identical assumptions about the priors for β and σ^2 are made and the X -matrix is always identical. Some decision rule, e.g., Jeffery's "weight of evidence" or Raftery's "grade of evidence", is then used to decide whether the magnitudes of these Bayes factors provide evidence that a model is favored. One additional advantage of this method is that a comparison between row-standardized and non-standardized weights matrices can be made for a given spatial model, as the weights matrix is involved in determining the marginal likelihood. Note that the derived marginal likelihood for each spatial model has two common terms, which can be omitted from the calculation of the Bayes factor. As the marginal likelihood for the SEC model is not included in Hepple's analysis, I derive that formula following the same procedure (see Appendix A).

Footnotes of Chapter 2:

1. Another popular specification is spatial moving average (SMA) process. The model with SMA error sees rare applications although a corresponding LM test has been developed. One reason is that it introduces extra difficulties in ML estimation (Sneek and Rietveld, 1997) and no other widely accepted estimation method is available.

2. The singularity problem refers to the possibility of zero determinant for $|I - \rho W|$ in the spatial lag model or $|I - \lambda W|$ in the SAR error model such that $I - \rho W$ or $I - \lambda W$ is not invertible.

3. The model with SMA error $\varepsilon = \gamma W u + u$ is limited to local spatial correlation. This can be seen from its error covariance structure $E(\varepsilon\varepsilon') = \sigma_u^2[I + \gamma(W + W') + \gamma^2 WW']$, in which the spatial covariance is zero for neighbors higher than second-order, and so is similar to the SEC model. Their subtle but important difference is that in SMA model, the location-specific error component u is considered to be same as those that are spatially correlated. These two specifications are non-nested.

4. See Kelejian and Robinson (1995, p76-77) for discussion about this point.

5. An exception is the two LM tests suggested by Mur (1999) and Trivezg and Mur (2004) to differentiating the SMA error process and the SAR error process. The second test requires calculating the eigenvalues and eigenvectors of the spatial weights matrix, which is difficult or impossible for very large sample size (also see Kelejian and Prucha, 1999, p513-14). The first LM test does not require calculating the eigenvalues (although they show up in the formula of test statistic) but involves highly computation-intensive manipulation of matrices. Neither of the two tests is

feasible for large sample problem.

6. There is an implicit assumption for defining the pseudo-R^2. That is, a stable spatial relationship is assumed to exist across the study area, i.e., the estimated parameters are constant for all spatial observations because the same parameter estimate is applied to “predict” the selling price of each in-sample house, This is also an implied assumption necessary for single-equation spatial econometric modeling although not always explicitly stated. As a contrast, the methodology of geographically weighted regression (GWR) focuses on the spatial heterogeneity of each observation, i.e., each spatial observation is characterized by distinct model parameter.

Chapter 3 Empirical analysis: Landfill externality

3.1 Data and variable description

The empirical application is conducted using a dataset on house sales near three landfills. Whether and how a landfill depresses nearby property values is of interest for many reasons. Several studies have estimated empirical relationships between residential property values and proximity to a landfill or set of landfills using the hedonic pricing technique. Many of these studies have found the existence of a negative externality from a landfill on nearby house prices. Hite et al (2001) combine an urban location choice model with hedonic pricing model to incorporating some spatial effects of disamenity in estimating the impact of landfill on nearby property values. Ready and Abdalla (2005) estimate the house price impact from landfills, among other local disamenities. Although some of the spatial aspects have been included in their studies, the spatial correlation of house prices is not explicitly accounted for. Using the data from the same study area as in Ready and Abdalla, I estimate the implicit price functions for three landfills using the spatial lag model, SAR error model and SEC model with both row-standardized and non-standardized weights matrices and focus on model test and selection.

The data set under study consists of 7493 arm-length house sales between 1998 and 2002 in south and east Berks County, Pennsylvania. Three landfills exist in the study area. Among the three landfills, Pioneer Crossing Landfill (PCL) is prominent on the landscape and directly visible from many nearby residential neighborhoods. Rolling Hills Landfill (RHL) is the largest and visible from at least some vantage points, although surrounding topology shields it from view from most directions. Western Berks Landfill (WBL) is the smallest and is physically isolated from residential areas and difficult to see from off property. The small size and isolation of WBL may prevent it from affecting nearby house price, while the more prominent landfills are expected to have larger house price impacts. For more details on the data acquisition, variable definition and landfill descriptions, see Ready (2004).

Houses included in the dataset are located within 10km of one of the three landfills. Data on sale price (natural log of real house price in 2002 real dollar), sale year, township, location and structural characteristics of the house are obtained. The structural characteristics of the house included in the spatial hedonic regressions are living area, living area squared, lot size, lot size squared, number of bedrooms, number of full bathrooms, number of half bathrooms, dummy variables for existence of a basement, stone exterior, brick exterior, masonry exterior, central air conditioning, whether the property is a single family detached house (as opposed to a duplex, etc), an index of physical condition, age of the structure at the time of sale, and age squared. The location-specific variables consist of distance to commuting waypoints to Allentown and Philadelphia, average slope within 100m of house, elevation, the difference between elevation at the house site and average elevation within 800m of house, average PSSA standardized school test scores, percentage of land within 400m of the house in industrial use, and percentage of land between 400m and 800m from the house in industrial use.

To control for regional differences in the housing market, a dummy variable is defined for each landfill equal to 1 if the house is located within 10 km of the landfill, and 0 otherwise. Based on a preliminary analysis using dummy variables to identify concentric rings around each landfill by Ready (2004), three proximity indices are defined as,

$$\text{PCLmi2} = \begin{cases} d_{PCL} & d_{PCL} < 2 \\ 2 & otherwise \end{cases},$$

$$\text{WBLmi2} = \begin{cases} d_{WBL} & d_{WBL} < 2 \\ 2 & otherwise \end{cases}$$

and

$$\text{RHLmi3} = \begin{cases} d_{RHL} & d_{RHL} < 3 \\ 3 & otherwise \end{cases},$$

where d_{PCL}, d_{WBL} and d_{RHL} are the distance measures (in miles) of the house from the

PCL, WBL and RHL, respectively; the 2-mile threshold represents the spatial limits of PCL's and WBL's impacts on nearby house prices, the 3-mile represents the spatial limit of RHL's impact on nearby house prices. These limits were determined by running an auxiliary OLS regression of house price on three dummy variables of concentric rings of width 1 mile around each landfill (totally 9 such dummy variables) and finding the significant dummy "farthest away" from each landfill. This idea is also used to determine the spatial limit of any impact from industrial land use. By including a measure of industrial land use near the house, the impact of landfill on house price can be estimated separately from that of industrial land. If a landfill does depress nearby house prices, a significantly positive marginal implicit price of the proximity measure for that landfill is expected.

All house price effects are therefore estimated relative to the average house price within 10 km of the landfill. Although the time period is short, it may contain price fluctuations. Four sale year dummy variables are therefore included to capture temporal variation. In addition, township dummy variables are included to represent different levels of public service that would affect house price; they may also capture some spatial aspects of the data.

I want to evaluate people's willingness to pay for living away from the landfills through their purchase decisions. Because of the existence of spatial correlation of house prices, the negative effect of landfills on house price may be hidden by the clusters of similar-valued houses. Alternatively, if similarly-valued houses are clustered near a landfill, a spurious coefficient on landfill proximity could be obtained in an OLS regression. Explicitly adding the spatial ingredient into hedonic model is expected to help recover people's "true" evaluation on the effect of landfill on house price.

3.2 Model selection and estimation

3.2.1 Choice of spatial weights matrix

For each residential parcel sold in the data set, the location of the house is determined as the centroid of the residential parcel with recorded x-coordinate and y-coordinate. A spatial weights matrix is constructed based on the distances between these centroids. It is common to use inverse distance as weights. One consequence is that weights are nonzero even for very remote neighbor. I prefer to choose a distance cutoff beyond which the neighborhood effect is set to be zero [1]. This practice facilitates the use of sparse matrix algorithms in Matlab [2] which provide a computationally efficient approach to solving for estimates and test statistics in models involving a large number of observations. With a distance cutoff, the constructed spatial weights matrix would contain a large proportion of zeros, as most houses are excluded from the neighborhood.

The choice of boundary for the definition of a neighborhood is not clear–cut. Two distance cutoffs (400m and 1600m) are tested. A cutoff distance of 400m yields many "islands" or observations with no neighbors, making the analysis of spatial correlation irrelevant for those houses. A 1600m cutoff produces many more neighbors for each house (average 320 neighbors per house) and just one "island" which is dropped from the sample. Ready and Abadalla (2005) find that nearby land use impacted house prices out to 1600m in the same study area. I use 1600m as the cutoff distance for construction of the spatial weights matrix, but recognize that this choice is somewhat ad hoc, and could have important implications.

Similar to the linear proximity index in Ready and Abadalla (2005, p317), the weights take the form [3]

$$W_{ij} = \begin{cases} 1 - d_{ij} / L & d_{ij} < L \\ 0 & d_{ij} \geq L \setminus i = j \end{cases} \tag{3.1}$$

where d_{ij} is the distance between house i and house j, and L is the cutoff distance

(1600 meter).

3.2.2 Spatial model selection

3.2.2.1 LM tests [4]

The procedure of using LM test statistics for the spatial lag model and the SAR error model follows the decision rule of Anselin (2005). Based on OLS regression errors, LM-Lag test statistics for a spatially lagged dependent variable and for SAR errors were calculated using both row-standardized and non-standardized weights matrices. All four tests show highly significant results and lead to reject the null hypothesis of no spatial correlation in each case (Table 3.1). The LM-SEC test for the SEC model is also highly significant for both weights matrices [5]. For completeness, the Moran's I test is also included.

Since both LM tests for the lag and SAR error structures reject the null hypothesis, we need to consider their robust forms. Before proceeding to next step, a brief look at Table 3.1 shows $\text{LM-Lag}(W_s) > \text{LM-Lag}(W_n)$ and $\text{LM-Error}(W_n) > \text{LM-Error}(W_s)$. The lag structure with row-standardized weights matrix is more significant while the error structure with non-standardized weights matrix is more significant. These results are consistent with the claims made in chapter 2 that a row-standardized weights matrix is more appropriate for the lag model and a non-standardized weights matrix may be more suitable for the SAR error model. We also have $\text{LM-SEC}(W_s) > \text{LM-SEC}(W_n)$, so for this data set, a row-standardized weights matrix appears to better represent the spatial error component structure. However, we cannot compare these test statistics from different spatial models to select the most "significant" one. For model comparison, we use robust LM statistics.

Table 3.1: LM and Moran's I tests for the spatial lag model, the SAR error model and the SEC model

Test Statistics	Spatial Weights Matrices	
	W_s [b]	W_n [c]
LM-Lag	314.37 (0.0000)	27.00 (0.0000)
LM-SAR	825.08 (0.0000)	1704.30 (0.0000)
LM-SEC	734.49 (0.0000)	452.50 (0.0000)
Moran's I	37.48 (0.0000)	72.47 (0.0000)

Notes: [a]. Values shown are test statistics with p-values in parentheses;

[b]. W_s denotes row-standardized weights matrix;

[c]. W_n denotes non-standardized weights matrix.

The robust LM test involves two weights matrices, one for the lag structure and the other for the SAR error structure. There exist four combinations of the two matrices. Referring to Table 3.2, all the eight robust test statistics are highly significant (some are much greater than the non-robust tests). Furthermore, I find Robust LM-Lag $(W_s, W_n) >$ Robust LM-Lag (W_n, W_n) and Robust LM-Lag $(W_s, W_s) >$ Robust LM-Lag (W_n, W_s), which imply that, in the presence of possible locally misspecified SAR error structure, the lag structure with a row-standardized weights matrix is more significant than that with a non–standardized weights matrix. I also find that Robust LM-Error $(W_n, W_n) >$ Robust LM-Error (W_n, W_s) and Robust LM-Error $(W_s, W_n) >$ Robust LM-Error (W_s, W_s), which imply that, in the presence of a possible locally misspecified lag structure, the SAR error structure with a non-standardized weights matrix is more significant than that with a row–standardized weights matrix. As in the non-robust tests, these robust LM tests support the previous claims that a row–standardized weights matrix better captures the lag structure, while a non-

standardized weights matrix better captures the SAR error structure.

Table 3.2: Robust LM tests for the spatial lag model and the SAR error model

Test Statistics	Spatial Weights Matrix Combination [b]			
	(W_s, W_s)	(W_s, W_n)	(W_n, W_s)	(W_n, W_n)
Robust LM-Lag	133.64 (0.0000)	166.56 (0.0000)	27.48 (0.0000)	28.96 (0.0000)
Robust LM-Error	644.36 (0.0000)	1556.50 (0.0000)	825.70 (0.0000)	1706.00 (0.0000)

Note: [a]. Values shown are test statistics with p-values in parentheses;

[b]. The first spatial weights matrix is for lag structure and the second one for SAR error structure, e.g., for (W_s, W_n), the row-standardized weights matrix corresponds to the lag term, the non-standardized weights matrix corresponds to the SAR error process.

Next, I compare the Robust LM-Lag (W_s, W_n) with the Robust LM-Error (W_s, W_n), with the favored weights matrices for each of the spatial structures. The latter (1556.50) is much greater than the former (166.56), indicating that, of these two models, the non-standardized SAR error model best models the spatial structure in our data. As the robust form of LM-SEC in the local misspecification of spatial lag structure is not available, we cannot compare the SEC model with the preferred SAR error model using a robust LM test.

3.2.2.2 Pseudo-R^2 criterion

Table 3.3 shows the pseudo-R^2 values of the SAR error model and the spatial lag model using both basic and enlarged information sets discussed in section 2.3.2 and the pseudo-R^2 value of the SEC model using the basic information set only. For the SAR error model, the pseudo-R^2 values are calculated for both the row-standardized and non-standardized weights matrices. For the lag model, only row-standardized weights matrix is involved since it has shown to be strongly favored over the non-

standardized one in our case. For the SEC model, I calculate the pseudo-R^2 for the row-standardized weights matrix only because the non-standardized weights matrix produces negative spillover error variance (σ_ϕ^2) estimate and so does not fit to our dataset (see footnote 12 of Anselin and Moreno (2003)).

Table 3.3: Pseudo-R^2 values for spatial regression models

Information Set	Spatial Lag [a]		SAR Error [b]		SEC [c]
	W_s	W_n	W_s	W_n	W_s
S^1	0.876	-	0.812	0.815	0.843
S^2	0.875	-	0.857	0.858	-

Notes: [a]. R^2 value based on ML estimation;

[b]. R^2 values based on GM estimation;

[c]. R^2 value based on MM estimation.

The focus is on the first row (basic information set S^1) of the table. First notice that the pseudo-R^2 (0.843) for the SEC (W_s) model is about 6.5% greater than that (0.812) for the SAR (W_s) model, showing that the SEC model outperforms the SAR error at prediction when using the same information set. Secondly, as expected, the spatial lag model yields a higher pseudo-R^2 value than the other two spatial models, implying that we cannot compare the lag model with the SAR error model or the SEC model using the pseudo-R^2 criterion. The reason is, as mentioned in section 2.3.2, the basic information set of the lag model contains more information than the basic information set of the SAR error model or the SEC model. Third, the pseudo-R^2 values in the second row show that, based on the second information set S^2, the spatial lag model achieves a better prediction goodness-of-fit than the SAR error model. Finally, as expected again, the spatial lag model produces almost the same pseudo-R^2 values using both information sets.

3.2.2.3 Bayes factor method

As in the case of the LM statistics, the Bayes factors are very large for our large microlevel sample. Because numerical integration is involved in the calculation of marginal likelihood and the integral over the whole integration range is close to zero, a numerical "problem" arises. Take the SAR error model for example, the marginal likelihood with the two common terms omitted is

$$p(data \mid M_{SAR}) = \frac{1}{D}\int_D \frac{|P|}{|X^{*'}X^{*}|^{1/2} * (s^2)^{(n-k)/2}} d\lambda \tag{3.2}$$

where $P = I - \lambda W$, $X^* = P * X$, $y^* = P * y$, $|\ |$ denotes determinant, s^2 is the residual sum of squares of the regression of y^* on X^*, n is the sample size and k is the number of explanatory variables (including the intercept term).

For the row-standardized weights matrix, the integration interval is $(-1,1)$, $|P|$ reaches its maximum of 1 at $\lambda = 0$. $|X^{*'}X^{*}|^{1/2}$ is decreasing over $(-1,1)$ and reaches its minimum of 3.3643×10^{47} at $\lambda = 1$. The minimum of s^2 is 151.52, then $\min((s^2)^{(n-k)/2}) \to \infty$ with $n = 7493$ and $k = 69$. For the non-standardized weights matrix, the integration interval is $(-0.2602, 0.0046)$ [6], $|P|$ reaches its maximum of 1 at $\lambda = 0$. $|X^{*'}X^{*}|^{1/2}$ reaches its minimum of 6.7317×10^{78} at $\lambda = 0.0046$. As the minimum of s^2 is 156.92, $\min((s^2)^{(n-k)/2}) \to \infty$. In both cases, the values of integrants are very close to zero, but one numerically dominates the other. The similar situation exists for the lag model and SEC model. As a result, the magnitudes of the calculated Bayes factors are extremely large and far beyond the "common" value [7], showing very strong evidence of favoring one model or one spatial weights matrix.

I first derive the Bayes factors of both row-standardized and non-standardized weights matrices for the lag model and the SAR error model. All calculated Bayes factors are presented in Table 3.4. Following Raftery's "grade of evidence" in which the Bayes

factor greater than 150 for model 1 against model 2 indicates strong evidence in favor of model 1, the spatial lag model with row-standardized weights matrix is favored over the one with a non-standardized weights matrix as expected. But the SAR error model also favors a row-standardized weights matrix, the opposite result from the LM test.

Table 3.4: Bayes factors of spatial regression models

Bayes factor	Value
Lag (W_s) / Lag (W_n)	1.5145×10^{25}
SAR (W_s) / SAR (W_n)	7.8269×10^{59}
SAR (W_s) / Lag (W_s)	1.1027×10^{53}
SEC (W_s) / SAR (W_s)	3.4117×10^{124}

Note: The calculation of SEC (W_s) / SAR (W_s) based on $\theta_{\max} = 20$ for SEC (W_s).

Next I calculate the Bayes factor for the SAR error model against the lag model with the same row-standardized weights matrix, resulting in a strong preference for the SAR error model. Last, the comparison between the SEC model and the SAR error model with same row-standardized weights matrix is made, showing that the SEC model is strongly preferred.

3.2.2.4 Summary of spatial model selection results

Even though the Bayes factor method yields surprisingly large statistics, all three model selection criteria results support the SEC model over the lag model and over the SAR error model, and support the use of a row-standardized weights matrix for the lag model and the SEC model. The choice of whether to use a non-standardized weights matrix for the SAR error model is ambiguous.

The LM tests and their robust forms are widely used in selecting among spatial models. The Bayes factor method, as an updated goodness-of-fit criterion, is

potentially useful but suffers from numerical difficulties when dealing with large-sample problems. Advances in this direction, e.g., alternative prior specifications on model parameters, estimation of the marginal likelihood by Markov Chain Monte Carlo (MCMC) technique, are expected. The pseudo-R^2 is more suitable for achieving a better spatial prediction and useful when spatial interpolation is needed to expand a (small) spatial sample.

3.2.3 OLS estimation

The preceding section showed that the SEC model with row-standardized weights matrix is the preferred model. However, it is of interest to explore whether the estimated MIP values are sensitive to model choice. The first column of Table 3.5 shows the OLS estimates of the five environmental variables of interest (all other MIP estimates for the OLS and the spatial models are included in Appendix C). It is found that property values are depressed within 2 miles and 3 miles of two prominent and visible landfills, but not affected by the less-prominent landfill (WBL). Further, the significantly negative MIP estimate of Pind400 and significantly negative MIP estimate of Pind400800 show that house price declines with the percentage measure of industrial land use within 400m of the house, but the impact decreases with distance.

While the three landfill MIP estimates and two industrial land use MIP estimates may be consistent with our expectations, the OLS estimates are biased if the spatial lag model is the "correct" specification or inefficient if a spatial error model is more appropriate. The LM tests and Moran's I test showed that there is spatial structure to prices and/or errors. In the next section, I explore how robust the MIP estimates are to the choice of spatial model.

3.2.4 Spatial regression model estimation

Spatial regression models are not easy to estimate because large microlevel data and corresponding high-dimensional spatial weights matrix are often involved. The maximum likelihood (ML) estimation under the assumption of normal errors is popular in part because estimation routines such as LeSage's are available for the lag model and the SAR error model. These routines require a row-standardized weights matrix. Since we have strong evidence in favor of the row-standardized weights matrix in the spatial lag model, only the estimation result for row-standardized lag model is presented [8]. To compare the row-standardized and non-standardized weights matrices for the SAR error model, we need an estimation method that accommodates both weights matrices. The generalized moments (GM) method suggested by Kelejian and Prucha (1999) is adopted to estimate the spatial parameter (λ), followed by a Cochran-Orcutt-type transformation to account for the estimated error structure for the SAR error model [9]. For the SEC model, a general method of moments (GMM) estimator for the two variance components is first obtained [10], followed by the same feasible GLS procedure as the SAR error model through Cholesky factorization.

Table 3.5 also shows the estimation results for the SAR error model with both types of weights matrix, and the SEC model and the spatial lag model with row-standardized weights matrix. Only the parameter estimates of three landfill proximity indices and two percentage measures of industrial land use of our interest are presented here. Estimated MIPs for all other house characteristics are similar across models (see Appendix B for variable definitions).

Table 3.5: Estimation results of the OLS and spatial regression models – the environmental variables and spatial parameters

Variable	OLS [a]	Lag [b](W_s)	SAR(W_s)	SAR(W_n)	SEC(W_s)
PCLmi2	0.0983 (6.85)[c]	0.0436 (3.03)	0.0257 (0.96)	0.0682 (3.84)	0.0674 (3.33)
WBLmi2	-0.0156 (-1.03)	-0.0459 (-3.08)	-0.0076 (-0.32)	-0.0011 (-0.06)	-0.0254 (-1.28)
RHLmi3	0.0716 (3.94)	0.0618 (3.48)	0.0908 (2.66)	0.0715 (3.96)	0.0807 (3.08)
Pind400	-0.2979 (-6.84)	-0.2508 (-5.88)	-0.3001 (-6.82)	-0.3038 (-6.91)	-0.3019 (-6.94)
Pind400800	-0.118 (-2.86)	-0.0573 (-1.42)	-0.0494 (-1.06)	-0.0707 (-1.61)	-0.0824 (-1.89)
ρ		0.2230 (16.80)			
λ [d]			0.6803	0.0040	
σ_ϕ^2					0.0639 (6.51)
σ_u^2					0.0212 (44.01)
Jarque-Bera [e]	476.54				

Notes: [a]. A White-heteroskedasticity-corrected estimation fails because of resulting short-rank of transformed design matrix X ;

[b]. Estimates are based on ML method and very close to the IV/2SLS estimates; for the same reason above, a 2SLS-Robust estimation by correcting for White-heteroskedasticity is not available either;

[c]. t-statistics in parentheses;

[d]. The GM estimation for SAR error model does not produce the variance estimate of spatial parameter;

[e.] Jarque-Bera test against the null hypothesis of error normality is rejected with high probability.

Note that, for the spatial lag model, the MIPs (marginal benefits) of house attribute i for all houses are $\beta_i (I - \rho W)^{-1}$, which is different from the estimate (β_i) of a spatial error model or traditional linear hedonic model because the lag model includes the induced effects of the characteristic change of all the houses (see Kim et al 2003, p35-36 for details).

Kim et al (2003) show that the "true" MIP of house attribute i is $\frac{\beta_i}{1-\rho}$ for row-standardized weights matrix if a unit change were induced at every location. The β generally underestimate the MIPs of the lag model. However, this story does not fit our situation so well. In their paper where the focus is on measuring the benefits of air quality improvement, a uniform change of air quality at all locations is possible. By contrast, such a uniform change of a landfill proximity index is impractical and is not of interest. Instead, I use an iterative method to find the average MIP estimate of the lag model. I first derive the predicted house prices that would occur if a landfill does not exist. Then I compute the difference between the predicted prices with the landfill and the predicted prices without the landfill. These differences are then regressed on distance to the landfill. The estimated slope serves as an estimate of the average MIP for the landfill. For PCL, the MIP estimate derived in this way is 0.05, which is greater than $\hat{\beta}_{PCL}$ (0.0436) but less than $\frac{\hat{\beta}_{PCL}}{1-\hat{\rho}}$ (0.0561). Note that this estimate is smaller than the estimated MIP from the favored SEC model (0.0674), the non-standardized SAR error model (0.0682) and the simple OLS (0.0983). The same calculation procedure can be used to find the "true" MIP estimate for the RHL. For WBL, the spatial lag model produces a significantly negative estimate, for which I don't have a good explanation, indicating possible model misspecification.

The two estimates of the spatial parameter λ for the SAR error model are not comparable because row-standardization changes the assumed spatial structure of original weights matrix. While the estimate of λ (0.6803) for SAR error (W_s) has an intuitive interpretation as the spatial correlation coefficient, it's hard to give an interpretation to the estimate (0.0040) for non-standardized weights matrix. The same argument holds for the spatial parameter ρ of the lag model. We see some differences in the parameter estimate in the SAR error model between the two weights matrices. For PCL, row-standardized weights matrix gives positive but insignificant estimate, while non-standardized weights matrix produces positive and significant estimate. For RHL, both weights matrices give positive and significant estimates, but the estimate from the non-standardized matrix is 21% smaller than that from row-standardized matrix.

The SEC model has similar estimate results (both in sign and in size) as the non-standardized SAR error model. They both show that PCL and RHL depress nearby house prices and WBL does not. This outcome is consistent with our expectation, as PCL and RHL are prominent and visible while WBL is small and not perceptible by many nearby residents. Specifically, the significantly positive MIP estimates of PCL and RHL from the SEC model show that a house 1 mile further away will increase its expected selling price (in natural log) by 6.74% and 8.07% respectively, while the negative estimated MIP for WBL is not significant. For surrounding land use, the log house price declines 3.02% with 1% industrial land use increase within 400m of the house, but the negative effect disappears out of this limit.

Interestingly, the naïve use of OLS regression provides parameter estimates that are same in sign for the three landfills and two industrial land use measures as the SEC model and the non-standardized SAR error models.

For the two percentage measures of industrial land use, all spatial models yield significantly negative estimates for Pind400, and negative but insignificant estimates for Pind400800. By contrast, the parameter estimates of three landfill proximity indices for different spatial models show a changing pattern. The underlying reason could be that the prices of houses close to industrial land are less spatially correlated than the prices of houses close to landfill. This less prominent spatial relationship makes the specification of spatial model not so critical in estimating the impact of industrial land on nearby house prices.

3.3 Summary

The arguments about spatial model specification and selection are illustrated and justified in this landfill example. Based on the LM test, pseudo-R^2 criterion and Bayes factor method, the SEC (W_s) provides the best estimates of the MIPs for these measures. It is also of interest to compare the results of the other spatial specifications to the SEC (W_s) result. First, all models do pretty well at industrial measures (Pind400m and Pind400800). Second, since a spatial error model appears to be more appropriate than a spatial lag model for our data, the OLS estimates are consistent (although inefficient) and have the same signs and significance as those of the preferred SEC model. Meanwhile, incorrect spatial models (LAG, SAR (W_s)) give misleading results. For example, the lag model overestimates the effect of WBL and SAR (W_s) underestimates the effect of PCL. These results indicate that, although the simple OLS may give spurious MIP estimates because it ignores the spatial correlation of house price, incorrect spatial specification may produce even worse estimate outcomes than OLS.

Footnotes of Chapter 3:

1. Goldsmith (2004) argues that a distance-based weights matrix is not feasible for rural studies as lot size may vary greatly in rural areas, resulting in an uneven number of neighbors from rural clusters or a small number of neighbors for larger lots. The area under our study is a mixture of rural and suburban areas, and the houses located on lots larger than 5 acres are excluded, showing a small variation in lot size, the distance-based weights matrix here therefore does not suffer from this problem.

2. LeSage (1999, p35-42) provides a good introduction and application of sparse matrix algorithm under the context of spatial econometrics.

3. This weight differs from the linear proximity index just by a factor d_{ij}. Another specification of bi-square weights produces similar estimation and test results for all spatial models considered. As have been shown in some studies, the estimation outcomes are not sensitive to the specification of weight function form.

4. LeSage's Spatial Econometrics Toolbox includes the routines of LM-Lag, LM-Error and Moran's I tests for row-standardized weights matrix. However, these routines are not feasible (or at least inefficient) for large-sample problem, as in our case, in that they require many manipulations of matrices, e.g., the inverse. Using Matlab, I develop routines for all LM tests used in this section, which are feasible and efficient for large-sample problem. These routines are available from the author.

5. The very large test statistic values may come from the large sample size. It can be reasonably inferred that, if we split evenly our sample into 10 sub-samples and build corresponding weights matrix, the LM tests for both the lag and error structures might be expected to be significant, say, at 5% level. In a manner similar to the Bonferroni correction of multivariate analysis, then, the "joint" significance level for the whole sample would be approximately $(0.05)^{10}$.

Footnotes of Chapter 3 (continue):

6. This interval is derived the same way as the row-standardized weights matrix, where 0.0046 corresponds to $1/k_{\max}$ ($k_{\max} = 219.667$) and -0.2602 corresponds to $1/k_{\min}$ ($k_{\min} = -3.844$).

7. There are two possible reasons. One is the unavoidable rounding error in numerical integration; the other is the large sample size, which is similar to our explanation of some very large LM test statistics.

8. The IV/2SLS estimation, in which WX acts as an instrument for Wy to avoid the introduced endogeneity problem, can be adopted for the lag model with both row-standardized and non-standardized weights matrices. For the landfill dataset, the ML estimates for the row-standardized lag model are very close to the IV/2SLS estimates. In addition, compared with the ML estimation, the IV/2SLS method allows for a 2SLS-robust estimation by correcting possible heteroskedasticity.

9. Curiously, of the five environmental variables of interest, the ML estimates and GM estimates are different in sign for PCLmi2 and WBLmi2 for the row-standardized SAR error model. This is an estimation issue that should be further explored in future research.

10. See Anselin and Moreno (2003, p600) for details.

Chapter 4 Modeling simultaneous determination of house price and surrounding open space

4.1 A nonlinear spatial simultaneous-equation model

4.1.1 Introduction

In this chapter, a nonlinear spatial simultaneous-equation approach is developed to simultaneously estimate the impact of developable open space on nearby house prices and the impact of local house price on the rate at which open space is developed. As mentioned in section 1.2.2, because developable open space is a part of the residential land market, its level responds to area house prices. An endogeneity problem arises in the hedonic regression model because the house price and surrounding open space are simultaneously determined. A weakness of the IV/2SLS approach is that it assumes that the quantity of open space can adjust upward or downward to reach equilibrium. But, in reality, the level of open space in a specific area can only adjust downward. It is therefore important to model an adjustment process that is dynamic but one-sided. Considering the limitations of simple IV/2SLS method, a new approach is suggested to improve the simple instrumental variable strategy in two ways. First, in addition to the hedonic modeling of house price, a dynamic process of developable open space conversion is modeled to capture historical information on land use, resulting in a simultaneous-equation model. Second, based on the arguments made in chapter 2 and 3, the SEC specification is applied to the errors in both the hedonic pricing and the open space loss equations to incorporate the spatial effects embedded in neighboring house prices and neighboring open space loss. As a result, we end up with a nonlinear spatial simultaneous-equation model to estimating the marginal implicit price of developable open space (and other land uses) around properties.

4.1.2 Model setup

The simultaneous-equation system consists of two equations shown below: one for hedonic price determination and one for developable open space conversion. The open space conversion equation models the loss of developable open space that has

occurred over a specific time period, in our case the 10 years ending in 2005. The hedonic house price equation explains variation in house prices for sales that occur at or near the end of that time period.

$$\ln P_j = \beta_0 + X_j^H \beta_H + X_j^L \beta_L + \beta * OS_j + \varepsilon_j, \varepsilon_j = W_j u + v_j, \ j = 1,....., n \quad (4.1)$$

$$OS_j\% = \alpha_0 + Z_j \alpha_Z + \gamma * \ln \tilde{P}_j + \phi_j, \phi_j = W_j \eta + \xi_j, \ j = 1,....., n \quad (4.2)$$

n : total number of observations

$\ln P_j$: natural log of sale price of house j

$\ln \tilde{P}_j$: "standardized" natural log sale price of house j (predicted price for a house at location j with average structural characteristics)

OS_j : percentage measure of land within 400 meters of house j that is developable open space at the time of the sale

$OS_j\%$: loss rate of surrounding developable open space in 2005 as a percentage of that 10 years before, defined as $1 - \frac{OS_{2005}}{OS_{1995}}$

X_j^H : observation vector of structural characteristics of house j

X_j^L : observation vector of location characteristics of house j, including other types of surrounding land use, such as protected open space, industrial land, etc

Z_j : observation vector of land characteristics that affect the loss of developable open space surrounding house j but not the house price

W_j : jth row of (first-order) spatial weights matrix W

ε_j : jth error term of hedonic regression equation

u : vector of spillover errors across neighboring houses

v_j : location-specific disturbance of house j

ϕ_j : jth error term for the regression equation of developable open space conversion

η : vector of spillover errors across neighboring developable open spaces

ξ_j : location-specific disturbance of surrounding developable open space conversion for house j

β_0 and α_0 are the two intercept terms, β_H , β_L , β , α_Z and γ are corresponding parameter/parameter vectors to be estimated, along with the variances $\sigma_u^2, \sigma_v^2, \sigma_\eta^2, \sigma_\xi^2$ and $\sigma_{u\eta}$.

As is commonly applied in hedonic price equation estimation, a semi-log functional form is assumed for the hedonic pricing regression (natural log of house price as the dependent variable). As the hedonic theory suggests, in equation 4.1, house price (in log form) is a function of structural characteristics and location characteristics of the house. Since developable open space is the variable of interest, it is considered separately from other location characteristics.

For open space modeling, the focus is on the conversion of developable open space to residential land. Area house prices are expected to affect the rate of conversion of developable open space. Areas with higher house prices will tend to lose open space more quickly because more houses will be built to capture the high profits. To measure the rate at which developable open space is being lost in an area, the percent of privately-owned developable open space lost in the 10 years prior to the sale is calculated. The amount of developable open space available 10 years prior is assumed to be exogenous to current price. Although the dependent variable in the open space conversion equation ($OS\%$) has a domain of [0,1], a simple linear regression equation of loss rate of developable open space ($OS\%$) is assumed for three reasons. Firstly, the focus here is not to exactly model the loss rate of open space but to more efficiently estimate the impact of developable open space on nearby house prices by including the information on surrounding land uses; a simple linear specification of

$OS\%$ fits to the end. Secondly, any "asymptotic" distributional assumption on $OS\%$, e.g., logistic distribution, is not suitable because about 17% houses in the sample lose all surrounding developable open space and have $OS\%$ value of 0, resulting in a sample cumulative density function (cdf) of $OS\%$ truncated at $F(0) = 0.17$ on vertical axis. Thirdly, a simple linear regression of $OS\%$ simplifies the 3SLS estimation procedure for the nonlinear two-equation model.

When making land use decisions, owners of developable open space will take into account the parcel's potential value for residential use. To capture relative differences in expected house price related to location, it is necessary to eliminate the effects of differences in house structural characteristics from house price, to reveal the "net" effect of house price on open space loss. A standardization procedure is suggested to eliminate the effects of house structural characteristics on the log house price, with the formula:

$$\ln \tilde{P}_j = \ln P_j - (X_j^H - \overline{X}^H)\hat{\beta}_H \qquad (4.3)$$

where X_j^H is observation vector of structural characteristics of house j, $\overline{X}^H$, a vector of structural characteristics for a "standardized" house, is set equal to the mean vector of structural characteristics of all houses, and $\hat{\beta}_H$ is a consistent parameter-vector estimate of the house structural characteristics in equation 4.1 (which can be obtained by applying 2SLS (GLS) procedure on equation 4.1).

It is assumed that the location-specific house character (X_j^L) affect open space conversion only through their impact on expected standardized house price, so these variables (X_j^L) will not appear as the explanatory variables in equation 4.2. Nevertheless, we still need to incorporate the land characteristics (Z_j) into the right-hand side of equation 4.2 that are thought to affect the conversion of developable open space but not the house price (i.e., Z_j may include zoning, parcel's soil type,

etc). These variables serve the role of instruments in the model.

In the model, $\ln P$ and $OS\%$ are the dependent variables (basic endogenous variables). There also exist two endogenous functions of $\ln P$ and $OS\%$, $\ln \tilde{P}$ in equation 4.2 and OS in equation 4.1. The variable transformation on developable open space and house price introduces nonlinearity into our simultaneous-equation system and thus results in a simultaneous-equation model nonlinear in the (basic) endogenous variables, but still linear in parameters [1]. The nonlinearity here means that at least one endogenous variable appear in the model in two or more linearly independent forms. Although the two transformed variables standardized log house price ($\ln \tilde{P}$) and measure of surrounding open space (OS) are linear function of log house price ($\ln P$) and open space loss rate ($OS\%$) respectively, OS is actually linearly independent of $OS\%$ since the measures of surrounding open space are different across observations. The same situation applies for $\ln \tilde{P}$ (see section 4.3 for detail). Therefore, it is in this sense that the model is nonlinear in the two endogenous variables ($\ln P$ and $OS\%$).

4.1.3 Modeling Spatial Correlation

A new ingredient or modeling improvement of the simultaneous-equation approach is that I explicitly take into account the spatial correlation of neighboring house prices and of neighboring open space conversion by specifying a spatial error component (SEC) structure for each equation.

The (positive) spatial correlation among neighboring house prices has been widely recognized and modeled in the literature. If house j is surrounded by high-price neighboring houses, its price would be likely higher than it otherwise would be (i.e., the systematic component $\beta_0 + X_j^H \beta_H + X_j^L \beta_L + \delta * OS_j$ in the hedonic price model), resulting in a positive estimate of error ε_j. If house j is surrounded by low-price neighboring houses, the estimate of error ε_j is then expected to be negative. The same

story occurs for all its neighbors, so their error estimates tend to share the same sign. The way in which neighboring errors are related to each other may come from a common factor which pushes all neighboring house prices up or down so that the positive/negative error effect on price of one house spills over to all its neighbors. At the same time, each house still keeps its individual specific disturbance on price, whose effect does not to spill over to any neighboring house. Although the common factor may act in a purely statistical way, it is more likely an outcome of omitted variables which matter in determining house price. If such omitted variables exist, their effects on price are included in the error term, and if they are positively related among neighbors, so do the neighboring errors.

As mentioned in section 2.1.3, the SEC structure captures the intuition about the process that drives spatial correlation in house prices: a spatial moving average component captures local unmeasured factors that affect house price, while each house is also subject to a second, house-specific disturbance. Specifically, for each house j, I specify its error as $\varepsilon_j = W_j u + v_j$, where u is the vector of common factors across neighboring houses and each element of u, say u_i, represents the unmeasured factor of house i whose effect spills over to its neighboring houses, so $W_j u$ summarizes the unmeasured spillover effect of all neighboring houses on the price of house j. Positive $W_j u$ may contribute to more loss of open space of house j since it pushes the house price up and so makes the surrounding open space more attractive to residential use. Negative $W_j u$ works the opposite way. Another advantage for adopting the SEC specification is computational. With the SEC structure, the parameters of the error covariance matrix can be easily estimated by general method of moments (GMM) and the specified covariance structure with fewer unknowns decreases the complexity of estimating a simultaneous system.

Compared with spatial correlation of house prices, the spatial correlation of neighboring developable open space (which is defined in the sense that, if two houses are neighbors, their surrounding open spaces are also neighbors) is seldom modeled. There are at least three situations in which open space may be spatially correlated.

Firstly, for neighboring houses, their surrounding developable open spaces will overlap; the spatial patterns of these open spaces tend to be similar. Thus, a spatial correlation among neighboring developable open space appears. Secondly, consider a situation that a new road is built nearby some neighboring houses. The house price is likely to increase by the location amenity from the new road and more houses will be attracted to be built in this area. A result of this process is that these houses will each lose some surrounding open space, so the conversion of neighboring open spaces is observed to be correlated. Even if the road is not actually built but only expected to be built in the foreseeable future, the effect of expectation will come into the house price and increase the value of residential use of this area, resulting in the same result above. This expectation is more relevant to our end since, in practice, we generally don't know such information. It acts as an omitted variable, when included in the error term, making the errors (ϕ) of equation 4.2 spatially correlated. The third explanation stems from the interacting-agents theory which emphasizes the interaction of decision-makers (see Irwin and Bockstael, 2002, for an application on modeling evolution of residential land use pattern). The landowners' decisions on land use are shown to be interacted; the spatial externalities from the neighbors' land uses influence "my" decision on open space conversion. If people find it desirable to live in close proximity to have social benefits of neighbors and attract public and private services to the area, the neighboring houses would be built with less surrounding (developable) open space by this positive externality between developed parcels, resulting in a positively correlated loss of open space. On the other hand, negative spillover effects occur between neighboring development if people are more concerned about congestion and favor open space amenity, then neighboring houses would be built

with large open space around. In either case, the conversion patterns of neighboring open spaces tend to correlate positively.

Therefore, we can model this spatial relationship by the same spatial error component structure as in the hedonic price equation, where η is similar to u in the sense that $W_j\eta$ summarizes the unmeasured spillover effect of neighboring developable open spaces on the developable open space around house j. Furthermore, since house price and surrounding developable open space are jointly determined and so are the u (spillover errors across neighboring houses) and η (spillover errors across neighboring developable open spaces), $W_j u$ (ε_j) is correlated with $W_j\eta$ (ϕ_j) for each house j. Then we have within-equation spatial error correlation for each observation as well as cross-equation spatial error correlation for each house and its surrounding developable open space across observations, which will help specify the error covariance structure of the simultaneous model.

Specifically, by assumption, the spatial errors ε and ϕ have zero means, each vector of errors consists of iid terms, and the components u, v of ε and η, ξ of ϕ are uncorrelated respectively, the resulting covariance matrices for ε and ϕ are $Var(\varepsilon) = \sigma_v^2 I_n + \sigma_u^2 WW'$ and $Var(\phi) = \sigma_\xi^2 I_n + \sigma_\eta^2 WW'$, where I_n is the $n \times n$ identity matrix, σ_v^2 and σ_ξ^2 are the variance components of house-specific disturbance and open space-specific disturbance, σ_u^2 and σ_η^2 are the variance components pertaining to the error spillover effects of the houses and their surrounding open spaces. In order to simplify the relationship between the ε and ϕ, we may reasonably assume that the spillover error components for the house and surrounding developable open space only correlate for the same observation with a common covariance, that is, $E(u_i\eta_i) = \sigma_{u\eta}$ and $E(u_i\eta_j) = 0$, $\forall i \neq j$, then

$\text{cov}(u,\eta) = \sigma_{u\eta} I_n$. As a result, $E(\varepsilon_i \phi_i) = E(W_i u * W_i \eta) = W_i \,\text{cov}(u,\eta) W_i' = \sigma_{u\eta} (WW')_{ii}$, where W_i and W_j are the i th and j th row of W respectively, $(WW')_{ii}$ is the i th diagonal element of WW', or equivalently, $\text{cov}(\varepsilon,\phi) = \sigma_{u\eta} WW'$. Therefore, the combined variance-covariance matrix for our model is shown to be

$$\Omega = Var\begin{pmatrix}\varepsilon \\ \phi\end{pmatrix} = \begin{pmatrix} Var(\varepsilon) & \text{cov}(\varepsilon,\phi) \\ \text{cov}(\varepsilon,\phi) & Var(\phi)\end{pmatrix} = \begin{pmatrix} \sigma_v^2 I_n + \sigma_u^2 WW' & \sigma_{u\eta} WW' \\ \sigma_{u\eta} WW' & \sigma_\xi^2 I_n + \sigma_\eta^2 WW'\end{pmatrix} \tag{4.4}$$

with parameters $\sigma_u^2, \sigma_v^2, \sigma_\eta^2, \sigma_\xi^2$ and $\sigma_{u\eta}$ to be estimated and of $2n \times 2n$ dimension.

A GMM estimator for the variance component parameters ($\sigma_u^2, \sigma_v^2, \sigma_\eta^2, \sigma_\xi^2$) can be obtained from the moment conditions for the diagonal elements of their variance-covariance matrices of $Var(\varepsilon)$ and $Var(\phi)$ (see Anselin and Moreno, 2003, p600 for details). To estimate $\sigma_{u\eta}$, first I find a consistent estimate $\hat{\varepsilon}_i * \hat{\phi}_i$ for $E(\varepsilon_i \phi_i)$, where $\hat{\varepsilon}_i$ and $\hat{\phi}_i$ are the residual estimates of ε_i and ϕ_i obtained from the second stage of 2SLS procedure (see section 4.3 for details). Then, a simple regression of $\hat{\varepsilon}_i * \hat{\phi}_i$ on $(WW')_{ii}$, $i = 1,...,n$ (without intercept term) leads to a consistent estimate of $\sigma_{u\eta}$.

In summary, the variable transformations on log house price ("standardized" log house price) and loss rate of developable open space (the percentage measure of developable open space around house) introduces the nonlinearity into the simultaneous system; the spatial correlation in both neighboring house prices and neighboring developable open spaces leads to the SEC specification of two error terms. Therefore, we end up with a nonlinear spatial simultaneous-equation model for joint determination of house price and surrounding developable open space.

4.2 Model identification

For linear simultaneous-equation system, the standard linear identification theorems assure that each original structural equation can be distinguished from a linear combination of the equations in the model. In the nonlinear case, unfortunately, the identification conditions need to be modified. Goldfeld and Quandt (1968) show that the basic difference between linear and nonlinear identifications comes from the observation that, in the linear case, only linear transformations of equations in the system generate a “new” equation which cannot be distinguished from, say, the first equation of the system; while in the nonlinear case, we also need to consider all possible nonlinear transformations of the original equations (what they call the implied equations). A linear combination of these implied equations and the original structural equations may not be distinguishable from, say, the first equation of the system, causing the identification problem. Accordingly, Fisher (1966) establishes two main results on nonlinear identification – a modified rank condition (sufficient and necessary) and a modified order condition (necessary only). The proposed spatial nonlinear model is identified by these two nonlinear identification conditions, as it can be shown that no such linear combinations of implied equations and the original structural equations exist.

4.3 Estimation methodology: 3SLS with strict 2SLS

The instrumental variables (IV) technique and full-information maximum likelihood (FIML) are widely used methods in the estimation of linear simultaneous equation systems. Nevertheless, the nonlinearity in the dependent variables (basic endogenous variables) introduces estimation difficulties. The FIML method works for both the linear and nonlinear cases; however, the computation gets more difficult for the nonlinear model in practice (see Goldfeld and Quandt, 1968, for an exposition). In the context of the proposed spatial nonlinear model, the assumed spatial error covariance structure brings further complexity into the numerical optimization procedure. In contrast, the IV method and related 2SLS or 3SLS are known for their operational convenience and robustness with respect to misspecifications. However, the automatic

application of IV methods appropriate for linear models does not produce consistent parameter estimates for the nonlinear models. As a result, some extensions of 2SLS to nonlinear models have been proposed. Kelejian (1971) demonstrates that a variant of the 2SLS technique can be used to estimate the parameters of a nonlinear model (nonlinear in the endogenous variables but linear in parameters).

Following Kelejian, consider an m-equation system with n observations:

$$y_{ij} = X_{ij}B_i + F_{ij}C_i + e_{ij}, i = 1,...,m; j = 1,...,n \quad (4.5)$$

where y_{ij} is the j th observation on the dependent variable (basic endogenous variable) in the i th equation; X_{ij} is a row vector of the j th observations on exogenous variables appearing in the i th equation, B_i is the corresponding column vector of parameters; F_{ij} is a row vector of the j th observations on endogenous functions (linear or nonlinear) of the dependent variables in the i th equation, each of which is assumed to depend on at least one dependent variable and any exogenous variables of the system, C_i is the associated column vector of parameters; e_{ij} is the j th disturbance in the i th equation.

For our 2-equation model, $y_{1j} = \ln p_j$, the natural log selling price of the j th house, $y_{2j} = OS_j\%$, the loss rate of developable open space; $X_{1j} = (1, X_j^H, X_j^L)$, the observations on all included structural and location characteristics of house j, $X_{2j} = (1, Z_j)$, the observations on the land-specific characteristics that affect open space loss but not house price. $F_{1j} = OS_j = f_1(y_{2j})$, the measure of developable open space surrounding house j when it is sold, is a nonlinear function of the dependent variable of equation 4.2, where $f_1(y_{2j}) = c_j(1 - y_{2j})$ (4.6), c_j is the measure of developable open space around house j 10 years before. Since c_j takes different

values for different houses, the endogenous function OS_j is linearly independent of $OS_j\%$ even though a linear relationship exists between $y_{2j} = OS_j\%$ and $f_1(y_{2j}) = c_j(1-y_{2j}) \quad = OS_j$ for each house j. $F_{2j} = \ln \tilde{p}_j = f_2(y_{1j}, X_j)$, the structural-characteristics-standardized natural log sale price of house j, is a function of the dependent variable of equation 4.1 and all exogenous variables in the system (since $\hat{\beta}_H$ is a function of all exogenous variables in the simultaneous model), where $f_2(y_{1j}, X_j) = \ln P_j - (X_j^H - \overline{X}^H)\hat{\beta}_H$. Also note that, as the consistent estimate of β_H ($\hat{\beta}_H$) from the second step of 2SLS is a linear function of $\ln p_j$, $\ln \tilde{p}_j$ is also linear in $\ln p_j$ for each j but with different "slope" across observations as in the case of OS_j and $OS_j\%$ above, $\ln \tilde{p}_j$ and $\ln p_j$ are therefore linearly independent $\forall j$; $e_{1j} = \varepsilon_j$ and $e_{2j} = \phi_j$ are the disturbances with specified SEC structure.

To estimate such a system, the endogenous functions $f_1(y_{2j})$ and $f_2(y_{1j}, X_j)$ on the right-hand side of the two equations are first regressed on a polynomial of all the exogenous variables X_{1j} and X_{2j} (including the exogenous variables that are excluded from the equation in question). The fitted values of $\hat{f}_1(y_{2j})$ and $\hat{f}_2(y_{1j}, X_j)$ then simply substitute for the two endogenous functions as new variables for a second step least square regression. Kelejian also emphasizes two additional points particular to the estimation of nonlinear models. First, the same degree polynomial must be used in the regression of endogenous functions; otherwise, the instruments will not be valid. Second, one has to use the entire endogenous function. We cannot regress the basic endogenous variables which comprise the function on a polynomial and then substitute these fitted values inside the endogenous functions, since it also leads to inconsistent estimates. That is, we cannot fit y_{1j} and y_{2j} to get $\hat{y}_{1j}$ and $\hat{y}_{2j}$ and then

use $f_1(\hat{y}_{2j})$ and $f_2(\hat{y}_{1j}, X_j)$ as the instruments in the second step regression of 2SLS.

The difference between this strict 2SLS method (regressing the entire endogenous function) by Kelejian and 2SLS in a linear system lies in that higher-order and corresponding interaction terms (so a polynomial) are included in the first-stage regression. This method has theoretical appeal in the sense that it is close to the ideal 2SLS instrument where the expectation of the endogenous function conditional on all exogenous variables in the system is used (nevertheless, we generally don't know the functional form of the expectation and it also depends on unknown model parameters). The strict 2SLS instrument $\hat{f}$ will be an efficient estimator provided that the order of the polynomial in stage-one regression is suitable. Furthermore, since I exploit the correlation between the errors of the two equations in the context of SEC structure, based on the residuals from the 2SLS procedure, a third-stage regression using GLS simultaneously for the two equations will be involved to incorporate the spatial information, resulting in a 3SLS estimation procedure. However, a disadvantage of this strict 2SLS method is, if a model is known for many exogenous variables, in order for the fitted value of the endogenous function to be a good instrument for the actual value, we have to include terms of second or higher order of the exogenous variables with all their cross-products. Moreover, stage-one high-order polynomial regression may increase the possibility of inverting a singular or badly-scaled matrix.

A question arises in the first two stage regressions of the 3SLS estimation procedure: does the specified SEC structure need to be included for each equation in the model? Following the estimation procedure of Kelejian and Prucha (2004), where two (IV/2SLS and IV/3SLS) estimators are suggested for a simultaneous system of spatially interrelated cross sectional equations (both including spatial lagged dependent variable and SAR error), the SAR error structure is not incorporated in the first two stage regressions. Similarly, here the SEC structure is incorporated only in the third stage regression when estimating the model.

To summarize, the steps of 3SLS estimation procedure with strict 2SLS are as follows:

1) Regress the endogenous function OS_j on the second-order polynomial of all exogenous variables (X_j^H, X_j^L, Z_j) in the model to find an instrument (OS_j^*) for OS_j. This is the stage-one regression for equation 4.1.

2) Regress $\ln p_j$ on X_j^H, X_j^L and OS_j^* to produce consistent estimates for all associated parameters. Note that the consistent estimate of house structural characteristics $(\widehat{\beta}_H)$ is used to construct the "standardized" log house price for equation 4.2. This is the stage-two regression of equation 4.1.

3) Construct the structural-characteristics-standardized log house price $\ln \tilde{p}_j$ following (4.3). Then regress the endogenous function $\ln \tilde{p}_j$ on the second-order polynomial of all exogenous variables (X_j^H, X_j^L, Z_j) in the model to find an instrument $(\ln \tilde{p}_j^*)$ for $\ln \tilde{p}_j$. This is the stage-one regression for equation 4.2.

4) Regress OS_j% on Z_j and $\ln \tilde{p}_j^*$ to produce consistent estimates for all associated parameters. This is the stage-two regression of equation 4.2.

5) Keep the residual estimates $\hat{\varepsilon}$ and $\hat{\phi}$ from the stage-two regressions of both equations and use GMM method to derive the variance component estimates of $\sigma_u^2, \sigma_v^2, \sigma_\eta^2, \sigma_\xi^2, \sigma_{u\eta}$ and thus the combined variance-covariance matrix 4.4 of the model.

6) Finally, a GLS regression [2] is applied simultaneously to both equations to produce the final estimates of all parameters in the model with OS_j and $\ln \tilde{p}_j$ replaced

by OS_j^* and $\ln \tilde{p}_j^*$. This is the stage-three regression.

4.4 Endogeneity revisited

This section discusses the relationship between residential land and developable open space and their implications on model estimation. In addition, two possible explanations are offered for the failure of the Hausman test in testing the endogeneity of developable open space and/or residential land in the model.

According to the category of land uses specified in section 5.3, most residential lands are evolved from developable open space (as shown in section 5.3, about 84% of developable open space within 400m of houses is zoned for residential use). It is reasonable to argue that, since the developable open space is endogenous and the amount of new residential land depends on the available developable open space, the resulting residential land is also endogenous. As a result, two endogenous variables exist in the model. In fact, there is only one implicit endogenous variable if we assume that all lost developable open spaces are converted into residential land. To explain this situation, define γ as the ratio of developable open space to the sum of developable open space and residential land in 2005. With $\gamma = \frac{OS\_\text{Res}05}{OS\_\text{Res}05 + \text{Res}05}$,

$OS\_\text{Res}05 = \gamma(OS\_\text{Res}05 + \text{Res}05)$ and $\text{Res}05 = (1-\gamma)(OS\_\text{Res}05 + \text{Res}05)$.

Note that the sum of developable open spaces and residential land is “fixed” since all other three land uses (industrial land, commercial land and protected open space) are assumed to be exogenous. This fact implies that how this “fixed” number is partitioned between the developable open spaces and residential land is endogenously determined. In other words, the only endogenous variable is the ratio γ.

To account for this assumption, a small modification is made to the stage-one regression of equation 4.1: the sum of estimated developable open spaces and residential land has to be "fixed" (although its value differs across observations), i.e., equal to the sum of observed developable open space and residential land. That is, in the stage-one regression of the hedonic pricing equation (equation 4.1), after the endogenous function (OS_j), percentage measure of surrounding developable open space, is regressed on the second-order polynomial of all exogenous variables in the model, an instrumented value for residential land is calculated by subtracting the instrumented developable open space (OS_j^*, here denoted as $OS\_\mathrm{Res}05^*$) from the "fixed" sum, that is, $\mathrm{Res}05^* = OS\_\mathrm{Res}05 + \mathrm{Res}05 - OS\_\mathrm{Res}05^*$. Subsequent estimation steps follow and are the same as in section 4.3.

In contrast to the standard 2SLS model, which assumes that the quantity of open space can adjust upward or downward to achieve equilibrium, the proposed model is built on the conceptual assumption that the quantity of open space can only adjust downward from historical level. Although the modification above is to account for the fact that open space loss is one-directional or irreversible, the empirical model does not technically restrict open space loss to be nonnegative (the econometric model allows negative open space loss). Therefore, it does not fully solve the problem since the irreversibility is not directly modeled. A simultaneous-equation econometric model that restricts open space loss to be nonnegative and that also incorporates spatial structure in the errors is an area for future research. Still, because the econometric model incorporates information on historical land use, it more realistically models the open space conversion process than does the standard 2SLS model.

Footnotes of Chapter 4:

1. There are two types of nonlinearities that are of interest in the simultaneous-equation models. One concerns the endogenous variables, while the other concerns nonlinear restrictions on the parameters. We're dealing with the first type of nonlinearity, which implies that there is at least one endogenous variable appearing in the model in two or more linearly independent forms.

2. There is an easy way to run the third stage of 3SLS (GLS) instead of using Orcutt-Cochran type transformation. Wonnacott R. and Wonnacott T. (1979, Chapter 20) show that, given a simultaneous-equation system in its structural form, after estimating the covariance matrix of the regression errors from the residual estimates of the second step of 2SLS, we can first transform (left-multiply) all the equations (both the left-hand side and right-hand side) by the transposed data matrix of all exogenous variables in the system, then simply run an GLS on the transformed equation system with transformed error covariance matrix to find all structural coefficient estimates. This method greatly facilitates the calculation of GLS since the dimension of the transformed system is just the number of all exogenous variables in the system, instead of the much larger number of observations.

Chapter 5 Model estimation

5.1 Data Development with GIS

The simultaneous-equation model is estimated using a data set that is developed using a Geographic Information System (GIS) and consists of information on individual house sales and land use in Berks County from 2002 to 2005. The database of land use is constructed from the 2005 parcel map of Berks County. This database, maintained by the Berks County Office of Assessment, is an ArcMap shape file (a digitized map) showing the boundaries of 151,913 parcels in the county, each of which is identified by a unique property id, as well as of roads and streams. A second database, the Office of the Assessment's 2005 residential CAMA file, includes information collected for each residential parcel for assessment purposes, such as characteristics of built structures on each parcel, which are necessary for the hedonic analysis, and information on the most recent sale of the parcel.

5.1.1 Residential property data

The digitized county parcel map is used for two purposes. First, it is used to identify residential parcels and so locate houses. For each residential parcel, the location of the house is assumed to be the centroid of the residential parcel.

Not all residential properties are included in the model. From the set of all houses, a subset is chosen to meet the criteria listed in Appendix D. This results in 4,007 single family houses. Some selection rules are worth mentioning here. For example, properties located in the City of Reading are excluded as the urban housing market is quite different from the suburban/rural housing market that is of interest here. More importantly, noting that where land is converted tends to be where sales occur (because almost every time a house gets built, it gets sold), more land use change is seen in places where house are selling, and houses with more development are therefore more likely to be selected. This creates a sampling problem (self-selection) as the houses are not randomly selected. This sample will overestimate the rate at

which land is converted. One way to avoid this problem is to only look at houses that were built no later than 1995, such that the parcel where a house is located is not converted at the sale time. Thus, the houses built after 1995 are excluded.

For each house in the sample, information is collected for all characteristics listed in Appendix E. The X,Y coordinates of each centroid are calculated to construct a distance-based spatial weights matrix and create a point map for many uses. One use is to calculate straight-line distances to commuting destinations. Another use is to create distance buffers around each house to produce some of the house location variables and land characteristics (see land use data section for details). For example, the commuting distances to two business centers, Reading and Philadelphia, are calculated as the distances from the house to the commuting waypoints to them. The commuting destination to Reading is set as the distance to downtown Reading. For Philadelphia, distance (in miles) is measured to the nearest of the two way-points, through (or near) which most commuting traffic would travel. The commuting waypoints for Philadelphia are the points where Rt. 422 and Rt. 23 cross the county border.

5.1.2 Land use data

The second use of the digitized county parcel map is to develop a county-wide map showing the land use of all parcels. This map is constructed based on 2005 land use pattern. Most land use change in this area is land conversion from developable open space to residential land as a result of residential property construction. The same land use map (for 2005) is used to measure surrounding land use for all sales between 2002 and 2005. It is assumed that the errors due to measuring surrounding land use one to three years after the sale are small.

All parcels in the county are categorized into five land uses. Specifically, they are

- Residential land
- Industrial land

- Commercial land
- Open space that can be developed for residential use (developable open space)
- Protected open space (publicly-owned open space)

The assignment starts with the land use codes given by the County Office of Assessment. According to those original codes, all parcels in the county fall into six categories: residential, industrial, commercial, institutional, governmental and vacant. Within each category, parcels are further assigned with land use codes according to specific land use or the type of building on the parcel. Note that these land use distinctions are derived from the perspective of the county tax assessor, which is based on the need of property tax collection but does not fit to the need of economic analysis of open space. On one hand, most of land use (codes) in institutional and governmental categories can be combined into commercial use as they impact nearby housing prices in a similar way. On the other hand, some land uses in commercial category such as landfill, quarry, meat packing plant, etc, are more suitable to be categorized into industrial use based on their expected impact on nearby house prices. More importantly, although some distinctions are made for vacant land (e.g., residential vacant land, institutional vacant land), the original codes do not distinguish different open spaces in the way as listed above.

Residential land is where a residential property is located. Industrial land is where the property is for industrial use, such as manufacture, warehouse, etc. Commercial land is a big category, including not only “conventional” commercial land use such as hotels, restaurants and stores but also governmental buildings, schools, hospitals and churches. Developable open space, the land use of most interest, is the open space that can be developed into residential use in the future. It includes vacant land, land in farms and privately-owned forests. Protected open space is publicly-owned open space that cannot be developed. Also note that there are a few parcels originally coded by the Office of Assessment as “vacant land” approved for commercial or industrial use. They are accordingly categorized into commercial land or industrial land. It is

assumed that these vacant lands are equivalent to commercial/industrial lands in the sense that people have expectation on their future usage.

Each parcel is assigned to one of the five land use categories to calculate the proportions of different land uses within a 400-meter buffer of each house at the time it was sold. These five measures enter the hedonic price equation (equation 4.1) as house location variables, and the measure of surrounding developable open space is used to calculate the dependent variable, the loss rate of developable open space, of the open space conversion equation (equation 4.2). It is important to obtain accurate measures of these surrounding land uses. Obviously, the measurement is closely related to how each parcel is assigned to different land uses.

One source of such information comes from the digitized county parcel map of 2002 with land use code assigned by the County Office of Assessment for each parcel. We may match the 2002 land use and 2005 land use for each parcel to see land conversion between different land uses. Specifically, for all residential parcels in 2005, the land uses in 2002 was examined to determine which land uses (codes) can be developed into residential land. Combining this information with the original vacant land codes, the category of developable open space parcel can be reasonably specified. For the category of protected open space, the 2005 county parcel map was matched with a digitized map of protected land where the overlapping parcels in the county are specified as protected open space (most of which are state game lands).

As mentioned before, the original land use codes are created for the purpose of tax collection, and some assignment error may still exist as the adjustments made above are based on the original land use codes. Further exploration is needed to decrease the assignment error. By carefully comparing the Metro Street Map of Berks County, which shows the main land uses in the county, with the county parcel map, it is found that all camping grounds, public parks and golf courses, which are more suitable to be categorized as protected open space, are designated with institutional land use codes

by the Office. Such specification errors are corrected. In spite of the endeavors to reducing the specification error in constructing new land use category, they cannot be totally eliminated. For example, one source of such error is the presence of agricultural easement. The parcels with agricultural easement are more suitable to be categorized as protected open space since further development is no longer allowed. Due to data limitations, they are not identified, and are treated as developable open space. When such information is available, corresponding re-categorization is suggested. By utilizing all available information, it is hoped that the error is reduced to an acceptable level such that the proportion measures of surrounding land uses are as accurate as possible.

In addition to the current land use measures, we also have to know the surrounding developable open space 10 year before the sale for each house in the sample to build the dependent variable of equation 4.2. To trace back this measure, a historical county parcel map is constructed for 1995, which shows all developable open spaces that existed at that time. Ideally, four historical maps are needed, one for each year during 1992-1995. However, considering the same reason of constructing just one parcel map for 2002-2005 and the fact that there is no more information on land conversion during this period, only one historical county parcel map is created. The historical map is constructed first by identifying the residential properties built after 1995 in the 2005 county parcel map. Since these residential parcels are developed after 1995, they can be reasonably specified as developable open spaces before the house construction, and so are assigned to the category of developable open space for the historical parcel map. Any developable open space that existed in 2005 is also considered to be developable open space in 1995. Here I focus on one type of land use - open space that can be developed in to residential land - because the historical parcel map is used to calculate the proportion of surrounding developable open space only (the other four measures of historical land uses are of no interest since they do not appear in the model).

All land use measures (five current and one historical) are calculated from the two digitized parcel maps using a program written in ArcView 3.1. Other digitized maps (grids) needed to assemble the complete data set include a county zoning map, a county elevation map, a county slope map and a county soils map. Combined with the historical developable open space map, the zoning map is used to calculate the proportion of developable open space within 400-meter buffer of the house in 1995 that is zoned for residential use. The elevation map provides average measures of land elevation within a 100-meter buffer and a 400-meter buffer of the house. The slope map provides average measures of the slope of developable open space within a 400-meter buffer of the house. The soils map provides average measures of suitability for building and agricultural productivity of the developable open space within a 400-meter buffer of the house. Those two soil-type variables, the slope of developable open space and the zoning variable affect the feasibility and profitability of land conversion from developable open space to residential land. They appear in equation 4.2 as Z variables and act as the instruments in estimating the system equations.

A few other location characteristic variables were constructed. A “Village” dummy variable denotes whether the house is located in a builtup village. A “Road” dummy variable is constructed by specifying if at least one state or interstate road crosses the 400-meter buffer of a house. Also included are two demographical variables, Minority and MedInc, which measure the proportion of minority (Hispanic and Black) and household median income of the block group where the house is located. These location variables are included in the X vector in equation 4.1.

5.2 One-market or two-market model?

A potential problem with hedonic models arises due to the underlying assumption that the study area represents one market for housing services. Because a hedonic price function is an equilibrium function describing a specific market, all properties used in a hedonic regression must be part of the same housing market. When choosing a sample frame for a hedonic analysis, one must consider the geographic coverage of

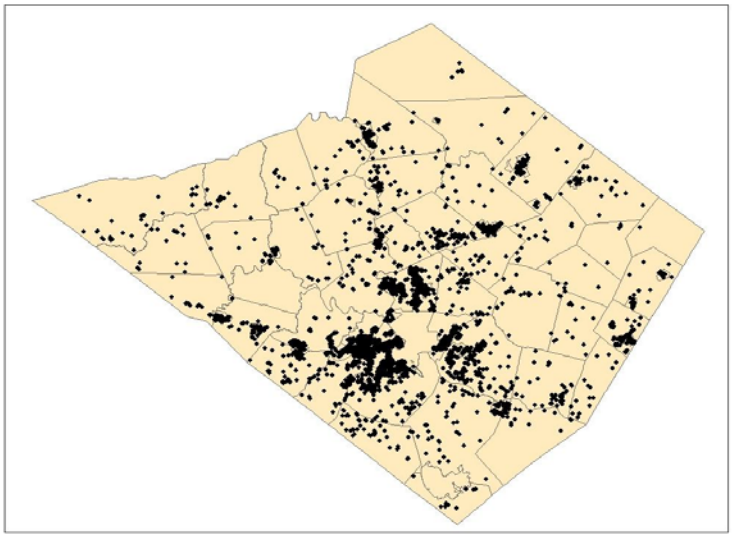
the data selected. If the sample is made up of properties from different housing markets, estimating one hedonic price function for the entire sample is inappropriate. The question then is how to determine what set of properties comprises a single market. Markets are separate if participants in one market do not consider houses in other markets when making purchase decision. It is commonly assumed that each urban area represents a separate housing market although there are evidences for market segmentation across urban areas. As there is no definite answers to such a problem, researcher's judgment along with supporting statistical tests are used as a guide to determine the market extent in a particular study.

To avoid multi housing markets and market niche, as mentioned in Appendix D, properties located in the City of Reading (considered an urban housing market compared with other areas of the County) and New Morgan Borough (the Borough has an unusually high proportion of land in industrial use) have been excluded. Does the area other than the City of Reading and New Morgan Borough represent a single housing market?

A brief look at the county map reveals that, excluding the City of Reading and New Morgan Borough, the county roughly consists of three areas: the densely-populated area surrounding the City of Reading where households are assumed to be tied through work or shopping to the City, the more rural area in the southeast part of the county where two highways to Philadelphia go through and people may commute to Philadelphia to work, and the remaining rural area in the north and west parts of the county. Loosely speaking, these three areas are distinguished by different geographical, demographical and economic characteristics and could be different housing markets. The area immediately surrounding the City of Reading is distinct from the other two areas in its suburban makeup as opposed to the rural nature. There are very few observations on house sales from the southeast part of the county (see Figure 5.1), so it is difficult to judge whether the rural southeast portions are distinct from the northern and western parts of the county. Consequently, the county will be

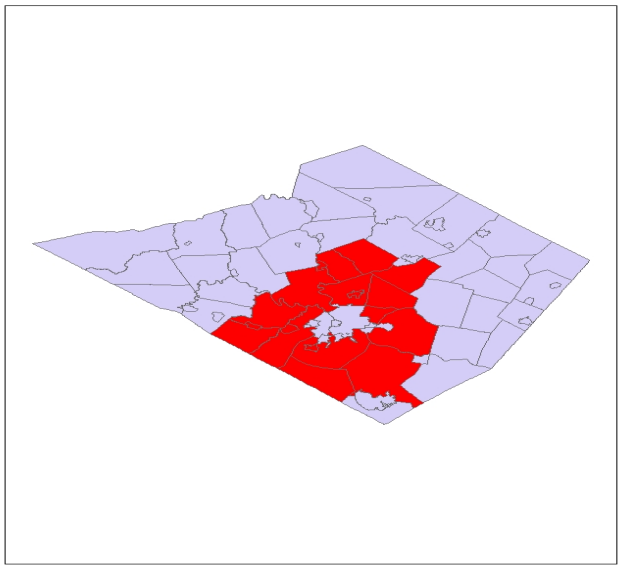
divided into two regions for analysis - the area around the City of Reading (area1) and the remaining area (area2), resulting in two potential housing markets.

Figure 5.1 Distributions of House Sales in Berks County



To formalize this idea, the assumed two housing markets need to be delineated definitely so that the sample observations can be split into each market according to their locations. The delineation is based on the township in the county. Since more urban areas are characterized by less developable open space compared with more rural areas, an ad hoc selection rule is adopted to find townships forming area1 according to the proportion of developable open space of each township, excluding protected open space. Specifically, the rule is set to be $0.20 \le \frac{\%DOS}{1-\%POS} \le 0.70$, where DOS denotes developable open space and POS denotes protected open space. To get a compact shape for area1, small adjustments are made such that the townships designated to the area are contiguous. The remaining townships comprise area2. This result is shown in Figure 5.2 where the dark color represents area1 (the City of Reading excluded) and the shadowed townships represent area2 (New Morgan Borough excluded).

Figure 5.2 Two housing markets in Berks County



The two-market distinction is also supported statistically from the results of the hedonic regression. The one-market model is a restricted model in the sense that it runs a single regression for all observations in the sample, and so restricts the regression coefficients to be the same for the two housing markets (if they are really separate). The two-market model is an unrestricted model since it does not restrict the coefficients to be the same. For the proposed simultaneous-equation model, as the focus is mainly on the hedonic equation, the unrestricted hedonic regression can be represented using dummy variable to allow for differences in regression coefficients for both assumed housing markets. Specifically, defining a dummy variable as below

$$D_j = \begin{cases} 1 & \text{if house j in area1} \\ 0 & \text{if house j in area2} \end{cases}$$

and following the same notation in section 4.1.2, we have

$$\ln P_j = \beta_0 + D_j\delta_0 + X_j^H\beta_H + D_j X_j^H\delta_H + X_j^L\beta_L + D_j X_j^L\delta_L + OS_j\beta + D_j OS_j\delta + \varepsilon_j$$

$$\varepsilon_j = W_j u + v_j, \quad j = 1,\ldots\ldots,n \qquad (5.1)$$

Under the null hypothesis of $H_0 : \delta_0 = 0, \delta_H = 0, \delta_L = 0, \delta = 0$ (the restricted model), the value of F statistic (with degrees of freedom 37 and 3925) 6.1455 rejects the restricted model in an highly significant way. The three asymptotic tests (Wald = 232.47, LR = 225.97 and LM = 219.72) also support the rejection. Also note that, with an assumed SEC structure for the regression error, the two-market model implies that the variance components of the spillover errors (σ_u^2 and σ_η^2) and idiosyncratic errors (σ_v^2 and σ_ξ^2), as well as the covariance between two spillover errors across equations ($\sigma_{u\eta}$) cannot be all the same for the two areas. If the two housing markets do exist, they should operate independently, then the error variance (covariance) estimates for each market are from two independent samples. Given the estimated variances and their standard deviation estimation from the second stage regression, a simple pairwise t test applies for testing the equality of each pair of corresponding error component variances (total 5 such tests). The test results reject the null hypothesis in three of the five cases, suggesting the inequality of error component variances across the two housing markets, and thus supporting the two-market model. Based on these results, all subsequent estimations are conducted independently for the two market areas.

5.3 Data description

In section 5.2, a two-market model is shown to be preferred to a single-market model. Accordingly, the sample is divided into two sub-samples, one for each housing market. As a result, of total 4007 houses in the sample, 2137 houses are located in the suburban area1 and 1870 houses are located in the rural area2. This section will give an overview of the data.

Table 5.1 shows the summary statistics of two dependent variables, house price (not log price used in the model) and loss rate of developable open space, and other independent variables in the model except surrounding land uses, for both areas.

On average, the house sale price of area1 is a bit higher than area2 and has lower dispersion. Since properties with sale price less than $25000 are excluded from the sample, the minimum prices are just a bit above this value. There is no limit on the upper bound and the maximum house price of area2 is far greater than that of area1. For open space loss rate, during the 10-year period, about 24% developable open space within 400m of house was lost in area1, only 20% lost in area2. In area1, 317 houses didn't lose any surrounding developable open space (317 zeros) and 38 houses lost all their surrounding developable open space (38 ones). In area2, 362 houses didn't lose any surrounding developable open space (362 zeros) and 29 houses lost all their surrounding developable open space (29 ones). Since area2 has smaller sample size, it has higher proportion of houses without losing any surrounding developable open space than area1. The mean loss rate and proportion of house with no open space loss coincide with the expectation of the difference between the two areas.

Generally speaking, there is no critical difference between the two areas in house structural characteristics although small discrepancy exists. Note that, as properties with lot size larger than 5 acres or less than 0.035 acres are excluded from the sample, the maximum and minimum lot sizes are confined to this range. The two distance measures (DistReading and DistPhilly) take distinct values between areas. Since area1 surrounds the City of Reading, it is obvious that the mean and maximum distances from houses in area1 to the city are much smaller than in area2. The distance to Philadelphia is measured as the distance to the commuting points located at eastern boundary of area2 towards Philadelphia, the mean distance measures are not so different for the two areas; however, as area2 contains houses in its western part and houses just beside the commuting points, we observe much larger maximum distance measure and a minimum of zero in area2.

Table 5.1: Summary statistics for hedonic price equation - house price (HP), open space loss rate (OS%), and house characteristics

	AREA1				AREA2			
	Mean	Std	Max	Min	Mean	Std	Max	Min
Dependent Variable								
HP	119526	52235	469800	25100	111086	59245	1214700	26600
OS%	0.244	0.258	1	0	0.196	0.239	1	0
Sale Year Dummy								
Sold03	0.328	0.470	1	0	0.326	0.469	1	0
Sold04	0.205	0.404	1	0	0.222	0.416	1	0
Sold05	0.054	0.227	1	0	0.064	0.245	1	0
House Structure								
Age	4.670	2.891	10.50	0.70	6.081	2.964	10.50	0.70
Livarea	1.608	0.666	11.388	0.60	1.604	0.662	7.336	0.616
Livarea2	3.029	4.073	129.69	0.36	3.010	3.316	53.817	0.379
Lotsize	0.398	0.604	4.78	0.04	0.343	0.587	4.85	0.04
Lotsize2	0.523	2.062	22.848	0.002	0.461	1.895	23.523	0.002
Bdrms	3.068	0.768	8	1	3.098	0.788	9	1
Flbth	1.326	0.549	4	0	1.260	0.532	4	0
Hfbth	0.475	0.512	3	0	0.381	0.513	3	0
Bsmnt	0.899	0.279	1	0	0.944	0.213	1	0
Stone	0.017	0.129	1	0	0.025	0.155	1	0
Brick	0.275	0.446	1	0	0.449	0.497	1	0
Mason	0.113	0.317	1	0	0.086	0.280	1	0
Aircond	0.402	0.490	1	0	0.223	0.416	1	0
Phycd	3.012	0.295	4	2	3.017	0.318	4	2
Attic	0.229	0.350	1	0	0.280	0.326	1	0

	AREA1				AREA2			
	Mean	Std	Max	Min	Mean	Std	Max	Min
Detached	0.954	0.209	1	0	0.814	0.389	1	0
Water	0.808	0.394	1	0	0.794	0.404	1	0
Sewer	0.831	0.375	1	0	0.805	0.396	1	0
House Location								
DistReading	4.983	1.940	10.53	1.88	8.420	6.033	25.91	5.26
DistPhilly	12.507	3.322	19.86	3.5	14.932	6.191	35.41	0
Road	0.844	0.363	1	0	0.898	0.302	1	0
Elev100	122.75	42.389	333.13	44.67	121.72	30.74	310.74	44.15
Elev100400	1.20	5.403	28.376	-25.27	0.338	4.544	26.997	-29.80
PSSA	13.523	0.448	13.87	11.97	13.725	0.337	14.20	11.97
Minority	0.033	0.037	0.331	0	0.044	0.061	0.331	0.
MedInc	4.860	1.235	10.271	2.604	4.961	1.189	8.751	2.911
Village	0.202	0.402	1	0	0.291	0.454	1	0

Surrounding land uses is the category of most interest. According to the land use classification defined in section 5.1, among total 151913 parcels of the study area, 118777 parcels are residential land, 1956 parcels are in industrial use, 11980 parcels are commercial land, 18606 parcels are developable open spaces and 594 parcels are protected open spaces. From Table 5.2, within 400m of house, residential land accounts for the largest land use proportion (55% in area1 and 49% in area2), commercial land has the second largest proportion (about 22% in area1 and 24% in area2), next is developable open space (17% in area1 and 19% in area2) and industrial land (about 5% in both areas), protected open space is the least (2% in both areas). A typical house in area1 has 12% more surrounding residential land than a typical house in area2. In contrast, a typical house in area2 is surrounded by 9% more developable open space and 12% more commercial land than a typical house in area1. As the

proportions of surrounding industrial land and protected open space are relatively small, the differences of these two land use types between the two areas are not significant (although area2 contains vast protected open spaces on the northwestern side, few houses in the sample are located in the neighborhood of that area).

Table 5.2: Summary statistics for hedonic price equation – proportion of surrounding land uses within 400m of house

	AREA1				AREA2			
	Mean	Std	Max	Min	Mean	Std	Max	Min
Residential	0.550	0.189	0.994	0.0068	0.493	0.204	0.985	0.015
Industrial	0.046	0.078	0.740	0	0.059	0.095	0.708	0
Commercial	0.215	0.160	0.826	0	0.241	0.163	0.847	0
OS_Res	0.170	0.184	0.950	0	0.185	0.232	0.985	0
OS_Pro	0.020	0.065	0.874	0	0.022	0.067	0.932	0

Out of 2137 houses in the sample of area1, 979 (46%) houses do not have industrial land within 400m, 119 (5%) houses do not have surrounding commercial land, 38 (1.8%) houses do not have surrounding developable open space and 1419 (66%) houses do not have surrounding protected open space. In area2, out of 1870 houses in the sample, these numbers are 631 (34%), 115 (6%), 29 (1.6%) and 1429 (76%), respectively. Residential land exists within 400m of 100% of the houses in both areas.

Table 5.3 shows the summary statistics for some important land characteristics of developable open space within 400m of house. The average building suitability, average slope, average agricultural productivity, and the residential zoning proportion are similar in both areas. In addition, the proportions of surrounding developable open space of 1995 are almost the same for the two areas, although area1 experienced a larger loss rate during the 10-year period.

Table 5.3: Summary statistics for open space loss equation

	AREA1				AREA2			
	Mean	Std	Max	Min	Mean	Std	Max	Min
Slope400	2.009	1.233	16.22	0	1.928	1.033	6.16	0.16
Buildsuit	1.939	2.554	10	0	1.519	2.033	9.72	0
Agprod	37.903	30.823	100	0	34.338	31.635	100	0
Zoning	0.832	0.286	1	0	0.851	0.298	1	0
OS95	0.186	0.190	0.918	0	0.184	0.229	0.925	0.0002

5.4 Estimation procedure and results

5.4.1 Choice of spatial weights matrix

In our spatial simultaneous-equation model (4.1 and 4.2), two spatial weights matrices have to be specified for each equation before model estimation, one for the spatial correlation of neighboring houses and the other for the spatial relationship between neighboring developable open spaces. Distance-based weights are adopted. As in the landfill case in Chapter 3, the weights are constructed based on a linear distance decay function with an ad hoc cutoff distance. In fact, the functional form for weight is not important as empirical studies have shown that different choices on functional form of weights have weak effects on model estimation. However, the choice of cutoff distance, which specifies the boundary of the “neighborhood”, is of importance. On one hand, a correctly specified neighborhood includes all neighboring houses whose sale prices affect the price of the house of interest. A large cutoff distance includes houses whose sale prices actually have no or negligible impacts. A small cutoff distance ignores some houses whose sale prices indeed matter. It has been shown that the choice of cutoff distance influences the model estimates in a significant way. On the other hand, with the SEC error structure, if the specified error structure is “reasonable”, a “wrong” cutoff distance may lead to insignificant or even negative variance estimates of the two error components, thus contradicting the SEC specification.

The choice of spatial weights matrix is largely an empirical problem. In fact, because two spatial weights matrices have to be specified for our model, the determination of cutoff distance becomes more complex than the case of single spatial weights matrix. Two problems are inherent in the specification of two spatial weights matrices: are they row-standardized or not and are the cutoff distances the same for both of them? It may happen that the data support row-standardization for one weights matrix but favor non-standardization for the other. It is also possible that the extent of neighborhood of houses is different from the extent of neighborhood of open spaces. In addition, these two problems are interrelated with each other; in other words, the determination of cutoff distance and weights structure has to be simultaneous.

As the study area is similar to that in the landfill example where a cutoff distance of 1600 meters is used, I will choose the cutoff distance from three alternatives: 1600m, 3200m and 4800m. Since there are three alternatives of cutoff distance and two choices on weights structure (row-standardization and non-standardization), six combinations of cutoff distance and weights structure exist. Two criteria are applied to guide the choice. First, the LM statistic of the SEC structure is calculated for each combination, the combination with the largest statistic value is preferred. Then, the preferred weights matrix is adopted in the second stage regression to find the two error component variance estimates. If both of the estimates are positive and statistically significant, the chosen cutoff distance is believed to describe the extent of underlying neighborhood and the weights structure is believed to correctly capture the extent of influences of all neighbors. If not, the weights matrix with the second largest LM statistic is chosen and the corresponding error variance estimates are computed. If both of those estimates are positive and statistically significant, then stop; otherwise, repeat the procedure above until a weights matrix is found to satisfy both criteria. In our case of two housing markets and two regression equations, for equation i of area j ($i = 1,2$, $j = 1,2$), six LM statistics need to be calculated for each combination of cutoff distance and weights structure.

Tables 5.4 – 5.7 show the six LM statistics for each combination of housing market and regression equation. First note that the LM statistic values for the open space loss equation are much higher than those for hedonic price equation in both areas, this result coincides with our expectation that the extent of spatial effect for open space is larger than that for houses.

Table 5.4: LM statistics for the hedonic price equation (equation 4.1) of area1

	1600m	3200m	4800m
row-standardization	140.08	62.78	13.19
non-standardization	70.21	19.25	4.13

Table 5.5: LM statistics for the open space loss equation (equation 4.2) of area1

	1600m	3200m	4800m
row-standardization	816.61	637.64	216.65
non-standardization	2211.90	1119.70	354.82

For the hedonic price equation of area1, since the first cell of Table 5.4 has the largest LM value (140.08) and the corresponding SEC error component variance estimates are both positive and statistically significant (see Table 5.8 for error component variance and covariance estimates of each chosen spatial weights matrix), the 1600m row-standardized weights matrix is selected. Since 2211.90 is the largest value in Table 5.5, the 1600m non-standardized weights matrix is first chosen; however, the spillover error variance estimate is negative, then I go to the cell with the second largest LM value (1119.70). The 3200m non-standardized weights matrix produces positive and significant error variance estimates and thus is chosen for the open space loss equation of area1.

For the hedonic price equation of area2, since 76.91 is the largest LM value in Table 5.6, the 1600m non-standardized weights matrix is first chosen. As in the case of open space loss equation of area1, the resulting spillover error variance estimate is negative. Then, the 1600m row-standardized weights matrix with second largest LM value (68.87) is selected since it yields positive and significant error variance estimates. For the open space equation of area2, a similar situation happens: the 1600m non-standardized weights matrix with the largest LM value is rejected and the 3200m non-standardized weights matrix with the second largest LM value is chosen.

Table 5.6: LM statistics for the hedonic price equation (equation 4.1) of area2

	1600m	3200m	4800m
row-standardization	68.87	19.27	9.79
non-standardization	76.91	58.90	9.65

Table 5.7: LM statistics for the open space loss equation (equation 4.2) of area2

	1600m	3200m	4800m
row-standardization	231.35	145.15	116.83
non-standardization	3728.50	2274.80	338.25

To summarize，different spatial weights matrices are used in hedonic price equation and open space loss equation, but both areas use the same set of weights matrices for each equation: 1600m row-standardized weights matrix in equation 4.1 and 3200m non-standardized weights matrix in equation 4.2.

5.4.2 Estimation results and analysis

The proposed simultaneous-equation model is estimated using a 3SLS with strict 2SLS method by taking into account of the endogeneity of developable open space and the spatial correlation of neighboring house prices and neighboring open spaces. We are interest in the impact of developable open space on nearby house price and comparisons of such impacts between developable open space and other surrounding

land use types. The inclusion of the open space loss equation in the model also provides a way of examining the factors influencing the conversion process from developable open space to residential land. Since the measures of the proportions of all five surrounding land use types sum to one, one type of land use has to be dropped in the estimation procedure under which the marginal implicit price estimates of the remaining four surrounding land uses are relative to the omitted land use. Thus, for a complete comparison of all types of surrounding land uses, the model is estimated five times, each time with a different omitted land use. In fact, each estimation procedure only differs in the final step regression (GLS regression) because the residual estimates and the resulting estimated error covariance matrix from the second stage regression should be the same no matter which land use is omitted. In addition, all coefficient estimates should be the same except those for the intercept term and surrounding land use variables in the hedonic price equation because their values are relative to the omitted land use.

Table 5.8 shows the SEC error component variance estimates of each equation and covariance estimates of between-equation error for both areas. All variance estimates are significantly positive as required in section 5.4.1. The covariance of between-equation error represents how unobserved factors influencing house price are interacted with the unobserved factors influencing surrounding developable open space loss. A negative and significant covariance estimate in area1 suggests that these unobserved or omitted factors are working in an opposite way in the suburban area. Consider some unobserved restrictions on house building. For example, farm easement protects the farm land from developing into residential use, resulting in a low loss rate of developable open space but increasing house price through the scarcity effect. This effect is more likely to occur in area1 where available developable open spaces are relative scarce, while in area2 this effect is not significant.

Table 5.8: Estimation results of the SEC error structure

		AREA1	AREA2
Equation 4.1	σ_v^2	0.0268	0.0290
		(31.86)	(21.09)
	σ_u^2	0.0233	0.0190
		(4.38)	(2.92)
Equation 4.2	σ_ξ^2	0.0434	0.0295
		(19.70)	(9.96)
	σ_η^2	0.00008	0.00018
		(4.69)	(5.39)
Between-Equation	$\sigma_{u\eta}$	-0.0025	-0.00027
		(-2.22)	(-0.25)

Note: t-statistics are in parentheses

The marginal implicit price estimates for surrounding land uses and intercept estimate are shown in Table 5.9. Other coefficient estimates of hedonic price equation are included in Appendix F.

In both areas, most house structural variables are significant and have the expected signs. For example, living area and lot size of a house increase the house price at a decreasing rate as the linear term is significantly positive and the quadratic term is significantly negative. The magnitudes of those estimates are also fairly close in each area. These results suggest that people have similar preference on house structural characteristics. It's a bit surprising that the variable "Bdrms" (number of bedrooms) is insignificant, but note the high correlation between this variable and living area, it can be inferred that part of impact of number of bedrooms on house price is included in the variable "Livarea", resulting in an insignificant estimate. It is not clear why the dummy variable sewer is significantly negative in area1, which implies that having public sewer decreases house price. Ready and Abdalla (2005) report an insignificant estimate for the same variable using a similar dataset.

For house location characteristics, different estimation results exist between the two areas. The distance to the City of Reading is found to be positive and significant for houses in area2 but insignificant for houses in area1. On the other hand, the distance to Philadelphia impacts house price in an opposite way as the distance to the City of Reading: a significantly positive effect in area1 and a significantly negative effect in area2. Generally speaking, the effects of these distance measures are hard to explain since the spatial structure (the distance-based spatial weights matrix in this case) is correlated with the distance variables that makes interpretation and inference problematic. Note that the dummy variable "Village" (whether the house is located in a village) is an area where many houses are clustered and can be considered as high-density residential land, its significantly negative coefficient estimate implies that house buyers do not favor high-density living where the negative externalities among neighbors are intensive.

From Table 5.9, as one of the five land uses is omitted each time, in each case the coefficient values should be viewed as the marginal difference between the land use in question and the omitted land use. Notice the symmetry of the estimates across each omitted land use case. For example, when industrial land is the base land use, the coefficient estimate of commercial land is 0.2038 (t = 3.37) in area1; when commercial land is the base land use, the coefficient estimate of industrial land is -0.2038 (t = -3.37) in the same area. All estimation results are combined to make a complete comparison of all surrounding land uses.

For area2, within 400m of house, residential land is found to be weakly (at 10% significance level) preferred to protected open space and highly (at 1% significance level) preferred to industrial land and commercial land; commercial land is found to be weakly (at 10% significance level) preferred to developable open space. All other comparisons between land uses are not even weakly significant. These results show that, in area2, (1) residential land is the most favorable land use type; (2) all other land uses are statistically indistinguishable.

By contrast, in area1, people's preferences on surrounding land uses can be nicely ordered. Within 400m of house, protected open space and residential land are superior to the other three land use types. Commercial land and developable open space are more favorable than industrial land. However, there is no evidence that people differentiate between protected open space and residential land or they value commercial land and developable open space differently. These results show that, in area1, (1) protected open space and residential land are the most favorable surrounding land use types; (2) industrial land is the worst favored surrounding land use type; (3) commercial land and developable open space are "neutral" in the sense that they are in the middle of the preference sequence. Obviously, the comparison of all surrounding land use types is more complete in area1 than in area2. This distinction is shown as below

$$\text{Area1: } \begin{matrix}\text{Protected open space}\\ \text{Residential land}\end{matrix} \succ \begin{matrix}\text{Developable open space}\\ \text{Commercial land}\end{matrix} \succ \text{Industrial land}$$

$$\text{Area2: } \text{Residential land} \geq \text{Protected open space}$$

$$\text{Residential land} \succ \text{Commercial land} \geq \text{Developable open space}$$

$$\text{Residential land} \succ \text{Industrial land}$$

where $\succ$ denotes "preferred to at 5% level" and $\geq$ denotes "preferred to at 10% level".

Table 5.9: Hedonic price equation estimation results – Surrounding land uses

	AREA1	AREA2
Industrial omitted		
Residential	0.2862 (5.53)	0.1478 (2.77)
Commercial	0.2038 (3.37)	-0.0244 (-0.41)
OS_Res	0.1937 (3.41)	-0.0922 (-1.62)
OS_Pro	0.3839 (4.69)	0.0152 (0.17)

	AREA1	AREA2
Commercial omitted		
Residential	0.0824	0.1722
	(2.87)	(4.86)
Industrial	-0.2038	0.0244
	(-3.37)	(0.41)
OS_Res	-0.0101	-0.0678
	(-0.26)	(-1.66)
OS_Pro	0.1801	0.0395
	(2.60)	(0.512)
Residential omitted		
Industrial	-0.2862	-0.1478
	(-5.53)	(-2.77)
Commercial	-0.0824	-0.1722
	(-2.87)	(-4.86)
OS_Res	-0.0926	-0.24
	(-2.69)	(-6.28)
OS_Pro	0.0977	-0.1326
	(1.50)	(-1.81)
OS_Pro omitted		
Residential	-0.0977	0.1326
	(-1.50)	(1.81)
Industrial	-0.3839	-0.0152
	(-4.69)	(-0.17)
Commercial	-0.1801	-0.0395
	(-2.60)	(-0.51)
OS_Res	-0.1902	-0.1073
	(-2.78)	(-1.39)
OS_Res omitted		
Residential	0.0926	0.24
	(2.69)	(6.28)
Industrial	-0.1937	0.0922
	(-3.41)	(1.62)
Commercial	0.0101	0.0678
	(0.26)	(1.66)
OS_Pro	0.1902	0.1073
	(2.78)	(1.39)

Note: t-statistics are in parentheses

As a result, people living in the two areas have different preferences for surrounding land uses. In area1, people like surrounding protected open space, but in area2 people do not prefer it to other land uses. Although developable open space is shown to be less desirable than residential land in both areas, it is at least “better” than industrial

land in area1; however, it is not the case in area2. Those differences in people's preference on surrounding land uses across areas may be attributed to the distinct landscape characteristics and the extent of scarcity of different land use types. On the other hand, there are some common findings for both areas. First, residential land is found to be most favored. Second, commercial land is equally preferred to developable open space. One reason to account for the large and significant marginal implicit price estimate of residential land is that part of negative externalities from clustered living has been captured by the dummy variable "Village" as mentioned before.

Focusing on the two open space types, the estimation results establish that people indeed treat protected open space and developable open space separately and prefer the open space that is protected from development in the future. In other words, uncertainty about future development offsets the amenity effect of open space. The fear of potential disamenity from future construction could be an important reason to account for house buyer's negative attitude toward developable open space. The significantly positive marginal implicit price estimate of protected open space suggests that, even accounting for possible negative externalities associated with open-space proximity such as noise and traffic, the amenity effects of protected open space dominate and increase nearby house price. Importantly, the estimate reflects only a fraction of the benefit from preserving open spaces. Benefits that have a strong "public good" element, such as improvement in water quality resulting from open space preservation, are unlikely to affect a house's sale price and so not reflected in the estimation result of the study.

Table 5.10 lists the estimation results of open space loss equation. What is of most interest is the coefficient estimate of the standardized log house price (St_logp) which is the house price net of house structural characteristics by construction. The significantly positive estimates ($t = 2.02$ and $t = 4.83$) for both areas suggest that, irrespective of the differences between areas, higher house price brings about larger

loss rate of surrounding developable open space, which is just what the endogeneity means. Therefore, the presupposed endogeneity of surrounding developable open space is supported by the simultaneous-equation modeling; otherwise, it does not make sense to model their simultaneous determination.

Table 5.10: Open space loss equation estimation results

Variable	AREA1	AREA2
Constant2	-3.5230	-7.1317
	(-1.90)	(-4.67)
Slope400	-0.0362	-0.0311
	(-3.95)	(-3.81)
Buildsuit	0.0142	0.0076
	(3.92)	(2.14)
Agprod	0.00004	-0.0026
	(0.12)	(-8.24)
Zoning	0.1210	0.0945
	(5.75)	(4.78)
OS95	-0.2345	-0.0025
	(-5.25)	(-0.07)
St_logp	0.3225	0.6396
	(2.02)	(4.83)

Note: t-statistics are in parentheses

Most land characteristics are found to significantly affect the rate at which developable open space is lost in the expected ways. The steeper slope of developable open space within 400m of house, the harder to build a house on it, leading to a lower loss rate. The better suited the soils are for building, the lower the construction cost, the more surrounding developable open space is lost. The larger proportion of the open space zoned for residential use, the more likely it is converted into residential land. Developable open space with high agricultural productivity is less likely to be converted in area2, but that is not the case in area1. This effect shows that opportunity cost matter. However, the proportion measure of surrounding developable open space in 1995 is found to be significant in area1 but not in area2. This result could be attributed to the scarcity of developable open space and the resulting high demand for residential land in the suburban area1.

5.4.3 Estimation result comparison: OLS, IV/2SLS and 3SLS/Strict 2SLS

In this section, I'll briefly compare the proposed simultaneous model with the OLS model and the simple IV/2SLS model on hedonic price equation estimation. The simple IV/2SLS model is estimated using all available instruments: the five land characteristic measures of surrounding developable open space (Z) in the open space loss equation (equation 4.2).

As implied in section 4.4, the presupposed endogeneity of surrounding developable open space may not be (fully) corrected by the simple IV/2SLS estimation. There are two possible interpretations. First, as mentioned in section 4.4, the simple IV/2SLS approach implicitly assumes that the quantity of open space can adjust upward or downward to reach equilibrium. Accordingly, the Hausman test is based on a two-way adjustment process. But, in reality, the level of open space in a specific area can only adjust downward. Second, the adjustment process is not instantaneous. It takes time for the residential land market to reach equilibrium and the land market may still be on the way to equilibrium at the time of data collection. This result can be easily seen from the closeness of the OLS estimates and the IV/2SLS estimates on all land use variables in both areas (see Table 5.11 for an illustration). In fact, all other coefficient estimates of the OLS and simple IV/2SLS are also fairly close (see Appendix G). The Hausman test based on simple IV/2SLS estimation fails to reject the null hypothesis of no endogeneity for developable open space and/or residential land (all p-values are above 0.5).

Theoretically, if surrounding open space is not endogenous, it only affects nearby house prices from the demand side. That is, surrounding open space would shift the demand curve for house outward if its amenity effect is appreciated, thus driving the house price up. If the open space variable is endogenous, it not only affects nearby house price from the demand side but also from the supply side since the amount of open space also responds to area house prices. The loss of open space will shift the house supply curve to the left. When both demand and supply for houses move in this

way, the resulting house price is higher than when there is only demand effect (or when the surrounding open space is not endogenous). Therefore, the endogeneity confounds the demand effect (amenity effect) with the supply effect. If it does exist but not accounted for, the estimate of open space variable, i.e., the OLS estimate, will be biased downward. In this case, a suitable instrumental variable method is expected to include the supply effect and result in a larger estimate.

Note that the scarcity of open space is a precondition for the analysis above. If surrounding open space is not scarce, as in the rural area2, its amenity effect on the demand side becomes small since it is not a well-appreciated surrounding land use type. For the supply side, if the open space is not in scarcity, the effect of open space loss on the availability of open space for future residential use will be limited. As a result, its supply effect in housing market is weak or even negligible. In essence, the endogeneity problem mainly refers to the effect of open space loss on house supply. When this supply effect is weak, the endogeneity problem will not be a concern.

Since the proposed simultaneous model has shown that the amount of developable open space indeed responds to nearby house prices and the simple IV/2SLS method does not correct for the endogeneity problem, the simple IV/2SLS estimate of surrounding developable open space is still biased downward. In contrast, the proposed simultaneous approach can better correct for the bias and produce consistent estimation.

Table 5.11 Comparison of the estimation results of surrounding land uses in area1 when industrial land is the base land use - OLS, IV/2SLS and 3SLS/Strict 2SLS

AREA1	OLS	IV/2SLS	3SLS/Strict 2SLS
Residential	0.2535 (4.93)	0.2503 (4.84)	0.2862 (5.53)
Commercial	0.1664 (2.78)	0.1660 (2.77)	0.2038 (3.37)
OS_Res	0.1304 (2.37)	0.1388 (2.48)	0.1937 (3.41)
OS-Pro	0.2991 (3.95)	0.3003 (3.96)	0.3839 (4.69)

Note: t-statistics are in parentheses

The estimation results of the simple IV/2SLS method shows a similar pattern of people's preference on different surrounding land uses as the proposed simultaneous model for both areas. However, the magnitudes of the two sets of estimates on surrounding land use are found to be different in area1, but this is not the case in area2 since endogeneity problem is not expected to be that prominent in this rural area. The difference in the estimate magnitude is not trivial since it represents house buyers' different valuation of surrounding land use. For example, from Table 5.11 where the base land use is industrial land, in area1, the estimate for protected open space from the proposed simultaneous model and the IV/2SLS model are 0.384 and 0.300, respectively. Since the dependent variable is the natural log of house prices, the estimated marginal implicit price measures the percentage change of house price from a one-unit change in a house attribute. For a house with sale price $200,000, if a 1% surrounding industrial land is converted to protected open space, this marginal change will increase the house price by $768 (0.384*0.01*20000) under the proposed simultaneous model and $600 (0.3*0.01*20000) under the simple IV/2SLS model. Consider a same 1% conversion from industrial land to developable open space, as the coefficient estimates are 0.194 and 0.139, respectively, this marginal change will increase the house price by $388 (0.194*0.01*20000) under the proposed simultaneous model and $278 (0.139*0.01*20000) under the simple IV/2SLS model. The higher the house price, the larger the magnitude of this real dollar difference

between the two estimates.

In summary, since the simple IV/2SLS method does not correct for the endogeneity of developable open space, the OLS estimation of the hedonic price equation is similar to the simple IV/2SLS estimation of the hedonic price equation using all available instruments, both underestimating people's willingness-to-pay for improvement of surrounding land use. By contrast, the proposed simultaneous model corrects for the estimation bias caused by endogeneity problem and yields consistent coefficient estimates for all surrounding land use variables.

Chapter 6 Conclusions and discussions

This final chapter concludes the thesis and discusses some important problems for further study.

6.1 Spatial regression model selection

This thesis discusses the spatial hedonic regression models and their applications in evaluating environmental amenity and disamenity, e.g., distance of house to landfills and the effect of open space on nearby house prices. Chapter 2 and 3 examine the specification of spatial hedonic pricing models – an improvement on traditional hedonic regression model by taking into account of spatial dimension of house price data. The focus is on how to incorporate the spatial correlation of observed house price or error into a hedonic model. Since the spatial weights matrix is an integral part of the spatial specification and the choice of weights matrix has important impacts on model estimation result, spatial model selection and the issue of row-standardization of spatial weights matrix are explored simultaneously. I examine the theoretical implications of three popular spatial specifications, the spatial lag model, the SAR error model and the SEC model, in the context of house price data. The implications of weights schemes, row-standardization or non-standardization, are discussed within each of the three spatial models. I argue that the spatial error components (SEC) structure provides a better model for house price data although it is infrequently used in hedonic analysis. The SEC model is preferred because it best represents the intuition about the process that drives the spatial correlation of house prices by explicitly specifying two sources of variations in house price given house characteristics: a spillover error that captures the common unobserved factors affecting neighboring house prices and an idiosyncratic error that only captures house specific unobserved error that will not “spill over” to neighboring house. I also question the “convention” of row-standardizing spatial weights matrix in practice and argue that, to model house price data, a row-standardized spatial weights matrix is more appropriate for the spatial lag model and probably the SEC model while a non-

standardizing spatial weights matrix might be more appropriate for the SAR error model. In addition to the theoretical consideration, there exist three empirical methods, the LM tests, a pseudo-R^2 criterion and the Bayes factor method, for choosing among these competing models.

An empirical study on the impact of landfills on nearby house prices is used to illustrate and justify these arguments by estimating different spatial models and applying the three spatial model selection methods. The estimation results also suggest that, although the simple OLS model may give spurious marginal implicit price estimates by ignoring the spatial correlation of house price, "incorrect" spatial specification may produce even worse estimate outcomes than OLS. Therefore, it is always important and necessary to select a "best" spatial model from some alternatives by considering their theoretical implications and employing these empirical methods to guide the choice.

Several points are worth mentioned here. First, the choice of spatial weights matrix is important. The extent of neighborhood (e.g., the cutoff distance) and the weights scheme (e.g., row-standardization or non-standardization) have to be carefully specified within each spatial model in empirical study. It cannot be separate from the model specification as the two concerns are interdependent. Second, on the preferred SEC model, some extensions have been proposed to include potential spatial heteroskedasticity in the error structure, e.g., Kelejian and Yuzefovich (2004). The argument favoring the SEC model is made in the context of house price data; whether this model is also favored for data set with spatial dimension other than house price data is not clear and needs further exploration. Researchers should explore whether the SEC model reasonably captures the mechanism that drives spatial correlation in the data of interest. It is possible that the spatial lag model or the SAR error model better represents the spatial process of the variables in question than the SEC model. Third, although three empirical methods are proposed to guide the model selection, they may not always lead to the same conclusion. Furthermore, for the most widely

used LM test, no robust test is available to distinguish the SEC model from the spatial lag model or the SAR error model. The choice between them could be ambiguous in practice. For the Bayes factor method, an efficient numerical integration routine needs to be developed to handle the common large-sample problem. Obviously, the calculated Bayes factors in Table 3.4 incur numerical imprecision although they may roughly guide the model selection.

6.2 Land use modeling analysis

Based on the arguments made in chapter 2 and 3, a simultaneous-equation model with SEC error structure is proposed to investigate the issue of estimating the impact of the endogenous developable open space on nearby property values. This new approach is expected to improve the traditional IV/2SLS methodology of dealing with endogeneity by including the information on the dynamic process of open space conversion and incorporating the spatial dimension in house price data. A 3SLS with strict 2SLS methodology is introduced to estimate the resulting spatial nonlinear simultaneous-equation system. Along with the information on individual house sale, Geographic Information System (GIS) is involved to develop the geography-related data required in the model. The geographic character of the study area and the resulting data set show that the study area, Berks County, can be divided into two housing markets. Accordingly, the model is estimated for each housing market independently. The estimation results confirm the presupposed endogeneity of developable open space, which, however, is not detected by the Hausman test (simple IV/2SLS estimation).

A clear distinction of people's preferences on different surrounding land uses appears between the two market areas. House buyers in the urban-rural fringe area (area1) show a nicely ordered preference sequence on surrounding land uses. In area1, residential land and protected open space are in the top tier of the list, next are commercial land and developable open space, industrial land lies in the bottom. This result is believed to be coincident with the social and geographic characteristics of

this area where protected open space is relatively scarce and people are assumed to be tied to the City of Reading for work and shopping. In the rural area2, no such a complete preference sequence exists. Except that the surrounding residential land is most favored, the other four surrounding land uses are almost incomparable in the sense that house buyers in area2 seem to be indifferent to each of them. Focusing on the two open space types, surrounding protected open space is a favored land use type in area1; but in area2, it is not significantly different from developable open space, commercial land and industrial land. Unlike protected open space, surrounding developable open space is not a favored land use type in both areas. This result may be due to house buyers' worry about negative externalities from future house construction on these parcels, or uncertainty over the nature of the building that could occur. Those estimation results suggest an amenity effect of open space when it is protected from future development and that this amenity effect may depend on how scarce the open space is. The finding of the amenity effect of protected open space is consistent with most empirical studies; however, developable open space is not found to provide a positive effect on nearby house price, although it is empirically detected by some other studies. For example, Ready and Abdalla (2005) find that the conversion of agricultural open space, i.e., farm land, to small-lot residential, commercial or industrial use would negatively impact nearby house price. Furthermore, the uncertainty about future development of developable open space is found to play an important role in people's valuation of this type of land use.

Land use modeling is a challenging research question because the spatial effect of land use prevails and the classification and measurement of different land uses are not uniform and in some cases difficult. In this study, the hedonic price equation classifies the land use around house into five exhaustive categories. Since there are a plenty of different land uses, subjective judgment has to be involved in the categorization process based on all available information. In addition, some coding errors may exist in the original land use code book. All of which are potential sources of measurement error in each land use category. For the two types of open space of most interest, they

are also combinations of different open spaces, especially the developable open space. Developable open space is a mixed category including farmland, forest, pastureland, cropland, etc, each of which may have different impact on nearby house price. For example, Irwin (2002) reports significant positive effects of cropland and pastureland but significant negative effect of forested land. Therefore, if information on this further classification of developable open space is available, examining their separate effects is expected to yield more information than examining only the combined effect.

One source of error in land use coding is the presence of agricultural easement. These are agricultural parcels where the development rights have been permanently sold, so that development is no longer allowed. Ideally, these parcels would have been identified and classified as protected open space. Due to data limitations, they were not identified, and are treated as developable open space. The large majority of these parcels occur in area2. If these protected parcels generate more positive impacts on nearby house prices than developable open space, then the potential error introduced by this mis-categorization is that the marginal implicit price for developable open space will be biased upward. Because the marginal implicit price for developable open space (relative to residential land) is negative in area2, this error is not likely to qualitatively change any of the conclusions.

As mentioned in section 4.4, due to the one-side (downward) adjustment process of developable open space, its presupposed endogeneity is not detected by the Hausman test. However, the estimation result of the proposed simultaneous-equation model shows that the endogeneity indeed exists. A further exploration is needed to address this problem, including theoretical formalization and development of new estimation and test methods. In section 4.1.2, I justify the use of a simple linear regression equation to model the loss rate of developable open space, leaving a room for exploring more suitable modeling method, which, of course, depends on the available data.

Last but not the least, the analysis includes observations from only one county. This spatial nonlinear simultaneous-equation approach should be extended to a broader area, or to other regions where the land use situations are quite different from Berks County or the open space is well- or bad-endowed compared with the study area. In addition, an attention must be paid to the question whether the study area constitutes a single house market. If not, an aggregate analysis is inappropriate.

Appendix A: Bayes Factor Method - Marginal likelihood for the SEC model

Consider the spatial error components model

$y = X\beta + \varepsilon$

$\varepsilon = W\phi + u$

$E(\varepsilon\varepsilon') = \sigma_u^2(I + \theta WW')$, $\theta = \sigma_\phi^2 / \sigma_u^2$,

Assumption1: ε is normally distributed.

Assumption2: Uniform, non-informative prior for β and σ_u^2

$$p(\beta, \sigma_u^2 \mid Model_{SEC}) \propto 1/\sigma_u$$

Assumption3: Uniform (and proper) prior for the spatial parameter θ with the range $D \in (0, \infty)$

$$p(\theta \mid Model_{SEC}) \propto 1/D$$

As a result, it can be shown that the marginal likelihood of the data given the SEC model is

$$p(y \mid Model_{SEC}) = \frac{1}{D}\Gamma(\frac{n-k}{2})(2\pi)^{-(n-k)/2}\int_0^m \frac{|L|^{-1}}{|X^{*'}X^*|^{1/2} * (s^2)^{(n-k)/2}} d\theta$$

$$\propto \frac{1}{D}\int_0^m \frac{|L|^{-1}}{|X^{*'}X^*|^{1/2} * (s^2)^{(n-k)/2}} d\theta$$

where $LL' = I + \theta WW'$ (Cholesky factorization, L is lower-triangular), $X^* = L^{-1}X$, $y^* = L^{-1}y$, $|\ |$ denotes determinant, s^2 is the residual sum of squares of the regression of y^* on X^*, $\Gamma(.)$ is the gamma function, n is the sample size and k is the number of explanatory variables (including the intercept term).

Appendix A: Bayes Factor Method - Marginal likelihood for the SEC model (Continued)

Note that $\Gamma(\frac{n-k}{2})$ and $(2\pi)^{-(n-k)/2}$ are two common terms in the marginal likelihoods for all three spatial models considered under the similar assumptions above, they cancel out each other when calculating the Bayes factor for any two competing spatial models.

The theoretical integration interval for the SEC model is $\theta \in (0,\infty)$, which is not operational for our purpose. As we will not expect a very large ratio (θ) of spillover error variance over local error variance in most cases and it can be shown that the marginal likelihood of SEC model is increasing with θ, I just pick $m = \theta_{\max} = 20$ for the landfill example of chapter 3 ($\widehat{\theta} = 3.014$ in this case). A larger $\theta_{\max}$ value would yield stronger evidence in favor of the SEC model.

For technical details, please refer to Hepple (2004).

Appendix B: Variable Definitions – Landfill Example

Name	Description
Year Sold	Dummy variables for year of house sale, omitted year is 1998
Sold99	Is the house sold in 1999? (1=yes, 0=no)
Sold00	Is the house sold in 2000? (1=yes, 0=no)
Sold01	Is the house sold in 2001? (1=yes, 0=no)
Sold02	Is the house sold in 2002? (1=yes, 0=no)
House Structural Characteristics	
Age	Age of structure at time of sale (10 years)
Age2	Age squared
Livarea	Living Area of house (1000 square feet)
Livarea2	Living Area squared
Lotsize	Lot Size (acres)
Lotsize2	Lot Size squared
Bdrms	# Bedrooms
Flbth	# Full Baths
Hfbth	# Half Baths
Bsmnt	Does it have a basement? (1=full, 0=no basement)
Stone	Stone Exterior (1=has stone exterior, 0=does not have stone exterior)
Brick	Brick Exterior (1=has brick exterior, 0=does not have brick exterior)
Mason	Masonry Exterior (1=has masonry exterior, 0=does not have masonry exterior)
Aircond	Does it have central air conditioning? (1=yes, 0=no)
Phycd	Physical Condition (1=Good, 4=Fair -- houses rated poor or unsound are excluded from analysis)
Detached	Is the house detached? (1=detached, 0=row house, duplex, etc)
House Location Characteristics	
DistAllenkm	Straight-line distance to Allentown(km)
DistPhillykm	Straight-line distance to commuting waypoints toward Philadelphia (km)
Slope100	Average slope within 100-meter of house (degree)
Elev100m	Average elevation within 100m of house (meters)
Elev100800	Average elevation within 100m of house minus average elevation within 800m of house (meters)
PSSA	Mean PSSA test score for school district (100 points)
Pind400m	% of land within 400m of house in industrial use
Pind400800	% of land between 400m and 800m from house in industrial use

Appendix B: Variable Definitions – Landfill Example (Continued)

Name	Description
PCLmi2	Proximity index for Pioneer Crossing landfill (PCL)
PCL10000dum	House is within 10 km of PCL (1=yes)
WBLmi2	Proximity index for Western Berks landfill (WBL)
WBL10000dum	House is within 10 km of WBL (1=yes)
RHLmi2	Proximity index for Rolling Hills Landfill (RHL)
RHL10000dum	House is within 10 km of RHL (1=yes)

Note: RHL is also called Delaware County Landfill

Appendix C: Other estimation results of landfill example – House structural and location characteristics, sale year and township dummies

Variable	OLS	Lag(W_s)	SAR(W_s)	SAR(W_n)	SEC(W_s)
Constant	0.8583 (0.55)	-0.3568 (-2.55)	3.6679 (2.09)	0.6170 (0.37)	3.4723 (2.06)
Sale Year Dummy					
Sold99	-0.0090 (-1.84)	-0.0085 (-1.78)	-0.0069 (-1.48)	-0.0087 (-1.82)	-0.0083 (-1.75)
Sold00	-0.0211 (-4.29)	-0.0219 (-4.57)	-0.0204 (-4.31)	-0.0215 (-4.41)	-0.0210 (-4.41)
Sold01	-0.0249 (-4.91)	-0.0259 (-5.22)	-0.0245 (-5.01)	-0.0241 (-4.87)	-0.0249 (-5.05)
Sold02	-0.0167 (-0.29)	-0.0223 (-0.40)	-0.0028 (-0.05)	-0.0177 (-0.32)	-0.0069 (-0.12)
House Structure					
Age	-0.0380 (-12.64)	-0.0288 (-9.64)	-0.0431 (-13.63)	-0.0414 (-13.00)	-0.0417 (-13.46)
Age2	0.0002 (0.82)	-0.0004 (-1.41)	0.0007 (2.52)	0.0006 (1.99)	0.0006 (2.06)
Livarea	0.3383 (30.37)	0.3287 (30.23)	0.3262 (30.05)	0.3238 (29.49)	0.3269 (29.94)
Livarea2	-0.0173 (-9.15)	-0.0174 (-9.46)	-0.0171 (-9.27)	-0.0158 (-8.46)	-0.0161 (-8.67)
Lotsize	0.3030 (29.53)	0.2643 (25.70)	0.2964 (27.13)	0.2948 (28.47)	0.3066 (28.09)
Lotsize2	-0.0547 (-21.03)	-0.0471 (-18.25)	-0.0521 (-19.76)	-0.0253 (-20.21)	-0.0550 (-20.27)
Bdrms	0.0141 (4.10)	0.0149 (4.45)	0.0145 (4.37)	0.0140 (4.17)	0.0128 (3.81)
Flbth	0.0626 (12.27)	0.0619 (12.42)	0.0568 (11.41)	0.0614 (12.24)	0.0599 (11.97)
Hfbth	0.0394 (8.56)	0.0370 (8.23)	0.0355 (7.95)	0.0370 (8.24)	0.0360 (7.80)
Bsmnt	0.0891 (11.11)	0.0775 (9.86)	0.0678 (8.32)	0.0755 (9.28)	0.0787 (9.85)
Stone	0.1795 (12.27)	0.1818 (12.72)	0.1761 (12.43)	0.1794 (12.55)	0.1788 (12.27)
Brick	0.0522 (9.34)	0.0493 (9.02)	0.0525 (9.54)	0.0546 (9.90)	0.0525 (9.54)
Mason	0.0284 (4.24)	0.0304 (4.65)	0.0309 (4.71)	0.0312 (4.73)	0.0307 (4.68)

Appendix C: Other estimation results of landfill example – House structural and location characteristics, and sale year dummies (Continued)

Variable	OLS	Lag(W_s)	SAR(W_s)	SAR(W_n)	SEC(W_s)
Aircond	0.0457 (8.65)	0.0434 (8.41)	0.0428 (8.31)	0.0433 (8.35)	0.0430 (8.33)
Phycd	-0.0774 (-10.37)	-0.0816 (-11.19)	-0.0751 (-10.43)	-0.0777 (-10.61)	-0.0736 (-10.04)
Detached	0.1054 (13.15)	0.1176 (14.97)	0.1238 (15.24)	0.1217 (14.91)	0.1169 (14.68)
House Location					
DistAllenkm	-0.0074 (-2.71)	-0.0076 (-2.85)	-0.0075 (-1.53)	-0.0098 (-2.77)	-0.0062 (-1.66)
DistPhillykm	-0.0048 (-1.67)	-0.0014 (-0.51)	-0.0077 (-1.53)	-0.0047 (-1.35)	-0.0068 (-1.70)
Slope100	-0.0045 (-4.81)	-0.0027 (-2.97)	-0.0052 (-5.26)	-0.0055 (-5.82)	-0.0054 (-5.69)
Elev100m	-0.0004 (-4.62)	-0.0005 (-5.73)	-0.0001 (-0.56)	-0.0003 (-3.51)	-0.0002 (-1.90)
Elev100800	0.0028 (15.03)	0.0025 (13.69)	0.0022 (9.38)	0.0027 (13.08)	0.0026 (12.18)
PSSA	0.7631 (6.51)	0.6743 (2.43)	0.5575 (4.24)	0.7875 (6.28)	0.5678 (4.49)
PCL10000dum	0.0103 (0.62)	-0.0016 (-0.10)	-0.0256 (-0.89)	0.0061 (0.34)	-0.0125 (-0.54)
WBL10000dum	0.0398 (1.33)	-0.0156 (-0.53)	-0.0021 (-0.06)	0.0347 (1.16)	0.0110 (0.34)
RHL10000dum	0.0403 (3.34)	0.0077 (0.64)	-0.0075 (-0.39)	0.0269 (1.79)	0.0241 (1.53)
Township Dummy					
ALSAC	-0.0170 (-0.46)	0.0182 (0.56)	0.0086 (-0.15)	0.0015 (0.04)	0.0072 (0.15)
AMITY	-0.1338 (-3.95)	-0.1494 (-4.52)	-0.0437 (-0.89)	-0.1143 (-3.03)	-0.0668 (-1.68)
BECTH	-0.0765 (-1.55)	-0.0556 (-1.23)	-0.0330 (-0.37)	-0.0780 (-1.43)	-0.0088 (-0.14)
BERNT	-0.2052 (-5.46)	-0.1833 (-5.68)	-0.0699 (-1.02)	-0.1972 (-4.42)	-0.0995 (-1.94)
BOYER	-0.0749 (-1.33)	-0.0705 (-1.33)	0.0203 (0.26)	-0.0576 (-0.97)	0.0076 (0.11)
BRECK	-0.1891 (-5.41)	-0.1781 (-5.80)	-0.1299 (-2.27)	-0.1485 (-3.66)	-0.1317 (-2.88)

Appendix C: Other estimation results of landfill example – House structural and location characteristics, and sale year dummies (Continued)

Variable	OLS	Lag(W_s)	SAR(W_s)	SAR(W_n)	SEC(W_s)
COLEB	-0.1936	-0.1764	-0.0984	-0.1752	-0.1064
	(-4.29)	(-4.33)	(-1.41)	(-3.55)	(-1.92)
CUMRU	-0.1641	-0.1291	-0.0860	-0.1159	-0.0932
	(-5.37)	(-5.06)	(-1.70)	(-3.18)	(-2.36)
DISTR	-0.0567	-0.0450	0.0758	-0.0538	-0.0474
	(-1.01)	(-0.82)	(0.73)	(-0.88)	(-0.58)
DOUGL	-0.2871	-0.2360	-0.1020	-0.2603	-0.1613
	(-5.40)	(-4.79)	(-1.26)	(-4.68)	(-2.39)
EARLT	-0.1171	-0.1070	-0.0800	-0.1061	-0.0461
	(-2.16)	(-2.13)	(-0.99)	(-1.86)	(-0.68)
EXETR	-0.0644	-0.0432	0.0179	-0.0427	-0.0015
	(-3.32)	(-2.42)	(0.47)	(-1.65)	(-0.06)
HEREF	-0.6530	-0.6888	-0.5301	-0.6469	-0.7100
	(-4.20)	(-4.53)	(-3.13)	(-4.21)	(-3.60)
KENHO	-0.2417	-0.1862	-0.1278	-0.1756	-0.1367
	(-6.94)	(-6.12)	(-2.35)	(-4.19)	(-3.14)
LAURE	-0.3061	-0.2128	-0.1988	-0.2861	-0.1953
	(-8.00)	(-7.23)	(-3.27)	(-6.10)	(-4.05)
LWRAL	-0.4069	-0.2867	-0.1842	-0.3768	-0.2208
	(-8.38)	(-12.12)	(-2.89)	(-6.83)	(-3.95)
LWRHD	-0.3114	-0.3290	-0.2101	-0.2749	-0.2040
	(-6.11)	(-9.30)	(-2.87)	(-4.74)	(-3.27)
MOHNT	-0.1206	-0.0828	-0.0049	-0.0661	-0.0321
	(-3.66)	(-2.92)	(-0.09)	(-1.64)	(-0.76)
MOUNT	-0.4103	-0.2930	-0.2474	-0.4095	-0.2646
	(-8.41)	(-11.78)	(-3.81)	(-7.33)	(-4.71)
MUHLE	-0.2772	-0.2167	-0.1829	-0.2644	-0.1754
	(-7.33)	(-7.57)	(-3.10)	(-5.93)	(-3.68)
OLEYT	-0.0105	-0.0147	-0.0016	0.0080	0.0076
	(-0.24)	(-0.37)	(-0.02)	(0.16)	(0.14)
PIKET	-0.0111	-0.0475	0.0051	0.0049	0.0146
	(-0.21)	(-0.98)	(0.05)	(0.09)	(0.20)
ROBES	-0.1274	-0.0937	-0.0386	-0.0877	-0.0623
	(-5.58)	(-4.18)	(-0.99)	(-3.11)	(-2.05)
ROCKL	-0.1319	-0.1637	-0.1232	-0.1262	-0.1060
	(-2.76)	(-3.55)	(-1.46)	(-2.35)	(-1.69)
RUSCO	-0.0424	-0.1057	-0.0897	-0.0262	-0.0511
	(-0.77)	(-2.07)	(-0.92)	(-0.44	(-0.68)
SHILL	-0.1900	-0.1149	-0.0982	-0.1225	-0.1216
	(-6.01)	(-4.24)	(-1.88)	(-3.00)	(-2.98)

Appendix C: Other estimation results of landfill example – House structural and location characteristics, and sale year dummies (Continued)

Variable	OLS	Lag(W_s)	SAR(W_s)	SAR(W_n)	SEC(W_s)
SINKI	-0.3531	-0.3283	-0.2224	-0.2809	-0.2207
	(-7.35)	(-10.70)	(-3.34)	(-5.14)	(-3.86)
SPRIN	-0.2797	-0.2378	-0.1513	-0.2090	-0.1605
	(-6.09)	(-8.56)	(-2.35)	(-4.00)	(-2.92)
STLAW	-0.1836	-0.1520	-0.1087	-0.1575	-0.1345
	(-7.10)	(-6.19)	(-2.54)	(-4.85)	(-4.06)
UNION	-0.0916	-0.1249	-0.0467	-0.0609	-0.0528
	(-2.59)	(-3.61)	(-1.11)	(-1.65)	(-1.36)
WASHI	-0.1894	-0.1544	-0.1065	-0.1829	-0.0912
	(-3.35)	(-2.94)	(-1.14)	(-3.05)	(-1.23)
WESTL	-0.3404	-0.2607	-0.1652	-0.1999.	-0.1789
	(-7.11)	(-8.39)	(-2.49)	(-3.61)	(-3.14)
WESTR	-0.5168	-0.3979	-0.2880	-0.4651	-0.3208
	(-8.17)	(-13.69)	(-3.57)	(-6.68)	(-4.44)
WYOMI	-0.2421	-0.1970	-0.0252	-0.1889	-0.0589
	(-4.08)	(-7.25)	(-0.33)	(-2.93)	(-0.85)

Note: t-statistics in parentheses; the omitted township is Birsdboro.

Appendix D: Selection rules for residential properties

Exclusion Rule	Rationale/Discussion
Excluded all mobile homes	- Regression models house values, not land values - Difficult to determine how much of the sale price is attributable to the structure.
Only included arms-length sales	- Used Office of the Assessment's validity codes
Excluded properties with lot larger than 5 acres	- Larger properties may have uses other than residential - Larger properties may be eligible for preferential use assessment
Excluded properties with lot smaller than 0.035 acres	- Avoid parcels with typographic errors in this field - Avoid condo and townhouse properties where the parcel is defined as the footprint of the building, but the owner has use of shared lawn/green space not included in the parcel
Only included properties sold in 2002 to 2005	- Want assessment data recorded in 2005 to be accurate at the time of the sale - Want land use map constructed based on 2005 data to be accurate at the time of the sale
Excluded properties where sale price is more than 25% different from assessed value	- Avoid parcels with typographic errors in sale price field - Avoid parcels where assessment data is incorrect at time of sale, due to improvements
Excluded properties with less than 600 square feet livable area	- Avoid parcels with typographic errors in this field - Avoid unique parcels, such as 1-bedroom condos split off from a larger house, that are difficult to value
Excluded properties with sale price less than $25,000	- Avoid parcels with typographic errors in sale price field - Avoid parcels with unknown adverse conditions
Excluded properties with physical condition rated as "poor" or "unsound"	- Used Office of the Assessment's physical condition code - Parcels that are damaged, condemned, or in poor condition are more difficult to value
Excluded properties with no delopable open space to start with in 1995	- The dependent variable, loss rate of developable open space as percentage of that in 1995, cannot be defined.
Excluded parcels located within 400 meters of county boundary	- To get an accurate measure of proportion of different land use within 400 meters of the property. Because the land use map covers only Berks County, we must restrict ourselves to properties where a circle of 400 meters radius around the house falls entirely within the county.

Appendix D: Selection rules for residential properties (Continued)

Exclusion Rule	Rationale/Discussion
Excluded properties built after 1995	- Avoid possible sampling problem (self-selection)
Excluded properties located in the City of Reading	- Avoid multi housing market (urban housing market quite different from rural/suburban housing market)
Excluded houses sold in New Morgan Borough	- Avoid market niche. The proportion of land in New Morgan Borough in industrial use is unusually high. Residential properties located in New Morgan Borough would have unusually large quantities of nearby industrial land, and would have a disproportionately large impact on the estimated relationship between nearby industrial land use and house sale price.

Appendix E: Variable definitions – Land use

Name	Description
lnp	Natural log of the real sale price, deflated to 2005 first half dollars using the annual northeast urban consumer price index (CPI-U), the dependent variable of equation 1
OS%	Loss rate of 2005 residential open space, defined as ($1-OS\_Res_{05}/OS\_Res_{95}$), the dependent variable of equation 2
Year Sold	Dummy variables for year of house sale, omitted year is 2002
Sold03	Is the house sold in 2003? (1=yes, 0=no)
Sold04	Is the house sold in 2004? (1=yes, 0=no)
Sold05	Is the house sold in 2005? (1=yes, 0=no)
House Structural Characteristics	
Age	Age of house at time of sale = (Year Sold – Year Built)/10 (10 years) - Houses built prior to 1900 are given build date of 1900
Livarea	Living Area (1000 square feet)
Livarea2	Living Area squared
Lotsize	Lot Size (acres)
Lotsize2	Lot Size squared
Bdrms	# Bedrooms
Flbth	# Full Baths
Hfbth	# Half Baths
Bsmnt	Does it have a basement? (1=full, 0.5=partial, 0=no basement)
Stone	Stone Exterior (1=yes, 0=no)
Brick	Brick Exterior (1= yes, 0=no)
Mason	Masonry Exterior (1= yes, 0=no)
Aircond	Does it have central air conditioning? (1=yes, 0=no)
Phycd	Physical Condition (1=Excellent, 2=Good, 3=Average, 4=Fair -- houses rated poor or unsound are excluded from analysis)
Attic	Does it have finished attic? (1=full finish, 0.5=part or unfinished, 0=none)
Detached	Is the house detached? (1=detached, 0=row house, duplex, etc)
Water	Does it have public water? (1=yes, 0=no)
Sewer	Does it have public sewer? (1=yes, 0=no)
House Location Characteristics	
DistReading	Straight-line distance to downtown Reading (miles)
DistPhilly	Straight-line distance to commuting waypoints toward Philadelphia (miles)
Road	Is a state or interstate road within 400-meter of the house? (1=yes, 0=no)

Appendix E: Variable definitions – Land use (Continued)

Name	Description
Elev100m	Average elevation within 100m of house (meters)
Elev100400	Average elevation within 100m of house minus average elevation within 400m of house (meters)
PSSA	Mean PSSA test score for school district (100 points)
Minority	Percentage of Minority (Hispanic and Black) of the block group where the house is located
MedInc	Household median income ($10,000) of the block group where the house is located
Village	Is the house located in a village? (1=yes, 0=no)
Residential	Proportion of land within 400-meter of house currently in residential use in 2005 (e.g., 1=all residential land, 0=no residential land)
Industrial	Proportion of land and vacant land within 400-meter of house currently in industrial use in 2005 and designated for future industrial use
Commercial	Proportion of land and vacant land within 400-meter of house currently in commercial use in 2005 and designated for future commercial use
OS_Res	Proportion of land within 400-meter of house currently being private open space for residential use in 2005
OS_Pro	Proportion of land within 400-meter of house currently being protected open space in 2005
Land Characteristics	
Slope400m	Average slope of 1995 residential open space (OS_Res) within 400m of house (degree)
Buildsuit	Average measure of building suitability of 1995 residential open space (OS_Res) within 400m of house (0 = poor, 10=good)
Agprod	Average measure of agricultural productivity of 1995 developable open space (OS_Res) within 400m of house (0 =low, 100=high)
Zoning	Proportion of developable open space (OS_Res) within 400m of house in 1995 zoned for residential use
St_logp	Standardized natural log of house price net of the effects of house structural characteristics
OS95	Proportion of land within 400m of house currently being private open space for residential use in 1995

Appendix F: Hedonic pricing equation estimation results – House structural and location characteristics

Variable	AREA1	AREA2
Sale year dummy		
Sold03	0.0152 (1.84)	-0.0128 (-1.31)
Sold04	-0.0058 (-0.61)	-0.0111 (-1.00)
Sold05	-0.0467 (-2.90)	-0.032 (-1.83)
House Structure		
Age	-0.040 (-21.53)	-0.0442 (-20.22)
Livarea	0.3913 (26.46)	0.3867 (16.45)
Livarea2	-0.0270 (-13.54)	-0.0261 (-6.16)
Lotsize	0.3829 (17.41)	0.3103 (10.69)
Lotsize2	-0.0733 (-13.49)	-0.0638 (-8.34)
Bdrms	0.0046 (0.79)	-0.012 (-1.83)
Flbth	0.0734 (7.74)	0.0627 (5.65)
Hfbth	0.0405 (4.65)	0.0490 (5.11)
Bsmnt	0.1467 (10.83)	0.1530 (7.09)
Stone	0.1588 (5.54)	0.1785 (6.35)
Brick	0.0870 (9.53)	0.0578 (5.65)
Mason	0.0404 (3.36)	0.0954 (5.97)
Aircond	0.0317 (3.39)	0.1038 (8.42)
Phycd	-0.0840 (-6.55)	-0.1110 (-8.13)
Attic	0.0362 (3.03)	0.0543 (3.78)
Detached	0.1337 (7.50)	0.1010 (9.17)

Appendix F: Hedonic pricing equation estimation results – House structural and location characteristics (Continued)

Variable	AREA1	AREA2
Water	0.0850 (4.54)	0.0659 (2.46)
Sewer	-0.0517 (-2.68)	-0.0194 (-0.67)
House Location		
DistReading	-0.0011 (-0.47)	0.0064 (3.88)
DistPhilly	0.0043 (3.29)	-0.0065 (-6.20)
Road	-0.0217 (-1.79)	-0.0212 (-1.33)
Elev100	-0.0008 (-6.67)	-0.000007 (-0.03)
Elev100400	-0.0057 (-7.86)	-0.0008 (-0.80)
PSSA	0.0710 (6.71)	0.1102 (6.40)
Minority	-0.3905 (-3.81)	-0.1253 (-1.56)
MedInc	0.0139 (4.01)	0.0005 (0.12)
Village	-0.0247 (-2.16)	-0.0366 (-2.01)

Note: t-statistics are in parentheses

Appendix G: The OLS and IV/2SLS estimation results of the hedonic price equation for both areas

	AREA1		AREA2	
	OLS	IV/2SLS	OLS	IV/2SLS
Const1	9.8832 (64.68)	9.8766 (64.47)	9.6129 (43.83)	9.6150 (43.73)
Sold03	0.0148 (1.78)	0.0147 (1.76)	-0.0130 (-1.32)	-0.0130 (-1.32)
Sold04	-0.0054 (-0.56)	-0.0056 (-0.57)	-0.0113 (-1.01)	-0.0113 (-1.01)
Sold05	-0.0467 (-2.85)	-0.0468 (-2.85)	-0.0320 (-1.80)	-0.0320 (-1.80)
Age	-0.0398 (-21.37)	-0.0398 (-21.32)	-0.0441 (-20.46)	-0.0441 (-20.43)
Livarea	0.3832 (25.97)	0.3832 (25.94)	0.3859 (16.31)	0.3859 (16.29)
Livarea2	-0.0262 (-13.32)	-0.0262 (-13.30)	-0.0258 (-6.05)	-0.0258 (-6.04)
Lotsize	0.3669 (17.40)	0.3671 (17.38)	0.3112 (11.57)	0.3111 (11.55)
Lotsize2	-0.0697 (-13.54)	-0.0698 (-13.54)	-0.0640 (-9.18)	-0.0640 (-9.17)
Bdrms	0.0064 (1.11)	0.0064 (1.10)	-0.0118 (-1.80)	-0.0118 (-1.79)
Flbth	0.0691 (7.33)	0.0692 (7.34)	0.0615 (5.55)	0.0615 (5.54)
Hfbth	0.0370 (4.21)	0.0370 (4.21)	0.0485 (5.05)	0.0485 (5.05)
Bsmnt	0.1430 (10.54)	0.1429 (10.51)	0.1532 (7.33)	0.1531 (7.31)
Stone	0.1606 (5.67)	0.1601 (5.65)	0.1779 (6.37)	0.1780 (6.37)
Brick	0.0888 (9.68)	0.0890 (9.68)	0.0579 (5.67)	0.0579 (5.66)
Mason	0.0419 (3.45)	0.0420 (3.45)	0.0954 (5.92)	0.0954 (5.91)
Aircond	0.0287 (3.04)	0.0287 (3.03)	0.1035 (8.41)	0.1035 (8.40)
Phycn	-0.0855 (-6.72)	-0.0855 (-6.72)	-0.1107 (-8.10)	-0.1107 (-8.08)
Attic	0.0337 (2.80)	0.0340 (2.82)	0.0538 (3.69)	0.0538 (3.68)
Detached	0.1317 (7.26)	0.1318 (7.25)	0.1102 (9.17)	0.1101 (9.15)

Appendix G: The OLS and IV/2SLS estimation results of the hedonic price equation for both areas (Continued)

	AREA1		AREA2	
	OLS	IV/2SLS	OLS	IV/2SLS
Water	0.0670 (3.77)	0.0687 (3.83)	0.0626 (2.52)	0.0624 (2.51)
Sewer	-0.0416 (-2.26)	-0.0415 (-2.26)	-0.0158 (-0.59)	-0.0161 (-0.59)
DistReading	0.0009 (0.38)	0.0007 (0.27)	0.0064 (4.54)	0.0064 (4.53)
DistPhilly	0.0023 (1.74)	0.0024 (1.76)	-0.0065 (-7.52)	-0.0065 (-7.50)
Road	-0.0244 -2.12)	-0.0236 (-2.05)	-0.0208 (-1.37)	-0.0209 (-1.37)
Elev100m	-0.0006 (-5.55)	-0.0006 (-5.56)	0.00008 (-0.10)	0.00002 (-0.10)
Elev100400	-0.0054 (-7.70)	-0.0055 (-7.74)	-0.0008 (-0.77)	-0.0007 (-0.77)
Pssa	0.0652 (6.31)	0.0656 (6.34)	0.1116 (7.65)	0.1114 (7.62)
Minority	-0.2753 (-2.56)	-0.2760 (-2.56)	-0.1274 (-1.70)	-0.1269 (-1.69)
MedInc	0.0141 (4.22)	0.0142 (4.23)	0.0001 (0.02)	0.0001 (0.03)
Village	-0.0395 (-3.45)	-0.0389 (-3.39)	-0.0374 (-2.32)	-0.0375 (-2.32)
Industrial omitted				
Residential	0.2535 (4.93)	0.2503 (4.84)	0.1469 (2.84)	0.1477 (2.85)
Commercial	0.1664 (2.78)	0.1660 (2.77)	-0.0213 (-0.36)	-0.0212 (-0.36)
OS_Res	0.1304 (2.37)	0.1388 (2.48)	-0.0877 (-1.61)	-0.0890 (-1.62)
OS_Pro	0.2991 (3.95)	0.3003 (3.96)	0.0157 (0.19)	0.0158 (0.19)
Commercial omitted				
Residential	0.0872 (3.17)	0.0844 (3.05)	0.1682 (4.97)	0.1689 (4.95)
Industrial	-0.1664 (-2.78)	-0.1660 (-2.77)	0.0213 (0.36)	0.0212 (0.36)
OS_Res	-0.0359 (-1.04)	-0.0271 (-0.75)	-0.0665 (-1.74)	-0.0677 (-1.74)

Appendix G: The OLS and IV/2SLS estimation results of the hedonic price equation for both areas (Continued)

	AREA1		AREA2	
	OLS	IV/2SLS	OLS	IV/2SLS
OS_Pro	0.1328 (2.16)	0.1343 (2.19)	0.0370 (0.53)	0.0371 (0.53)
Residential omitted				
Industrial	-0.2535 (-4.93)	-0.2503 (-4.84)	-0.1469 (-2.84)	-0.1477 (-2.85)
Commercial	-0.0872 (-3.18)	-0.0844 (-3.05)	-0.1682 (-4.97)	-0.1689 (-4.95)
OS_Res	-0.1231 (-4.17)	-0.1115 (-3.42)	-0.2346 (-6.94)	-0.2366 (-6.64)
OS_Pro	0.0456 (0.80)	0.0500 (0.87)	-0.1312 (-2.00)	-0.1318 (-2.00)
OS_Res omitted				
Residential	0.1231 (4.17)	0.1115 (3.42)	0.2347 (6.94)	0.2366 (6.64)
Industrial	-0.1304 (-2.37)	-0.1388 (-2.48)	0.0877 (1.61)	0.0890 (1.62)
Commercial	0.0359 (1.04)	0.0271 (0.75)	0.0665 (1.74)	0.0677 (1.74)
OS_Pro	0.1687 (2.85)	0.1615 (2.70)	0.1035 (1.50)	0.1048 (1.51)
OS_Pro omitted				
Residential	-0.04561 (-0.80)	-0.050 (-0.87)	0.1312 (2.00)	0.1318 (2.00)
Industrial	-0.2991 (-3.95)	-0.3003 (-3.96)	-0.0157 (-0.19)	-0.0158 (-0.19)
Commercial	-0.1328 (-2.16)	-0.1343 (-2.19)	-0.0370 (-0.53)	-0.0371 (-0.53)
OS_Res	-0.1687 (-2.85)	-0.1615 (-2.70)	-0.1035 (-1.50)	-0.1048 (-1.51)
R^2	0.830		0.802	

Appendix H: Matlab code for estimating the spatial error components (SEC) model

```
function results = secres(y,x,w)

% PURPOSE: computes spatial error components model estimates
%          y = XB + e, e = W*u + v, using sparse algorithms
% ---------------------------------------------------
%  USAGE: results = secres(y,x,W)
%  where:  y = dependent variable vector
%          x = independent variables matrix
%          W = sparse spatial weights matrix
%              (standardized or non-standized)
% ---------------------------------------------------
%  RETURNS: a structure if the two error variance estimates are
%          significantly positive; otherwise return error message
%         results.meth  = 'sec'
%         results.beta  = ols.beta
%         results.tstat = ols.tstat
%         results.bstd  = ols.bstat
%         results.yhat  = ols.yhat   (nobs x 1)
%         results.resid = ols.resid  (nobs x 1)
%         results.sige  = ols.sige   scalar
%         results.rsqr  = ols.rsqr   scalar
%         results.rbar  = ols.rbar   scalar
%         results.dw    = ols.dw     statistic
%         results.nobs  = ols.nobs
%         results.nvar  = ols.nvars
%         results.y     = ols.y      (nobs x 1)
%         results.bint  = ols.bint   (nvar x2) vector with 95%
%                                    confidence intervals on beta
%---------------------------------------------------
%  Note: OLS.m is a function executing OLS regression;
%          OLSE.m is a function returning only OLS residual vector.
%         Both are available from the Econometrics Toolbox by
James
%         LeSage at http://www.spatial-econometrics.com/.
% ---------------------------------------------------
% REFERENCES: Anselin L. and Moreno R. 2003. Properties of tests
% for spatial error components,Regional Science and Urban
% Economics, 33 p600, the general method of moments (GMM)
% estimation.
```

Appendix H: Matlab code for estimating the spatial error components (SEC) model (Contiue)

```
n=size(y,1);
e=olse(y,x);
e2=e.^2;
d=full(diag(w*w'));
res=ols(e2,[ones(n,1) d]);

if res.tstat>[1.96;1.96]
  % Covariance matrix of the spatial error
  sigma=res.beta(1)*speye(n)+ res.beta(2)* w*w';
  % Cholesky decomposition
  L=chol(sigma);
  y=sparse(y);
  y=L\y;

  x=sparse(x);
  x=L\x;
  y=full(y);
  x=full(x);
  secres=ols(y,x);

elseif res.beta(1)<0
   error('sec: negative estimate of idiosyncratic error');
elseif res.beta(2)<0
  error('sec: negative estimate of spillover error');
elseif res.tstat(1)<1.96 & res.tstat(1)>0
   error('sec: insignificant estimate of idiosyncratic error');
elseif res.beta(2)<1.96 & res.tstat(2)>0
  error('sec: insignificant estimate of spillover error');
end;
```

References

Anselin L. 1988. *Spatial Econometrics: Methods and Models*. Dordretcht: Kluwer.

Anselin L. 2001. Rao's score test in spatial econometrics, *Journal of Statistical Planning and Inference*, 97: 113-139.

Anselin L. 2002. Under the hood: Issue in the specification and interpretation of spatial regression models, *Agricultural Economics*, 17: 247-267.

Anselin, L. 2005, *Exploring Spatial Data with GeoDa: A Workbook*, in https://www.geoda.uiuc.edu/documentation.php#manuals

Anselin L. and Bera A.K. 1998. Spatial dependence in linear regression models with an introduction to spatial econometrics, in Ullah A. and Giles D. (eds.) *Handbook of Applied Economics Statistics*, Marcel Dekker, New York, NY, 237-289.

Anselin L. and Moreno R. 2003. Properties of tests for spatial error components, *Regional Science and Urban Economics*, 33: 595-618.

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

Basu, S. and Thibodeau, T. 1998. Analysis of spatial autocorrelation in house prices, *Journal of Real Estate Finance and Economics*, 17: 61-86.

Bell, K.P. and Bockstael, N.E. 2000. Applying the generalized moments estimation approach to spatial problems involving microlevel data, *The Review of Economics and Statistics*, 82: 72-82.

Bolitzer, B. and Netusil, N. 2000. The impact of open spaces on property values in Portland, Oregon. *Journal of Environmental Management*, 59: 185-193

Bowden, R.J. and Turkington, D.A. 1981. A comparative study of instrumental variables estimators for nonlinear simultaneous model, *Journal of the American Statistical Association*, 76: 988-995

Bowen, W.M., Mikelbank, B.A. and Prestegaard, D.M. 2001. Theoretical and empirical considerations regarding space in hedonic housing price model application, *Growth and Change*, 32 No.4: 466-490

Can, A. and Megbolugbe, I 1997. Spatial dependence and house price index construction, *Journal of Real Estate Finance and Economics*, 14: 203-222

Durbin R.A. 1998. Spatial autocorrelation: A Primer. *Journal of Housing Economics*, 7, 304-327

Fisher, F.M. 1966. *The identification problem in econometrics*. New York: McGraw-Hill

Geoghegen J. 2002. The value of open spaces in residential land use. *Land Use Policy* 19: 91-98

Goldfeld, S.M. and Quandt, R.E. 1968. Nonlinear simultaneous equations: estimation and prediction. *International Economic Revi*ew, 9: 113-136

Goldsmith, P. 2004. Using spatial econometrics to assess the impact of swine production on residential property values. Submission to *American Journal of Agricultural Economics*

Graaff T., Florax R.J.G.M. and Wijkamp P. 2001. A general misspecification test for spatial regression models: dependence, heterogeneity and nonlinearity. *Journal of Regional Science*, 41(2): 255-276

Hepple .L.W. 2004. Bayesian model choice in spatial econometrics, in Lesage, J.P. and Pace, R.K. (eds) *Spatial and Spatiotemporal Econometrics*, Advances in Spatial Econometrics, Vol.18. Elsevier , 101-126

Hite, D., Chern W., Hitzusen F., and Randall A. 2001. Property-Value Impacts of an Environmental Disamenity: The Case of Landfills. *Journal of Real Estate Finance and Economics*, 22(2/3), 185-202.

Irwin, E.G. 2002. The effects of open space on residential property values. *Land Economics*, 78(4): 465-480

Irwin, E.G. and Bockstael, N.E. 2001. The problem of identifying land use spillovers: measuring the effects open space on residential property values. *American Journal of Agricultural Economics*, 83(3): 698-704

Irwin, E.G. and Bockstael, N.E. 2002. Interacting agents, spatial externalities and the evolution of residential land use patterns. *Journal of Economic Geography*, 2: 31-54

Kelejian, H. 1971. Two stage least squares and econometric systems linear in parameters but nonlinear in the endogenous variables. *Journal of the American Statistical Association*, 66: 373-374

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*, 17: 99-121.

Kelejian H.H. and Prucha I.R. 1999. A generalized moments estimator for the autoregressive parameter in a spatial model. *International Economics Review*, 40: 509-533.

Kelejian, H. and Prucha, I. 2004. Estimation of simultaneous system of spatial interrelated cross sectional equations. *Journal of Econometrics*, 118: 27-50

Kelejian H.H. and Robinson D.P. 1993. A suggested method of estimation method for spatial interdependent model with autocorrelatted errors, and an application to a county expenditure model. *Papers in Regional Science*, 72: 297-312.

Kelijian and Robinson 1995. Spatial correlation: a suggested alternative to the autoregressive model. in Anselin, L. and Florax R.J.G.M. (eds) *New Directions in Spatial Econometrics*, Spring-Verlag Germany, p75-96

Kelejian, H. H. and Yuzefovich, Y. 2004. Properties of Tests for Spatial Error Components: A Further Analysis in Getis A., Mur J. and Zoller H.G. (eds) *Spatial Econometrics and Spatial Statistics.* , 135-149, Palgrave MacMillan

Kim C.W., Phipps T.T. and Anselin L. 2003. Measure the benefits of air quality improvement: A spatial hedonic approach, *Journal of Environmental Economics and Management*, 45: 24-39.

Lutzenhiser M. and Netusil N. 2001. The effect of open spaces on a home's sale price. *Contemporary Economic policy*, 19(3): 291-298

LeSage J.P. 1999. *The Theory and Practice of Spatial Econometrics*. Unpublished manuscript available at: http://www.spatial-econometrics.com.

Mur J. 1999.Testing for spatial autocorrelation: Moving average versus autoregressive processes. *Environment and Planning A*, 31:137-1382.

Pace R.K. and Gilley O. 1997. Using the spatial configuration of the data to improve estimation, *Journal of Real Estate Finance and Economics*, 14: 333-340.

Ready R.C. 2004. Do landfills always depress nearby property values? Working paper, Dept. of Agricultural Economics and Rural Sociology, Pennsylvania State University

Ready R.C. and Abdalla C.W. 2005. The amenity and disamenity impacts of agriculture: Estimates from a hedonic pricing model, *American Journal of Agricultural Economics* 87(2):314-326

Smith V.K., Poulos C. and Kim H. 2002. Treating open space as urban amenity. *Resource and Energy Economics* 24: 107-129

Sneek J.M. and Rietveld P. 1997. On the estimation of the spatial moving average model. Tinbergen Institute Discussion Paper, 97-049/4.

Tyrvainen L. and Miettinen A. 2002. Property prices and urban forest amenities. *Journal of Environmental Economics and Management* 39: 205-223

Trivezg F.J. and Mur J. 2004. Some proposals for discriminating between spatial process, in Getis A., Mur J. and Zoller H.G. (eds) *Spatial Econometrics and Spatial Statistics*, 150-175, Palgrave MacMillan

Wonnacott R.J. and Wonnacott T.H. 1979. *Econometrics*, 2nd ed. John Wiley&Sons

VITA
LI WANG

Date of Birth: Nov. 29, 1974, P.R.China

Education:

Ph.D. Department of Agricultural Economics and Rural Sociology. The Penn State University. U.S.A. 2006. Specialization: Environmental Economics and Quantitative Research

M.A. College of Business Administration. Shanghai University of Finance and Economics. P.R. China. 1998

B.A. College of Business Administration. Shanghai University of Finance and Economics. P.R. China. 1995

Employment:

Research Assistant. Department of Agricultural Economics and Rural Sociology. The Penn State University. U.S.A. 2002-2005.

Research Work:

Wang. L. and Ready. R. 2005 Spatial Econometric Approaches to Estimating Hedonic Property Value Models. Paper presented at AAEA annual meeting, Providence, RI

Affiliation:

American Agricultural Economics Association