

Article title

A Spatio-Temporal Analysis of Single-Family Home Values in the Greater Orlando Region

Author names

Susmita Sarker*

Department of Civil, Environmental and Construction Engineering

University of Central Florida, Orlando, Florida, United States, 32816

Email: susmita.sarker@ucf.edu

Sudipta Dey Tirtha

Department of Civil, Environmental and Construction Engineering

University of Central Florida, Orlando, Florida, United States, 32816

Email: sudiptadey.tirtha@ucf.edu

Naveen Eluru

Department of Civil, Environmental and Construction Engineering

University of Central Florida, Orlando, Florida, United States, 32816

Email: naveen.eluru@ucf.edu

Acknowledgement

The authors are grateful to Florida Department of Revenue (FDOR), Orange County Property Appraiser (OCPA), Osceola County Property Appraiser, Seminole County Property Appraiser (SCPA), Bureau of Economic Analysis (BEA), Florida Department of Education (FLDOE), School District of Osceola County, Florida, Orange County public schools (OCPS), Seminole County Geographic Information System (GIS), Florida Department of Transportation (FDOT), and Central Florida Regional Planning Model (CFRPM) for providing access to their datasets. The authors also appreciate the insightful feedback from two anonymous reviewers on a previous version of the paper.

***Corresponding author**

A Spatio-Temporal Analysis of Single-Family Home Values in the Greater Orlando Region

ABSTRACT

Home equity represents the largest component of wealth for households in the US. The accurate estimation of home values is important for different sectors of the economy. The current study develops a new method - grouped generalized ordered probit (GGOP) model - for examining the impact of independent variables on single-family home values. The proposed methodology is developed using home property value data for the greater Orlando region from 2013 to 2022. The host of independent variables considered include structural characteristics, demographic characteristics, land use characteristics, transportation infrastructure characteristics, and temporal characteristics. The model estimation process includes the estimation of a GGOP model, spatial lag GGOP and spatial error GGOP models. The spatial lag model offers significantly improved fit compared to other models. The independent variables such as parcel area, living area, number of bathrooms, population, median income, number of jobs, school quality, proximity to downtown, distance to SunRail stations and year of analysis significantly affect single-family home values in the Greater Orlando region. The model results were validated using data not considered for estimation. An elasticity analysis was also conducted to offer insights on independent variable impact magnitudes. The model findings can be employed to quantify property tax revenues for new developments and homeowner/ community losses in the event of natural disasters.

Keywords: Home values, Single-family, Spatial econometrics, Grouped generalized ordered probit, Weight matrix.

1 INTRODUCTION

1.1 Background Motivation

Housing location, and neighborhood quality have strong impact on humans' physical, emotional, and social well-being (Palimaru et al., 2023; Streimikiene, 2015; "The Development of the World Health Organization Quality of Life Assessment Instrument (the WHOQOL)," 1994). In the US, 62% of the households live in a home they own, and home equity represents the largest component of wealth for households (*Value of Assets U.S. Households Own and the Debt They Carry* | Pew Research Center, n.d.). For individuals to set list prices, tax authorities to appraise taxes, banks to determine eligibility for mortgages, and investors to make educated selections, home value prediction is very essential (Liu, 2013). The determination of home values is important for long-term residential location and daily travel pattern modeling. These components form a critical component of integrated land use model frameworks and influence long range transportation planning models. Household residential location decisions have an outsized impact on transportation demand and supply needs. With the exacerbation of global climate crisis, the incidence of disasters in the future is likely to increase (*Is Climate Change Increasing the Risk of Disasters?*, n.d.). Accurate models of home values can contribute to determining the losses to households and communities from such disasters. The estimation modules can also provide a proactive mechanism to conduct a benefit cost analysis of various climate change alleviations strategies (such as building a levee to protect from flood water) on reducing disaster impacts. Another area of application is toward addressing the housing affordability crisis in the US (Luque et al., 2019). The accurate estimation of home values is critical to identifying urban pockets that are unaffordable proactively (*Florida's Affordable Housing Crisis*, n.d.). This is also helpful in understanding the incentive structures to be considered by policy makers to increase

inventory for affordable homes. Finally, accurate estimation of home values can also contribute to identifying existing inequities across the urban regions by different attributes including spatial location, ethnicity and education attainment.

In this paper, we contribute to home value estimation along two directions. First, we develop a novel modeling framework for estimating home values. The most common approach for home value estimation employs a linear regression model structure with different advanced variants (such as linear mixed models, spatial error or spatial lag models). While these models are versatile, they implicitly assume that the impact of the independent variables on the dependent variable is linear in nature. In this study, we apply an ordered discrete model structure that allows for non-linear variable interactions while allowing for home value predictions in continuous scale (similar to traditional linear regression models). The consideration of non-linear variation allows us to reduce misspecification bias in model estimation and ensure that the accurate impact of independent variables is retrieved from the models. Second, the research is conducted in the Central Florida region, home to Orlando metro area, a fast-growing region affected by several changes/challenges including growing diverse population, rapidly increasing home building, significant transportation investments, climate change impacts (flooding and intensified hurricanes), housing affordability challenges and provides an ideal testbed for home value analysis. The analysis is conducted using home price value data for Single-Family homes from Florida Department of Revenue (FDOR), for the years 2013 through 2022 focusing on Orange, Osceola and Seminole counties. The home value data is augmented with rich set of independent variables including structural characteristics (such as parcel area, total living area and number of baths), demographic characteristics (such as population, median income and job density), land use characteristics (such as proportion of land use, network distance to downtown,

school quality and land mix), transportation infrastructure characteristics (such as distance to nearest highway, distance to nearest SunRail stations and sidewalk density) and temporal characteristics (years since 2013). The results from our models are reinforced with elasticity effects to quantify the impact of various independent variables considered in the model development.

1.2 Earlier Work

Given the importance of determining home values, the subject has received significant attention in various domains including transportation, land use and urban form, environment, marketing and public policy (Allen et al., 2015; Chernobai & Ma, 2022; Choi et al., 2023; Del Negro & Otrok, 2007; J. Kim et al., 2024; Koster & Rouwendal, 2012; J. Wu et al., 2014). A detailed paper by paper review of earlier work on home value/price evaluation is beyond the scope of this paper (see M. Boyle & Kiel, 2001; Geerts et al., 2023; L. Krause & Bitter, 2012 for literature reviews). In this paper, we provide a concise review of earlier research efforts on home price estimation of single-family homes focusing on study coverage areas, home price analysis resolution, methodological approaches employed for analysis, independent variables considered in the analysis and summary of the variable impacts identified.

The ubiquitous relevance of home value variation can be garnered based on research conducted across a wide range of countries including USA, Canada, United Kingdom, Netherlands, China, and Australia (for example A. Boyle et al., 2014; Caprioli et al., 2023; Choi et al., 2023; Levkovich et al., 2016; Liu, 2013; Thompson et al., 2023; Yang et al., 2023). The home value analysis has been explored generally at the disaggregate resolution using individual home value data (Choi et al., 2023; Helbich et al., 2014; J. Kim et al., 2024;

Levkovich et al., 2016). Home value variation has also been studied at various aggregate resolutions such as residential community (Wen et al., 2017), block group level (Feng & Humphreys, 2012), Metropolitan Statistical Area (MSA) level (Cohen et al., 2016), regional level (Holly et al., 2010) and state (Gabauer et al., 2024).

Several analysis methods have been adopted to estimate the home value with respect to contributing factors using statistical and machine learning approaches. Statistical methods considered include hedonic price model (Allen et al., 2015; A. Boyle et al., 2014; Chernobai & Ma, 2022; Peng & Knaap, 2023; Yang et al., 2023), repeat-sales models (Levkovich et al., 2016; Peng & Knaap, 2023), quantile regression models (Yang et al., 2023), geographically and/or temporally weighted regression (GTWR) model (Huang et al., 2010), structural equation model (J. Kim et al., 2024), agent based modelling (ABM) (Zhuge et al., 2016). Hedonic price models include ordinary least squared regression (OLS) (Helbich et al., 2014; Huang et al., 2010; Rey-Blanco et al., 2024), multilevel price model (Choi et al., 2023; Yang et al., 2023), spatial lag and spatial error models (Allen et al., 2015; Li et al., 2015; Wen et al., 2017), spatial autoregressive (SAR) model (Helbich et al., 2014), spatiotemporal autoregressive (STAR) model (Liu, 2013), spatial autocorrelation models (Basu & Thibodeau, 1998; Rey-Blanco et al., 2024), difference-in-differences (DID) estimators (Levkovich et al., 2016; Wen et al., 2017) and time and space indicators in regression analysis (Golub et al., 2012; Turnbull et al., 2018). To capture the spatio-temporal variance in home value, the methods employed include spatial weight matrix using inverse distance (Allen et al., 2015), contiguity spatial weight matrix (Li et al., 2015), and semi variogram (Basu & Thibodeau, 1998). Machine learning approaches employed include lasso and elastic-net, recursive partitioning tress (Rey-Blanco et al., 2024), decision tree and gradient boosted tree (Soltani et al., 2021), random forest and other variants (Rey-Blanco et al., 2024).

Earlier studies on single-family home value considered structural characteristics, school quality, neighborhood/built environmental characteristics, and transportation infrastructure attributes. The structural characteristics such as the number of baths, living area are found to positively affect home value (Li et al., 2015) and building age is likely to have a negative impact (Golub et al., 2012; Li et al., 2015). In earlier work, school quality has been found to significantly affect residential home values. School quality was represented based on test scores and was considered for primary and secondary schools (Turnbull et al., 2018; Wen et al., 2017). The distance of the property to various landmarks was found to significantly influence home values such as distance to city hall/ central business district (CBD) (J. Kim et al., 2024; Li et al., 2015), and newly constructed highways (Chernobai et al., 2011). The studies also found evidence of the negative impact of highways due to increase in traffic and noise (Allen et al., 2015). Walking infrastructure (sidewalk presence), walkability and walk scores have been found to affect home values (Li et al., 2015) . Walkability in the neighborhood has been found to have a varied effect on home values ranging from positive influence (Chernobai & Ma, 2022) to insignificant impact (A. Boyle et al., 2014). Transit infrastructure and distance to transit is observed to impact home values (Peng & Knaap, 2023). The general impacts for residential home values are positive but some studies have identified reductions for locations very close to metro and light rail stations.

Earlier studies have also highlighted the impact of temporal and spatial factors on home price values. It is likely that proximally close locations are likely to share amenities, property and other taxes, and structural commonalities and thus are most likely priced similarly (Basu & Thibodeau, 1998; Liu, 2013). The features of the neighborhood and changes in the surrounding land use also contribute towards the large variation in home values (Koster & Rouwendal, 2012;

Yang et al., 2023). Finally, over time, homes sold in the same time period are likely to exhibit similar price trends (Liu, 2013) .

1.3 Contributions of the current study

The literature review highlights the significant amount of research conducted for home price modeling. While several efforts have been undertaken, majority of the studies from econometric modeling approaches employ linear/log-linear assumption in examining the relationship between home value and independent variables. It is possible to address non-linearity in variable impacts using machine learning or artificial intelligence approaches for home value modeling. However, these approaches have limited interpretability. In this study, we propose a framework that will naturally accommodate for independent variable non-linearities on home price using a grouped generalized ordered probit (GGOP) model for analysis (Rahman et al., 2019; Tirtha et al., 2023, 2024). The proposed model provides prediction similar to a linear regression while relaxing the linear in parameter restriction on the independent variables (Rahman et al., 2019; Tirtha et al., 2023, 2024). Introducing non-linearity in an econometric approach enables users to clearly quantify the impact of independent variables aiding in policy analysis.

In our proposed framework, the propensity of home price is directly linked to observed values, allowing us to estimate the variance of the property price. The variance terms can also be parameterized to allow for the impact of independent variables accounting for heteroscedasticity. The framework is also flexible to account for any potential alternative specific effects. The proposed GGOP kernel is also embedded within a spatial framework to accommodate potential spatial and temporal effects. **The spatial variants considered include spatial lag and spatial error models.** Within the spatial grouped generalized ordered probit (SGGOP), several alternative

weight matrices to reflect spatial and temporal differences (i.e., we recognize records that are spatially and temporally closer have a larger impact on the records) are generated and tested in the model based on spatial and temporal distances between different observations. In earlier research efforts, the prevalent approach for weight matrix consideration was to assume a specific parameterized form and run the model for only that one form. In our study, we build on earlier research by systematically testing the model performance by varying the structure of the weight matrix. We also test the model for different distance and time thresholds. We selected the model that offered the best model fit as the chosen weight matrix. The proposed spatial model is estimated employing a log-composite likelihood method that serves as a computationally feasible approach compared to the traditional likelihood model (see Bhat et al., 2010). The SGGOP model is estimated drawing home price value data from Florida Department of Revenue (FDOR), for the years 2013 through 2022. In addition to the methodological approach, the current paper generates a comprehensive set of independent variables to explain the home value variations including: a) structural characteristics, b) demographic characteristics, c) land use and neighborhood characteristics d) transportation infrastructure characteristics and e) temporal characteristics. The model estimates are augmented with an elasticity analysis exercise to illustrate the impact of different independent variables on home prices.

The rest of the paper is organized as follows: Section 2 provides a mathematical overview of the model framework. A discussion of the data compilation procedures and an overview of the data used in this study are presented in Section 3. Section 4 presents the model fit details and discusses the parameter interpretations. Section 5 and section 6 provide the model validation measures and discuss the findings of the elasticity analysis, respectively. Finally, in section 7 the

summary of the study is presented, with a discussion of potential applications where the proposed model can be applied and other directions for future research.

2 ECONOMETRIC METHODOLOGY

In this section, we first present the details of the grouped generalized ordered probit (GGOP) model employed to model home price values without considering any spatial dependencies between the parcels. In the subsequent sub-sections, we present the formulations of spatial lag and spatial error GGOP models, respectively. Finally, we present model estimation procedure.

2.1 Grouped Generalized Ordered Probit Model

Let k ($k = 1, 2, \dots, K$) be an index to represent single-family residential parcels, t ($t = 1, 2, 3, \dots, T = 10$) represent the different years and j ($j = 1, 2, 3, \dots, J = 15$) be an index to represent ordered parcel home values. For ease of presentation, we represent unique combination of k and t using q ($q = 1, 2, 3, \dots, Q$). We consider fifteen categories for the home value and these categories are: Bin 1= $\leq$ \$90,000, Bin 2= $\Rightarrow$ \$90,000-\$125,000, Bin 3= $\Rightarrow$ \$125,000-\$150,000, Bin 4= $\Rightarrow$ \$150,000-\$175,000, Bin 5= $\Rightarrow$ \$175,000-\$200,000, Bin 6= $\Rightarrow$ \$200,000-\$225,000, Bin 7= $\Rightarrow$ \$225,000-\$250,000, Bin 8= $\Rightarrow$ \$250,000-\$275,000, Bin 9= $\Rightarrow$ \$275,000-\$300,000, Bin 10= $\Rightarrow$ \$300,000-\$325,000, Bin 11= $\Rightarrow$ \$325,000-\$350,000, Bin 12= $\Rightarrow$ \$350,000-\$400,000, Bin 13= $\Rightarrow$ \$400,000-\$450,000, Bin 14= $\Rightarrow$ \$450,000-\$500,000 and Bin 15= $\Rightarrow$ \$500,000)¹.

¹ The reader would note that we tested different price bins in our analysis including 12 bins, 15 bins and 18 bins. The results across the three alternatives provided similar results (presented in APPENDIX A). Given the stability across the different bins, we employed 15 bins in our analysis.

The continuous latent propensity in grouped ordered probit framework for an observation q can be written as follows:

$$y_q^* = \boldsymbol{\alpha}' \mathbf{x}_q + \varepsilon_q, \quad y_q = j \text{ if } \psi_{j-1} < y_q^* < \psi_j \quad (1)$$

In equation (1), y_q^* is the latent (continuous) propensity for home value of parcel k for the year t . This latent propensity y_q^* is mapped to the price category j by the ψ thresholds, in the usual ordered-response modeling framework. In our case, we consider $J=15$ and thus the 16 known threshold values are as follows: $-\infty$, \$90,000, \$125,000, \$150,000, \$175,000, \$200,000, \$225,000, \$250,000, \$275,000, \$300,000, \$325,000, \$350,000, \$400,000, \$450,000, \$500,000 and $+\infty$. $\mathbf{x}_q$ is a matrix of exogenous attributes that influence home value (including the constant); $\boldsymbol{\alpha}$ is the vector of coefficients corresponding to the attributes. Further, ε_q is an idiosyncratic random error term assumed independently normally distributed with mean 0 and variance λ^2 .

The variance vectors for price values are parameterized as a function of independent variables as follows: $\lambda_q = \exp(\boldsymbol{\theta}' \mathbf{m}_q)$ where $\mathbf{m}_q$ is a matrix of attributes that influence the variance of home value and $\boldsymbol{\theta}$ is the vector of coefficients corresponding to the attributes. The parameterization allows for the variance to be different across the parcels accommodating for heteroscedasticity. Finally, to allow for alternative specific effects, we also introduce threshold specific deviations in the model as $\rho_j = \boldsymbol{\tau}_j' \mathbf{l}_q$. Here, $\boldsymbol{\tau}_j$ is a vector of coefficients and $\mathbf{l}_q$ is a set of independent variables including constant. If ρ_j is positive, the threshold demarcating alternatives $j-1$ and j shifts to the left and the probability of the lower-level category (higher level category) decreases (increases).

The probability for a parcel q to have price in category j is given by:

$$Pr(y_q = j) = \Lambda \left[\frac{\Psi_j - (\alpha' x_q + \rho'_j)}{\lambda_q} \right] - \Lambda \left[\frac{\Psi_{j-1} - (\alpha' x_q + \rho'_{j-1})}{\lambda_q} \right] \quad (2)$$

where $\Lambda(\cdot)$ is the cumulative standard normal distribution.

2.2 Spatial Lag GGOP Model

The spatial lag GGOP model follows the structure of the spatial lag linear model with minor adaptations for the discrete ordered categories. The spatial lag formulation includes spatial correlation in the latent propensity of property price presented in Equation 1 as follows (Castro et al., 2013; Tirtha et al., 2024):

$$y_q^* = \delta \sum_{q'=1}^Q w_{qq'} y_{q'}^* + \alpha' x_q + \varepsilon_q, y_q = j \text{ if } \psi_{j-1} < y_q^* \leq \psi_j \quad (3)$$

Where, $w_{qq'}$ is an element of an exogenously defined distance-year difference based space and time weight matrix $\mathbf{W}$ calculated based on locations and year of analysis for parcels k and k' (with $w_{qq} = 0$ and $\sum_{q'} w_{qq'} = 1$), and δ ($0 < \delta < 1$) is the spatial autoregressive parameter (please see Tirtha et al. 2024 for detailed procedure). In space-time weight matrix $\mathbf{W}$, we employ different functional forms of $w_{qq'}$ including $1/\sqrt{\text{distance} * \text{year}}$, $1/(\sqrt{\text{distance}} * \text{year})$, $1/(\text{distance} * \sqrt{\text{year}})$, $1/(\text{distance} * \text{year})$, $1/(\text{distance}^2 * \text{year})$, $1/(\text{distance}^3 * \text{year})$ and $1/(\text{distance} * \ln(\text{year}))$. Further, we restrict 3 groups of elements of $\mathbf{W}$ to be zero: a) diagonal elements to avoid self-inclusion of the observations ($w_{qq} = 0$), b)

off-diagonal elements for the same parcel (as we have repeated records for the parcels across the years), and c) any element for future time points where year difference is negative. Then, we perform row normalization of $\mathbf{W}$ to restrict $\sum_{q'} w_{qq'} = 1$.

Finally, to restrict δ between 0 and 1, we represent δ using a function: $\frac{e^\zeta}{1+e^\zeta}$ and estimate the parameter ζ . The latent price propensity presented in Equation 3 can be re-written using vector notation as follows:

$$\mathbf{y}^* = \delta \mathbf{W} \mathbf{y}^* + \mathbf{x} \boldsymbol{\alpha} + \boldsymbol{\varepsilon} \quad (4)$$

Where $\mathbf{y}^* = (y_1^*, y_2^*, \dots, y_Q^*)'$, $\mathbf{x} = (x_1, x_2, \dots, x_Q)'$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_Q)'$ are vector notations representing latent price propensities, exogenous attributes influencing home value, and idiosyncratic random errors for all Q units, respectively.

Now, the Equation 4 can be re-written as follows (Castro et al., 2013):

$$\mathbf{y}^* = \mathbf{S}(\mathbf{x} \boldsymbol{\alpha} + \boldsymbol{\varepsilon}) \quad (5)$$

where $\mathbf{S} = [\mathbf{I}_Q - \delta \mathbf{W}]^{-1}$ is a $(Q \times Q)$ matrix and $\mathbf{I}_Q$ is an identity matrix of size Q . The vector $\mathbf{y}^*$ is multivariate normally distributed as, $\mathbf{y}^* \sim MVN_Q(\mathbf{B}_{lag}, \boldsymbol{\Sigma}_{lag})$. We represent $\mathbf{B}_{lag}$ and $\boldsymbol{\Sigma}_{lag}$ as follows

$$\mathbf{B}_{lag} = \mathbf{S} \mathbf{x} \boldsymbol{\alpha} \text{ and } \boldsymbol{\Sigma}_{lag} = \mathbf{S} \mathbf{I}_Q \mathbf{S}' \quad (6)$$

2.3 Spatial Error GGOP Model

The spatial error GGOP model follows the structure of the spatial error linear model with adaptations necessary for the discrete ordered model. In spatial error model formulation, continuous latent propensity is expressed as follows (Castro et al., 2013; Tirtha et al., 2024):

$$y_q^* = \delta \sum_{q'=1}^Q w_{qq'} \varepsilon_{q'} + \alpha' x_q + \varepsilon_q, y_q = j \text{ if } \psi_{j-1} < y_q^* \leq \psi_j \quad (7)$$

The latent price propensity presented in Equation 7 can be re-written as follows:

$$\mathbf{y}^* = \delta \mathbf{W} \boldsymbol{\varepsilon} + \mathbf{x} \boldsymbol{\alpha} + \boldsymbol{\varepsilon} \quad (8)$$

Where the different matrices $\mathbf{x}$, $\mathbf{W}$, and $\boldsymbol{\alpha}$ follow earlier definitions.

We can re-write Equation 8 as follows:

$$\mathbf{y}^* = \mathbf{x} \boldsymbol{\alpha} + \mathbf{S} \boldsymbol{\varepsilon} \quad (9)$$

where $\mathbf{S} = [\mathbf{I}_Q - \delta \mathbf{W}]^{-1}$ as defined earlier.

The vector $\mathbf{y}^*$ is multivariate normally distributed as, $\mathbf{y}^* \sim MVN_Q(\mathbf{B}_{error}, \boldsymbol{\Sigma}_{error})$.

We represent $\mathbf{B}_{error}$ and $\boldsymbol{\Sigma}_{error}$ as follows:

$$\mathbf{B}_{error} = \mathbf{x} \boldsymbol{\alpha} \text{ and } \boldsymbol{\Sigma}_{error} = \mathbf{S} \mathbf{S}' \quad (10)$$

2.4 Estimation Procedure

The vector of parameters to be estimated in both spatial lag and spatial error GGOP model is $\boldsymbol{\gamma} = (\boldsymbol{\alpha}', \boldsymbol{\rho}_j', \lambda, \zeta)$. In this study, as maximum likelihood estimation is infeasible, we employ the composite maximum likelihood approach and adopt the pairwise method to compute the log-

composite likelihood as follows (see Castro et al., 2013; Tirtha et al., 2024 for similar formulation):

$$\begin{aligned}
L_{CML}(\mathbf{Y}) &= \prod_{q=1}^Q \prod_{q'=1, q' \neq q}^Q \Pr(y_q = z_q, y_{q'} = z_{q'}) \\
&= \prod_{q=1}^Q \prod_{q'=1, q' \neq q}^Q [\Phi(\varphi_q, \varphi_{q'}, \nu_{qq'}) - \Phi(\varphi_q, \mu_{q'}, \nu_{qq'}) - \Phi(\mu_q, \varphi_{q'}, \nu_{qq'}) \\
&\quad + \Phi(\mu_q, \mu_{q'}, \nu_{qq'})]
\end{aligned} \tag{11}$$

Where, $\varphi_q = \frac{\psi_{q,z} - ([\mathbf{B}]_q + \rho'_{q,z})}{\sqrt{\lambda_q * [\boldsymbol{\Sigma}]_{qq}}}$, $\mu_q = \frac{\psi_{q,z-1} - ([\mathbf{B}]_q + \rho'_{q,z-1})}{\sqrt{\lambda_q * [\boldsymbol{\Sigma}]_{qq}}}$, $\nu_{qq'} = \frac{[\boldsymbol{\Sigma}]_{qq'}}{\sqrt{[\boldsymbol{\Sigma}]_{qq} * [\boldsymbol{\Sigma}]_{q'q'}}$ and

$z_q(z_1, z_2, \dots, z_Q)$ is the observed price level.

In the equation above, $[\mathbf{B}]_q$ represents the q th element of vector $\mathbf{B}$, and $[\boldsymbol{\Sigma}]_{qq'}$ represents the qq' th element of matrix $\boldsymbol{\Sigma}$. In computing marginal likelihood function presented in Equation 11, we need to calculate $Q(Q - 1)$ numbers of joint probabilities. In Equation 11, $\nu_{qq'}$ represents correlation parameter in bivariate normal cumulative density function which is stronger for observations in proximity in terms of time and space. $\nu_{qq'}$ is considerably small for observations with larger distance and year difference. In this study, we assume that spatial correlations are present within a distance band and a time threshold. Based on spatial distribution of the parcels, we select 5 miles as the distance band and 6 years as the time threshold for our study. Therefore, we can re-write Equation 11 as follows:

$$L_{CML}(\mathbf{Y}) = \prod_{q=1}^Q \prod_{q'=1, q' \neq q}^Q \Pr(y_q = z_q, y_{q'} = z_{q'}) \tag{12}$$

$$\begin{aligned}
&= \prod_{q=1}^Q \prod_{q'=1, q' \neq q}^Q [\Phi(\varphi_q, \varphi_{q'}, R_{qq'} \nu_{qq'}) - \Phi(\varphi_q, \mu_{q'}, R_{qq'} \nu_{qq'}) \\
&\quad - \Phi(\mu_q, \varphi_{q'}, R_{qq'} \nu_{qq'}) + \Phi(\mu_q, \mu_{q'}, R_{qq'} \nu_{qq'})]
\end{aligned}$$

Where, $R_{qq'} = 1$ if $d_{qq'} \leq 5$ miles and $0 < t_{qq'} \leq 6$ years, 0 otherwise

In above Equation 12, $R_{qq'}$ is a dummy variable indicating the presence of spatial correlation between a pair of parcels. $d_{qq'}$ and $t_{qq'}$ represent distance in miles and time difference in years between observations, q and q' . Finally, covariance matrix of the parameters is estimated by the inverse of Godambe's (Godambe, 1960) sandwich information matrix (see Bhat et al., 2010; Castro et al., 2013 for the details of covariance matrix).

3 DATA COMPILATION AND DESCRIPTIVE ANALYSIS

3.1 Dependent Variable

The home value data is drawn from Florida Department of Revenue (FDOR). Home values from 2013 to 2022 from three counties - Orange, Osceola and Seminole – are processed to prepare the final dataset. The data compiled included 5,316,952 data records for single-family land parcels for the ten-year period. After removing data records based on data availability for other independent variables, 3,763,696 records were retained. The home value was gathered from parcel data based on the tax assessment value of the parcel (not the sale value). While these assessments can vary from the actual sale price in some transactions, the widespread availability of the assessment data across the study region prompted us to use this variable. Further, given the

role of tax assessment in revenue generation for counties (as property taxes), accurately predicting assessed values is of great interest to state and county agencies².

As described in section 2, the home values are categorized into 15 bins. The share of the 15 categories considered in our analysis is presented in Figure 1. From Figure 1, we can see that a large share of the single-family homes in our study region range between \$150,000 - \$275,000 accounting for about 45% of the homes.

3.2 Independent Variables

We generate a comprehensive set of independent variables to explain the home value variations including a) structural characteristics, b) demographic characteristics, c) land use and neighborhood characteristics d) transportation infrastructure characteristics and e) temporal factors. The structural characteristics information is sourced from the Orange County Property Appraiser (OCPA), Osceola County Property Appraiser, Seminole County Property Appraiser (SCPA) and Florida Department of Revenue (FDOR). The variables include parcel area, total living area and the number of baths. Demographic information was sourced from United States Census Bureau and Bureau of Economic Analysis (BEA). The variables compiled include population, median income and job density. For some block groups the median income information was missing. The records with income missing for several years were removed. For

² If data on home sales is widely available in a study region, it would be beneficial to develop models with this information. The primary reason for adoption of tax assessment data is its ease of availability across the study region (and several other regions). The widespread availability of these data will allow for easy transferability of the proposed model across jurisdictions.

cases where income was missing intermittently, we imputed the income using a growth factor approach using income data for previous years. The imputed data was only a small share of the overall records (< 0.75%).

The land use and neighborhood characteristics information are sourced from Florida Department of Revenue, Florida Department of Education (FLDOE), School District of Osceola County, Florida, Orange County public schools (OCPS), Seminole County Geographic Information System (GIS) and Google Maps. The variables include the proportion of commercial, vacant, single-family residential, multi-family residential and industrial areas, school quality in terms of school grades, network distance to the nearest downtown, land use mix and distances among the parcels. The school quality variable was generated for each parcel by mapping the school zone areas from the three counties to the parcel data. The exercise allowed us to identify the assigned elementary, middle and high school for each parcel. For each of these schools, the corresponding grades from Florida Department of Education were compiled. The letter grades assigned by the Department of Education were converted to numerical values (A=4, B=3, C=2 and D=1, F=0) and used to create a single score for the three schools. Based on the distribution of the school quality variable, we created three specified grade ranges: 1.00 to 2.99 as fair, 3.00 to 3.66 as good and 3.67 to 4.00 as very good.

The transportation infrastructure characteristics are sourced from the Florida Department of Transportation (FDOT), ArcGIS Hub and Central Florida Regional Planning Model (CFRPM). The variables include distance to the nearest highway, distance to the nearest SunRail stations and sidewalk density. The temporal factors considered include the number of years passed from 2013 (2013 being the base year).

For the study area, the distances to the nearest downtowns were considered for Orlando city center, Sanford city center, Kissimmee city center, Melbourne city center, Deland city center and Cocoa city center. The SunRail stations' distance information was prepared for 12 stations from year 2014 to 2017, and for 16 stations including the new 4 stations from year 2018 to 2022. The descriptive statistics for the independent variables are presented in **Table 1**.

4 MODEL ESTIMATION AND RESULTS

4.1 Model estimation strategy

The model estimation process started with the simple GGOP model estimation. After finalizing the GGOP model, we developed the spatial error and spatial lag variants. In spatial models, based on the econometric methodology outlined, the reader would note that as the sample size increases the number of probability computations required for model estimation increase non-linearly. The computational burden is particularly high for computing the variance covariance matrices needed for inference. Given the computational complexity involved in estimating spatial models, we employed a sample of 3000 records for our analysis³. At the same time, we recognize that sampling only 3000 records could potentially result in a skewed and unrepresentative sample for our population. Hence, we systematically test across various samples to ensure that the variables identified in our sample are representative across the population sample. We separate the prepared dataset into two groups (a) an estimation sample and (b) a validation sample. We conduct this parameter stability analysis using 25 samples of 3000 records

³ We tested the stability of the model results from 3000 sample with larger samples including 6000, 12000 and 18000 records. The comparison results are presented in the APPENDIX B. Given the relative stability across the different sample sizes and the computational resources needed for larger samples, we adopted the 3000 sample for subsequent analysis.

from the estimation sample and compare the model parameters across these samples using a modified Wald test.

We generate the Wald test t-statistic using the following equation (Parvez et al., 2023) :

$$t = abs\left[\frac{(sample\ parameter-population\ benchmark)}{\sqrt{SE_{sample}^2+SE_{population}^2}}\right] \quad (13)$$

The test is conducted for each parameter in the model using Equation 13. The population benchmark and the standard error of the population are the average of the parameters and the average of the standard errors respectively, across the 25 samples. The Wald test variations for all the parameters across 25 samples are presented in Figure 2. Of the 550 (22 parameters x 25 samples) test values generated from our comparison, only 9 violate the 95% confidence level value (1.96). Thus, the Wald t-test analysis confirms the presence of significant parameter stability across the models for the sample considered in our analysis.

4.2 Model specification and overall measure of fit

For various categorical variable models, log-likelihood and BIC measures are used as criteria for model fitness comparison. For the estimation of Spatial GGOP model variants, it is infeasible to employ the maximum likelihood approach (see Castro et al., 2013). Hence, we resort to using composite maximum likelihood (CML) estimation. The models estimated using CML are compared using measures analogous to log-likelihood and BIC. These measures are referred to as log-composite likelihood and modified BIC (Tirtha et al., 2024). Thus, when we are comparing two models estimated using CML model fitness is compared using these two measures. The modified BIC is calculated using the following equation:

$$\text{Modified BIC} = -\frac{LL}{q-1} + \text{No. of parameters} * \ln(q) \quad (14)$$

For the spatial GGOP model estimation, we consider different weight matrix (W) structures. In the computation of W, we allow for spatial dependency within a five-mile distance⁴ and temporal dependency within a six-year time period. The reader would note that time difference is computed as the year difference +1 to allow for dependencies across parcels in the same year. The appropriate W for the model system was determined based on the log-composite likelihood for different weight matrix formulations presented in **Table 2**. From the various W formulations tested we can observe that the W matrix with the $1/(\text{distance}^2 * (\text{year difference} + 1))$ structure offers the best fit. The log-composite likelihood and the modified BIC values using Equation 14 for simple GGOP, and the best spatial lag and spatial error models are presented in **Table 3**. From **Table 3**, we can conclude that spatial lag model offers the best fit for our data. Going forward, we only discuss the spatial lag GGOP model.

4.3 Estimation results

The final model results with statistically significant variables at 95% confidence level are presented in **Table 4**. The variable interpretations are presented by variable group. The reader will note that a positive (negative) value indicates a positive (negative) influence on home value.

⁴ The five-mile distance was selected based on a comparison of model fit across different mileages. The results from the comparison exercise indicated that model fit stabilizes around the five-mile distance and beyond the five-mile distance there is no improvement (see results presented in APPENDIX C). Thus, a five-mile distance was selected to minimize computational overhead for model estimation.

4.3.1 Structural characteristics

The impact of structural characteristics on single-family home values is along expected lines. The area of the property parcel is positively correlated with home value. Similarly, an increase in living area is associated with an increase in home value. The number of baths, as expected, has a positive impact on home values (please see Chernobai & Ma, 2022; Choi et al., 2023; Li et al., 2015 for similar findings for living area and number of baths).

4.3.2 Demographic characteristics

Among demographic characteristics, the parcel block group population and median income and county job density influence home values. In our study region, we observe that an increase in population in the block group is associated with a reduction in home values. Given the preference for single-family homes in Florida, the presence of high population in the block group is indicative of a presence of apartments and other housing units. The impact of median income for the block group on home values is as expected. The parcels located in block groups with higher median income are likely to be of higher value. An increase in job density around the parcel is associated with an increase in the price of the property. The result is intuitive as increasing job density can contribute to higher demand for housing in the vicinity contributing to increase in home values (see J. Kim & Long, 2024). It is important to recognize that while the parameters of the independent variables remain constant across the region, the actual impact on the propensity is realized as the product of the coefficient and the independent variable. Thus, as median income reduces in a block group (say in rural areas), the impact of income on propensity reduces for this variable.

4.3.3 Land use and neighborhood characteristics

Several land use and neighborhood characteristics were considered in our analysis. The presence of industrial areas in the block group is negatively associated with single-family home values. The areas with industrial infrastructure are typically associated with increased noise, undesirable traffic movements (such as large trucks in late nighttime periods) and possible toxic waste disposal and thus likely to be less preferred areas. The proximity to downtown serves as an indication of proximity to urban centers. As expected, increasing network distance to downtown is associated with a reduction in home values (all else being equal). The findings are similar to several studies examining these impacts (J. Kim et al., 2024; Li et al., 2015). As explored in previous studies, we also find that school quality (in terms of school grade) has a significant impact on the home values. The properties associated with lowest school grade (fair) are likely to have a lower home value compared to properties associated with other school grades (good and very good) (see Wen et al., 2017 for similar results).

4.3.4 Transportation infrastructure characteristics

In our analysis, increasing distance to highways is positively correlated with home values. The results highlight that the negative consequences of proximity to highways such as traffic, air and noise pollution affects pricing. Given the low dense growth in Florida, the finding is not surprising. Given the availability of space in Florida, single-family developments can be built away from highways. On the other hand, proximity to SunRail stations offer a contrasting result i.e. proximity to SunRail stations contributes to higher home value. The presence of SunRail stations contributes to various amenities that can enhance the quality of residents in its proximity. The sidewalk density variable offers a positive correlation with single-family unit

price. The variable might be serving as a surrogate for access to walking (and biking) infrastructure in the neighborhood. Increased access to walking can contribute to neighborhood desirability and thus contribute to home value appreciation.

4.3.5 Temporal characteristics

The time elapsed variable (since 2013) has a positive coefficient. It confirms that the home value has an increasing trend over time – as has been observed across the entire country.

4.3.6 Threshold specific effects

In the model formulation, threshold specific parameters can be modified to contribute to improved data fit. In our analysis, we find 3 adjustments for our thresholds. The estimates are presented in second-row panel of **Table 4**. The threshold parameters estimated are analogous to fixed effects for the different alternatives and do not have any substantive interpretation.

4.3.7 Variance components

The results for the variance component are presented in the third-row panel of **Table 4**. The variance component is parameterized in our model to accommodate for potential heteroscedasticity. The constant parameter represents the overall variance across the observations. The other variance components indicate that the home value variance is found to be higher for parcels with larger parcel area, higher job density and increased land use mix while price variance is lower for higher population block groups. Incorporating for heteroscedasticity in the model offers additional flexibility in representing the true impact of model parameters (reduces bias) and offers enhanced precision.

4.3.8 Spatial autoregressive parameter

As expected, the spatial autoregressive parameter is statistically significant and highly influential. The statistically significant parameter confirms that home values are *highly* influenced by the home values of the spatially closer parcels within the five miles range, along with parcels within the six-year period. To represent the home value correlation in space and time, we present a variation of the spatial parameter with distance and time in Figure 3. From the figure, we can see the impact of the parameter to be substantially higher for neighboring parcels in both space and time dimensions, with dissipation in impact with increasing spatial or temporal distance. Further, the reduction in parameter over time is slower relative to changes in distance. Beyond 1.5 miles, the spatial parameter reduces by nearly 4 orders of magnitude (0.6 to 0.15).

5 MODEL VALIDATION

A validation exercise is conducted to understand how the model performs on data not used in the model estimation process. The dataset prepared for analysis was separated into (a) an estimation sample and (b) a validation sample. In validation exercise, we present two comparisons. First, we compare the spatial lag GGOP model to the simple GGOP model. Second, we compare the spatial lag GGOP model to the linear regression model benchmark model.

5.1 Spatial lag GGOP Vs. simple GGOP

We conduct the validation analysis among spatial lag GGOP and simple GGOP models with 25 random samples of 3000 records each, drawn from the previously separated validation sample.

Using the 25 samples, the simple GGOP and spatial lag GGOP model prediction are obtained at the individual record resolution. The log-composite likelihood and modified BIC values are calculated for each of these samples using the prediction values. The differences in model fit (computed as log-composite likelihood, and modified BIC) would inform us if the model developed is overfitted to the data or not. The LL and modified BIC values for simple and spatial lag GGOP models across the different samples are presented in Figure 4 (a and b). From the figure, we can see that the spatial models consistently outperform the aspatial models for each of the 25 samples. The results highlight the superior performance of the spatial model on data not considered within the estimation sample. Overall, the results indicate that incorporating spatial effects contribute considerably to an improved model.

5.2 Spatial lag GGOP Vs. Linear regression benchmark

Spatial lag GGOP and the linear regression models are estimated using different estimation approaches. Hence, when two models estimated using different estimation approaches are compared, it is necessary to use measures that are not reliant on the estimation approach. Thus, for comparing the proposed spatial lag GGOP model with the linear regression model⁵, we employ several measures including (a) adjusted R^2 , (b) root mean squared error (RMSE), (c) mean absolute error (MAE) and (d) error percentage distribution. The mathematical formulae for computing the metrics are provided below.

The formulations used for the evaluation metrics are defined below:

⁵The estimation results of the linear regression model are presented in the APPENDIX D (**Table 7**).

$$Adjusted R^2 = 1 - \frac{\sum_{m=1}^M (p_m - o_m)^2 (1 - n)}{\sum_{m=1}^M (\bar{o} - o_m)^2 (n - l - 1)} \quad (15)$$

Where, m = observation no. ($m = 1, 2, 3, \dots M$)

p_m = predicted home value for observation m for the model under consideration

o_m = observed home value for observation m

$\bar{o}$ = average of observed home value for observation $m = 1, 2, 3, \dots M$

n = number of observations

l = number of parameters used

$$MAE = \frac{\sum_{m=1}^M |p_m - o_m|}{n} \quad (16)$$

$$RMSE = \sqrt{\frac{\sum_{m=1}^M (p_m - o_m)^2}{n}} \quad (17)$$

Error Percentage Distribution Indicators:

$$No. of records with error \leq 10\% = \sum_{m=1}^M \left(\frac{|p_m - o_m|}{|o_m|} * 100 \leq 10 \right) \quad (18)$$

$$No. of records with error \leq 20\% = \sum_{m=1}^M \left(\frac{|p_m - o_m|}{|o_m|} * 100 \leq 20 \right)$$

$$No. of records with error \leq 30\% = \sum_{m=1}^M \left(\frac{|p_m - o_m|}{|p_m|} * 100 \leq 30 \right)$$

The metrics are computed using five samples for linear regression and spatial lag GGOP models. The results from the comparison are summarized in **Table 5**.

The results from **Table 5** show that spatial lag GGOP model outperforms linear regression in MAE and error percentage distribution metrics. The linear regression model outperforms the spatial lag GGOP model in adjusted R^2 and RMSE values. It might appear from these results that the performance is similar across the two models. However, when we look at the error percentage distribution measures, we can see that linear regression model offers less than or equal to 30% error in only 64.5% of the records while spatial lag GGOP improves the percentage to 80.7%. It is possible that there are outliers in spatial lag GGOP worsening the RMSE but the error percentage distribution clearly shows useful predictive power for the GGOP model. These results highlight how spatial lag GGOP can serve as a complementary tool for home value estimation.

6 ELASTICITY ANALYSIS AND APPLICATION

In linear regression models, the coefficient value directly represents the magnitude of the variable impact on home value. In the GGOP model, the coefficients do not directly represent the variable impact. The coefficient represents the impact on the underlying propensity that impacts the alternative probability. Thus, it is common to estimate elasticity effects to identify the impact of independent variables. In our analysis, we conduct an elasticity analysis by observing the percentage change in the home values for pre-defined changes to independent variables. For continuous independent variables, we consider a 10% change in the independent variable value and examine the average change in predicted value across all records. For ordinal variables, such

as the number of baths and years since 2013, the change is observed for the values by increasing the value by one. For categorical variables such as school grade, the variable values are reversed (such as 0 to 1 and 1 to 0) and the changes across the two groups are appropriately aggregated to arrive at an elasticity measure (see Tirtha et al., 2023 for similar elasticity analysis). The results from the elasticity exercise are presented in Figure 5. The elasticity impacts are easy to interpret. For instance, the elasticity value of the living area variable indicates that a 10% increase in living area will contribute to about 9.5% increase in home price value. All other variables can be interpreted in the same way.

From Figure 5, we can see that, increase in parcel area, living area, number of baths, median income, number of jobs, and years passed since 2013 increase the single-family home values. In contrast, home value is negatively affected where there are increases in population, distance to downtown, school grade quality being fair (compared to good and very good) and proximity to SunRail stations.

The elasticity values provide clear estimates for agencies on how different variables impact home values. For example, if the median income in the census block group is increased by 10%, the single-family home value increases by 2.57%. As newer developments are considered across different census block groups, these values provide estimates of potential new home prices, and the revenue generated in turn through property taxes.

The results also offer a mechanism to compute the impact of school quality on home prices. Given the impact of school quality on home prices, county school boards might consider proactively investing in schools to improve their quality as a means to improve property tax revenues in the region. It is also possible to consider the impact of several other variables such as job availability, proximity to downtown and public transit accessibility around the area for such

analysis. The model also can help with further development of similar home value model in disaster affected areas and unaffected areas to execute comparative home value analysis. In this scenario, case and control areas creation for affected and unaffected areas can be used for the analysis (Eluru et al., 2020).

To further illustrate the applicability of the proposed model system, we employ the model to predict home values for three block groups in the study region and compare them to the observed values (see Figure 6). Figure 6 (a), (c) and (e) illustrate the observed home values while Figure 6 (b), (d) and (f) illustrate the predicted home value in the selected block groups in year 2022. The results highlight how the proposed model predictions reflect the home price value trends well. Further, the results also show the clear spatial relationship in home price value across block groups. Various agencies can employ our model to generate home values across their jurisdictions.

7 CONCLUSION

Home value represents the largest component of wealth for households in the US. With increasing occurrence of natural disasters, an accurate estimate of home values is a useful metric for different agencies. The current study develops a new methodology for home price estimation employing a non-linear price model while providing price value predictions in continuous form. The proposed model is developed for data drawn from 2013 to 2022 on single-family homes in the Greater Orlando region. A wide range of potential influencing variables such as structural characteristics, demographic characteristics, land use and neighborhood characteristics, transportation infrastructure characteristics and temporal factors are considered in the analysis. The proposed grouped generalized ordered probit model and its spatial variants (spatial lag and

error model) are estimated and compared using log-composite likelihood and Bayesian Information Criterion. The spatial models are estimated using different spatio-temporal weight matrices. The model analysis reveals that the spatial lag GGOP model performs better than the simple GGOP and the spatial error GGOP model. The validation analysis also reinforces the superior performance of the spatial (lag) model relative to its aspatial counterparts. The elasticity analysis identifies that characteristics such as parcel area, living area, number of baths, population, median income, number of jobs, distance to downtown, school quality, proximity to SunRail stations and parcel year of analysis have significant influence on the single-family home values.

The findings from our analysis offer useful inputs for policy makers. The model developed can be employed to estimate property damage costs at the individual home resolution and/or community resolution contributing to advanced tools for disaster analysis. Further, the developed price values can be integrated with tools for proactive decision making. As many Florida counties rely on property taxes for revenue, natural disasters can significantly reduce property values and impact county finances. The pricing model developed can be incorporated within a benefit cost analysis tool that will allow counties to identify potential solutions to reduce flooding and hurricane wind impacts in a community. The analysis can allow counties to select treatments that offer savings to life and property. The model findings may also help in understanding how different transportation infrastructure investments can affect housing affordability in the region.

The current study is focused on a single region – Greater Orlando. The region offers significant variability in population, income, housing structural characteristics and several other variables. Thus, while the model is not representative of every home market in the country, the

findings should be applicable to a large share of home markets with a mix of urban, sub-urban and rural developments in the southern part of the country. Thus, the model developed would serve as a starting point for different urbanized regions. It would be beneficial to supplement the modeling efforts in our analysis using data from other urbanized regions. Given the effort essential for data compilation at the parcel resolution for different urbanized regions, this would be an ideal future research exercise for consideration. Considering data from multiple urbanized regions will also allow the consideration of model transferability and development of models' representative of broader geographic regions. The analysis can also benefit from the consideration of sale records, to develop similar models for comparison. Finally, it would also be useful to extend the price value analysis for other residential properties including apartments, condominiums and multi-family homes.

DECLARATION STATEMENT

The authors confirm that the authors have no financial or non-financial competing interests.

The authors confirm contribution to the paper as follows: study conception and design: Naveen Eluru, Sudipta Dey Tirtha, and Susmita Sarker; data collection: Sudipta Dey Tirtha, Susmita Sarker; analysis and interpretation of results: Susmita Sarker, Naveen Eluru, and Sudipta Dey Tirtha; draft manuscript preparation: Susmita Sarker and Naveen Eluru. All authors reviewed the results and approved the final version of the manuscript.

REFERENCES

- Allen, M. T., Austin, G. W., & Swaleheen, M. (2015). Measuring Highway Impacts on House Prices Using Spatial Regression. *Journal of Sustainable Real Estate*, 7(1), 83–98. <https://doi.org/10.1080/10835547.2015.12091876>
- Basu, S., & Thibodeau, T. G. (1998). Analysis of Spatial Autocorrelation in House Prices. *Journal of Real Estate Finance and Economics*, 17(1), 61–85. <https://doi.org/https://doi.org/10.1023/A:1007703229507>
- Bhat, C. R., Sener, I. N., & Eluru, N. (2010). A flexible spatially dependent discrete choice model: Formulation and application to teenagers' weekday recreational activity participation. *Transportation Research Part B: Methodological*, 44(8–9), 903–921. <https://doi.org/10.1016/j.trb.2010.02.003>
- Boyle, A., Barrilleaux, C., & Scheller, D. (2014). Does Walkability Influence Housing Prices? *Social Science Quarterly*, 95(3), 852–867. <https://doi.org/10.1111/ssqu.12065>
- Boyle, M., & Kiel, K. (2001). A Survey of House Price Hedonic Studies of the Impact of Environmental Externalities. *Journal of Real Estate Literature*, 9(2), 117–144. <https://doi.org/10.1080/10835547.2001.12090098>
- Caprioli, C., Bottero, M., & De Angelis, E. (2023). Combining an agent-based model, hedonic pricing and multicriteria analysis to model green gentrification dynamics. *Computers, Environment and Urban Systems*, 102, 101955. <https://doi.org/10.1016/j.compenvurbsys.2023.101955>

- Castro, M., Paleti, R., & Bhat, C. R. (2013). A spatial generalized ordered response model to examine highway crash injury severity. *Accident Analysis & Prevention*, 52, 188–203. <https://doi.org/10.1016/j.aap.2012.12.009>
- Chernobai, E., & Ma, Z. (2022). The Effect of Walkability on House Prices. *Journal of Housing Research*, 31(1), 53–73. <https://doi.org/10.1080/10527001.2021.2007583>
- Chernobai, E., Reibel, M., & Carney, M. (2011). Nonlinear Spatial and Temporal Effects of Highway Construction on House Prices. *The Journal of Real Estate Finance and Economics*, 42(3), 348–370. <https://doi.org/10.1007/s11146-009-9208-9>
- Choi, K., Park, H. J., & Uribe, F. A. (2023). The impact of light rail transit station area development on residential property values in Calgary, Canada: Focus on land use diversity and activity opportunities. *Case Studies on Transport Policy*, 12, 100924. <https://doi.org/10.1016/j.cstp.2022.11.008>
- Cohen, J. P., Ioannides, Y. M., & (Wirathip) Thanapisitikul, W. (2016). Spatial effects and house price dynamics in the USA. *Journal of Housing Economics*, 31, 1–13. <https://doi.org/10.1016/j.jhe.2015.10.006>
- Del Negro, M., & Otrok, C. (2007). 99 Luftballons: Monetary policy and the house price boom across U.S. states. *Journal of Monetary Economics*, 54(7), 1962–1985. <https://doi.org/10.1016/j.jmoneco.2006.11.003>
- Eluru, N., Hasan, S., Dey Tirtha, S., Kumar Dey, B., Zhang, J., & Tanmoy Bhowmik, M. (2020). *Evaluating Community Building Effectiveness of Transportation Investments: Using Traditional and Big Data Oriented Analytical Approaches (Phase 4)*.

University of Central Florida. https://people.cecs.ucf.edu/neluru/wp-content/uploads/2022/04/Final_Report_Oct_06.pdf

Feng, X., & Humphreys, B. R. (2012). The impact of professional sports facilities on housing values: Evidence from census block group data. *City, Culture and Society*, 3(3), 189–200. <https://doi.org/10.1016/j.ccs.2012.06.017>

Florida's Affordable Housing Crisis. (n.d.). Retrieved July 20, 2024, from <https://www.floridapolicy.org/posts/floridas-affordable-housing-crisis>

Gabauer, D., Gupta, R., Marfatia, H. A., & Miller, S. M. (2024). Estimating U.S. housing price network connectedness: Evidence from dynamic Elastic Net, Lasso, and ridge vector autoregressive models. *International Review of Economics & Finance*, 89, 349–362. <https://doi.org/10.1016/j.iref.2023.10.013>

Geerts, M., vanden Broucke, S., & De Weerd, J. (2023). A Survey of Methods and Input Data Types for House Price Prediction. *ISPRS International Journal of Geo-Information*, 12(5), 200. <https://doi.org/10.3390/ijgi12050200>

Godambe, V. P. (1960). An Optimum Property of Regular Maximum Likelihood Estimation. *The Annals of Mathematical Statistics*, 31(4), 1208–1211. <https://doi.org/10.1214/aoms/1177705693>

Golub, A., Guhathakurta, S., & Sollapuram, B. (2012). Spatial and Temporal Capitalization Effects of Light Rail in Phoenix. *Journal of Planning Education and Research*, 32(4), 415–429. <https://doi.org/10.1177/0739456X12455523>

- Helbich, M., Brunauer, W., Vaz, E., & Nijkamp, P. (2014). Spatial Heterogeneity in Hedonic House Price Models: The Case of Austria. *Urban Studies*, 51(2), 390–411. <https://doi.org/10.1177/0042098013492234>
- Holly, S., Pesaran, M. H., & Yamagata, T. (2010). A spatio-temporal model of house prices in the USA. *Journal of Econometrics*, 158(1), 160–173. <https://doi.org/10.1016/j.jeconom.2010.03.040>
- Huang, B., Wu, B., & Barry, M. (2010). Geographically and temporally weighted regression for modeling spatio-temporal variation in house prices. *International Journal of Geographical Information Science*, 24(3), 383–401. <https://doi.org/10.1080/13658810802672469>
- Is climate change increasing the risk of disasters?* (n.d.). Retrieved July 20, 2024, from <https://www.worldwildlife.org/stories/is-climate-change-increasing-the-risk-of-disasters#:~:text=Climate%20change%20is%20increasing%20ocean,storm%20surge%2C%20and%20rainfall%20rates.&text=Nature%20can%20help%20reduce%20the%20damage%20caused%20by%20those%20impacts.>
- Kim, J., Ewing, R., & Rigolon, A. (2024). Does green infrastructure affect housing prices via extreme heat and air pollution mitigation? A focus on green and climate gentrification in Los Angeles County, 2000-2021. *Sustainable Cities and Society*, 102, 105225. <https://doi.org/10.1016/j.scs.2024.105225>
- Kim, J., & Long, D. (2024). Working from home, commuting time, and intracity house-price gradients. *Journal of Regional Science*, 64(3), 866–895. <https://doi.org/10.1111/jors.12693>

- Koster, H. R. A., & Rouwendal, J. (2012). THE IMPACT OF MIXED LAND USE ON RESIDENTIAL PROPERTY VALUES*. *Journal of Regional Science*, 52(5), 733–761. <https://doi.org/10.1111/j.1467-9787.2012.00776.x>
- L. Krause, A., & Bitter, C. (2012). Spatial econometrics, land values and sustainability: Trends in real estate valuation research. *Cities*, 29(SUPPL.2), S19–S25. <https://doi.org/10.1016/j.cities.2012.06.006>
- Levkovich, O., Rouwendal, J., & van Marwijk, R. (2016). The effects of highway development on housing prices. *Transportation*, 43(2), 379–405. <https://doi.org/10.1007/s11116-015-9580-7>
- Li, W., Joh, K., Lee, C., Kim, J.-H., Park, H., & Woo, A. (2015). Assessing Benefits of Neighborhood Walkability to Single-Family Property Values: A Spatial Hedonic Study in Austin, Texas. *Journal of Planning Education and Research*, 35(4), 471–488. <https://doi.org/10.1177/0739456X15591055>
- Liu, X. (2013). Spatial and Temporal Dependence in House Price Prediction. *The Journal of Real Estate Finance and Economics*, 47(2), 341–369. <https://doi.org/10.1007/s11146-011-9359-3>
- Luque, J. P., Ikromov, N., & Noseworthy, W. B. (2019). Housing Affordability Crisis: The United States. In *Affordable Housing Development*. Springer, Cham. https://doi.org/https://doi.org/10.1007/978-3-030-04064-2_1
- Palimaru, A. I., McBain, R., McDonald, K., Batra, P., & Hunter, S. B. (2023). The relationship between quality of housing and quality of life: evidence from

- permanent supportive housing. *Housing and Society*, 50(1), 13–34.
<https://doi.org/10.1080/08882746.2021.1928853>
- Parvez, D. A., Tirtha, S. D., Bhowmik, T., & Eluru, N. (2023). Joint Econometric Model Framework for Transportation Network Company Users' Trip Fare and Destination Choice Analysis. *Transportation Research Record: Journal of the Transportation Research Board*, 2677(7), 545–557. <https://doi.org/10.1177/03611981231155157>
- Peng, Q., & Knaap, G. (2023). When and Where Do Home Values Increase in Response to Planned Light Rail Construction? *Journal of Planning Education and Research*, 0739456X2211330. <https://doi.org/10.1177/0739456X221133022>
- Rahman, M., Yasmin, S., & Eluru, N. (2019). Evaluating the impact of a newly added commuter rail system on bus ridership: a grouped ordered logit model approach. *Transportmetrica A: Transport Science*, 15(2), 1081–1101.
<https://doi.org/10.1080/23249935.2018.1564800>
- Rey-Blanco, D., Zofio, J. L., & González-Arias, J. (2024). Improving hedonic housing price models by integrating optimal accessibility indices into regression and random forest analyses. *Expert Systems with Applications*, 235, 121059.
<https://doi.org/10.1016/j.eswa.2023.121059>
- Soltani, A., Pettit, C. J., Heydari, M., & Aghaei, F. (2021). Housing price variations using spatio-temporal data mining techniques. *Journal of Housing and the Built Environment*, 36(3), 1199–1227. <https://doi.org/10.1007/s10901-020-09811-y>

- Streimikiene, D. (2015). Quality of Life and Housing. *International Journal of Information and Education Technology*, 5(2), 140–145. <https://doi.org/10.7763/IJiet.2015.V5.491>
- The Development of the World Health Organization Quality of Life Assessment Instrument (the WHOQOL). (1994). In *Quality of Life Assessment: International Perspectives* (pp. 41–57). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-79123-9_4
- Thompson, J. J., Wilby, R. L., Hillier, J. K., Connell, R., & Saville, G. R. (2023). Climate Gentrification: Valuing Perceived Climate Risks in Property Prices. *Annals of the American Association of Geographers*, 113(5), 1092–1111. <https://doi.org/10.1080/24694452.2022.2156318>
- Tirtha, S. D., Bhowmik, T., & Eluru, N. (2023). Understanding the factors affecting airport level demand (arrivals and departures) using a novel modeling approach. *Journal of Air Transport Management*, 106, 102320. <https://doi.org/10.1016/j.jairtraman.2022.102320>
- Tirtha, S. D., Bhowmik, T., & Eluru, N. (2024). Accommodating spatio-temporal dependency in airline demand modeling. *Journal of Air Transport Management*, 116, 102572. <https://doi.org/10.1016/j.jairtraman.2024.102572>
- Turnbull, G. K., Zahirovic-Herbert, V., & Zheng, M. (2018). Uncertain School Quality and House Prices: Theory and Empirical Evidence. *The Journal of Real Estate Finance and Economics*, 57(2), 167–191. <https://doi.org/10.1007/s11146-017-9611-6>

- Value of assets U.S. households own and the debt they carry* | Pew Research Center. (n.d.). Retrieved July 20, 2024, from <https://www.pewresearch.org/2023/12/04/the-assets-households-own-and-the-debts-they-carry/>
- Wen, H., Xiao, Y., & Zhang, L. (2017). School district, education quality, and housing price: Evidence from a natural experiment in Hangzhou, China. *Cities*, 66, 72–80. <https://doi.org/10.1016/j.cities.2017.03.008>
- Wu, J., Deng, Y., & Liu, H. (2014). House Price Index Construction in the Nascent Housing Market: The Case of China. *The Journal of Real Estate Finance and Economics*, 48(3), 522–545. <https://doi.org/10.1007/s11146-013-9416-1>
- Yang, L., Liang, Y., He, B., Yang, H., & Lin, D. (2023). COVID-19 moderates the association between to-metro and by-metro accessibility and house prices. *Transportation Research Part D: Transport and Environment*, 114, 103571. <https://doi.org/10.1016/j.trd.2022.103571>
- Zhuge, C., Shao, C., Gao, J., Dong, C., & Zhang, H. (2016). Agent-based joint model of residential location choice and real estate price for land use and transport model. *Computers, Environment and Urban Systems*, 57, 93–105. <https://doi.org/10.1016/j.compenvurbsys.2016.02.001>

APPENDIX A NUMBER OF BINS SELECTION

In our analysis, we employed 15 price bins to ensure reasonable distribution of observations across the price bands. To ensure that the results are not unduly influenced by the price bin selection, we evaluated the model development with 12 and 18 price bins. For the three price bin alternatives, we developed the home price model with the same independent variable specification. A cursory examination of the model results highlights that for the three models, the variables offer the same sign and similar magnitude with some differences. To further investigate the differences, we computed the percentage change in the coefficient value for the 12-bin and 18-bin specification relative to the 15-bin specification. The values computed are presented in Figure 7. From the comparison, we can see that the differences across the various specifications are very small highlighting the stability of the model coefficients across the three price bin alternatives. Overall, the analysis clearly supports the consideration of 15-bin alternative as a reasonable approach for our analysis. The reader would note that across the three specifications, the number of threshold parameters can vary. Hence, we only focus on the independent variables' coefficients.

APPENDIX B SAMPLE SIZE ANALYSIS

Prior to settling on the final sample size for our model of 3000 records, we conducted an exhaustive analysis to ensure sample size does not influence the model specification significantly. We developed models for five samples each for sample sizes of 3,000, 6,000, 12,000 and 18,000. It is important to note that across all specifications the model parameters present the same sign with minor differences in the parameter magnitude. We compare the specifications across the different sample sizes with results for 3,000 records as the base

specification. The comparison results defined as percentage difference are presented in Figure 8. From the comparison, we can see that across all the variables the parameters exhibit minor differences. For propensity variables, the average error is quite low with only two parameters crossing 15% error. The comparison clearly illustrates that the model results for 3,000 are quite stable and are representative of the population samples.

APPENDIX C DISTANCE THRESHOLD SELECTION IN THE WEIGHT MATRIX

The log-composite likelihood values for different distances in the weight matrix are presented in **Table 6**.

APPENDIX D LINEAR REGRESSION MODEL RESULTS

The linear regression home value model with an average of five samples estimates is presented in **Table 7**.

Table 1 Variables description and distribution from the sample dataset

Variable	Variable description	Min.	Max.	Mean	Std.Dev.
Independent Variables (Continuous)					
<i>Structural Characteristics</i>					
Parcel area	The parcel area in acres	0.015	11.404	0.318	0.628
Total living area	The living area in the parcel in acres * 10 ³	12.718	231.680	49.467	20.744
<i>Demographic Characteristics</i>					
Population	Population in the block group * 10 ⁻⁴	0.017	3.423	0.513	0.523
Median income	Median income in \$ in the block group * 10 ⁻⁴	1.474	25.103	7.085	3.161
Job density	The number of jobs per acre in a county	0.124	2.228	1.496	0.632
<i>Land use and neighborhood characteristics</i>					
Vacant area	Proportion of vacant area in the block group * 10	0.000	798.300	142.150	115.470
Commercial area	Proportion of commercial area in the block group * 10	0.000	630.010	59.700	84.920
Single-family residential area	Proportion of single-family residential area in the block group * 10	9.640	1000.000	410.250	223.240
Multi-family residential area	Proportion of multi-family residential area in the block group * 10	0.000	777.320	26.190	50.600
Other residential areas	Proportion of other residential areas in the block group * 10	0.000	872.040	109.470	123.320
Industrial area	Proportion of industrial area in the block group * 10	0.000	520.530	20.230	55.230
Recreational area	Proportion of recreational area in the block group * 10	0.000	716.430	18.650	68.630
Land use mix	Land use mix = $\left[\frac{-\sum_k (P_k * (\ln P_k))}{\ln N} \right]$ Where, k = land use category P = land use proportion for a specific category N = the total number of land use categories	0.000	0.851	0.541	0.152
Distance to downtown	Network distance to the nearest downtown in miles	0.331	24.689	9.804	3.928
<i>Transportation infrastructure characteristics</i>					
Distance to highways	Distance to the nearest highway in miles	0.014	5.591	0.886	0.790
Distance to rail stations	Distance to the nearest SunRail station in miles	0.118	17.155	5.876	3.674
Sidewalk density	Sidewalk length in feet per acre area of each block group * 10 ⁻¹	0.000	11.296	1.083	1.220
Independent Variable (Ordinal)					
<i>Structural Characteristics</i>					
Number of baths	The number of baths * 10	0.000	150.000	23.720	8.300
Temporal Variable					
Years since 2013	The number of years passed since 2013	0.000	90.000	47.010	28.880

Variable	Variable description	Min.	Max.	Mean	Std.Dev.
* 10					
Independent Variables (Categorical)				Freq.	Percent
<i>Neighborhood Characteristics</i>					
School grade					
Very good grade	The parcel is in the school zone of a very good grade school			785	26.167
Good grade	The parcel is in the school zone of a good grade school			811	27.033
Fair grade	The parcel is in the school zone of a fair to poor grade school			1404	46.800

Table 2 Log-composite likelihood for different weight matrices

Weight matrix formulations	Log composite likelihood (Spatial lag model)
$1/\sqrt{\text{distance}} * (\text{year difference} + 1)$	-28675590
$1/(\sqrt{\text{distance}} * (\text{year difference} + 1))$	-28705260
$1/(\text{distance} * \sqrt{(\text{year difference} + 1)})$	-28460400
$1/(\text{distance} * (\text{year difference} + 1))$	-28514040
$1/(\text{distance}^2 * (\text{year difference} + 1))$	<u>-28446469</u>
$1/(\text{distance}^3 * (\text{year difference} + 1))$	-28473390
$1/(\text{distance} * \ln((\text{year difference} + 1)))$	-28468770

Table 3 Measures of fit in terms of log-composite likelihood and modified BIC

Model	Number of parameters	Log composite likelihood	Modified BIC
Simple GGOP model	21	-28761945.0	9758.64
Spatial lag model	22	<u>-28446469.0</u>	<u>9661.45</u>
Spatial error model	22	-28761922.0	9766.64

Table 4 Spatial lag model estimation results

Variables	Estimates	t statistics
Propensity Components		
Constant	-10.7164	-30.58
<i>Structural characteristics</i>		
Parcel area in acres	0.9606	2.71
Total living area in acres * 10 ³	0.3225	47.50
Number of baths * 10	0.1420	10.02
<i>Demographic characteristics</i>		
Population in the block group * 10 ⁻⁴	-0.4303	-6.41
Median income in \$ in the block group * 10 ⁻⁴	0.6274	26.28
Job density (number of jobs per acre in a county)	1.6807	23.76
<i>Land use and neighborhood characteristics</i>		
Land use proportions (Base: other land use proportions)		
Industrial areas in the block group * 10	-0.7402	-11.09
Network distance to the nearest downtown in miles	-0.0939	-3.16
School grade for individual parcels (Base: Other grades)		
Fair grade * 10	-0.1888	-18.23
<i>Transportation infrastructure characteristics</i>		
Distance to the nearest highway in miles	0.1753	2.30
Distance to the nearest SunRail station in miles	-0.1802	-7.33
Sidewalk length in feet per acres block group * 10 ⁻¹	0.1725	37.36
<i>Temporal characteristics</i>		
Years since 2013 * 10	0.1009	37.36
<i>Threshold Specific Effects</i>		
Threshold 13 specific parameter	2.2138	12.66
Threshold 14 specific parameter	5.1122	20.13
Threshold 15 specific parameter	7.9789	21.90
<i>Variance components</i>		
Constant	1.0693	18.14
Parcel area in acres	0.3685	3.48
Population in the block group * 10 ⁻⁴	-0.3139	-9.65
Job density (number of jobs per acre in a county)	0.0860	3.71
Land use mix	0.2497	3.05
Spatial autoregressive parameter		
δ	0.1868	10.63

Table 5 Comparison of evaluation metrics among linear regression and spatial lag GGOP model

Metrics	Linear regression	Spatial lag GGOP
Adjusted R ²	0.680	0.634
MAE	5.916	4.684
RMSE	11.441	12.240
No. of records with error $\leq 10\%$	25.133	36.213
No. of records with error $\leq 20\%$	47.613	63.553
No. of records with error $\leq 30\%$	64.480	80.693

Table 6 Log-composite likelihood for different distance threshold values

Distance threshold (in miles)	Log composite likelihood (Spatial lag model)
4.0	-28446471
4.5	-28446470
5.0	<u>-28446469</u>
5.5	-28446469
6.0	-28446469
7.0	-28446469

Table 7 Linear regression model estimates

Variables	Estimates	t statistics
Propensity Components		
Constant	-22.3667	-20.92
Structural characteristics		
Parcel area in acres	1.2321	3.93
Total living area in acres * 10^3	0.5720	38.21
Number of baths * 10	0.2521	6.95
Demographic characteristics		
Population in the block group * 10^{-4}	-2.1198	-4.96
Median income in \$ in the block group * 10^{-4}	1.0478	12.40
Job density (number of jobs per acre in a county)	3.1285	8.99
Land use and neighborhood characteristics		
Land use proportions (Base: other land use proportions)		
Industrial areas in the block group * 10	-0.8958	-2.19
Network distance to the nearest downtown in miles	-0.3267	-3.83
School grade for individual parcels (Base: Other grades)		
Fair grade * 10	-0.1843	-3.83
Transportation infrastructure characteristics		
Distance to the nearest highway in miles	1.1476	3.87
Distance to the nearest SunRail station in miles	-0.3732	-4.04
Sidewalk length in feet per acres block group * 10^{-1}	0.5430	2.76
Temporal characteristics		
Years since 2013 * 10	0.0991	11.70
Variance components		
Constant	11.2422	77.46

Figure 1 Dependent variable distribution

Figure 2 Wald t-stat distribution for the parameters used in the simple GGOP model

Figure 3 Spatial autoregressive parameter variations over the selected space and time constraints

Figure 4 Validation analysis in the simple GGOP and spatial lag models:

- (a) Log-composite likelihood of the validation samples and**
- (b) Modified BIC values of the validation samples**

Figure 5 Elasticity effects of the exogenous variables on home values

Figure 6 Observed and predicted home value distribution in three randomly selected census block groups for year 2022:

- (a) Observed home value in census block group 1,**
- (b) Predicted home value in census block group 1,**
- (c) Observed home value in census block group 2,**
- (d) Predicted home value in census block group 2,**
- (e) Observed home value in census block group 3 and**
- (f) Predicted home value in census block group 3**

Figure 7 Percentage change in GGOP model estimates with change in number of bins

Figure 8 Percentage change in GGOP model estimates with change in sample size