Proof-of-Concept Land Value Modeling in Baltimore, Maryland

Chester Harvey, Ph.D., NCSG

Eli Knaap, Ph.D., COGS

Nicholas Finio, Ph.D., NCSG

Casey Dawkins, Ph.D., NCSG

Technical Report
prepared for the
Robert Schalkenbach Foundation

March 2026 Revision

National Center for Smart Growth (NCSG)
University of Maryland, College Park

Center for Open Geographical Science (COGS)
San Diego State University

Introduction

This report summarizes our efforts to develop a regression-based hedonic price model to estimate land values in Baltimore City, Maryland. Our team developed a series of regression-based statistical models to estimate land values as a function of property-level variables. Below, we summarize theoretical and methodological background on hedonic price measurement of vacant land from research literature. Next, we document our modeling approach, including data we collected to train the model, the form we used to specify it, and our approach to estimation, validation, and prediction. We then present and interpret our final model specification. Finally, we offer suggestions for further research, suggestions for applicability in other geographies, and ideas for methodological improvement.

This revised version of our original 2024 report reflects additional work to address an issue with prediction of negative land values when decomposed from total property value, inconsistent with the logical expectation that the vast majority of parcels have positive value. We address this with updated model specifications that prioritize logical decomposed predictions with limited reductions in overall model fit. This revision demonstrates how simpler models may more effectively capture how key components of property value are represented in real-world assessment—both land and improvements contribute positively to overall value. It also reinforces how complex models cannot effectively substitute for high-quality data.

Theoretical & Methodological Background

If vacant land sales were randomly distributed across space, we could predict the value of vacant land in any given location using the predicted values from a traditional hedonic regression, where the dependent variable is the transaction price of vacant land, and the independent variables are the observable characteristics of vacant land at different locations within the city. Unfortunately, there are at least two distinct problems with this approach:

1. In many cities, there may be few or no sales of vacant land. In the extreme case of zero vacant land sales, researchers have typically relied on one of two approaches to infer the value of land from the value of improved properties (Bourassa and Hoesli, 2022):

    a. Using the “residual” or “land leverage” approach, researchers have deflated the observed value of improved properties by an estimate of the cost of constructing the structure on improved land. The residual (total property value minus construction costs) yields the value of land, and land leverage is equal to the ratio of vacant land value to overall property value. Papers relying on this method include Davis and Palumbo (2008), Davis et al. (2017), Davis et al. (2021), and Kuminoff and Pope (2011). An obvious limitation of this method is that it requires precise estimates of construction costs, which may not be available or may require unrealistic assumptions. However, in contrast to most states, Maryland is unique in requiring tax assessors to separately assess the value of land and structures, so a full account of estimated land and structure costs is available in the state tax assessor’s database.

    b. Using matching methods, researchers have compared the sale price for a vacant land parcel with the subsequent inflation-adjusted sale price of that same parcel after it has been improved. Examples of this approach include Thorsnes (1997) and Ahlfeldt and McMillen (2018). A limitation of this method is that it requires a long time series of repeat sales transactions data, which may not be readily available. Also, for reasons described at the end of this review, the price of a vacant land parcel evolves over time in relation to its option value and the presence of development on adjacent parcels. Simply subtracting the value of vacant land at some prior time from its later improved value would fail to capture these effects.

Bourassa and Hoesli (2022) compare these two approaches with the hedonic approach (described below) and conclude that the matching method performs best and is the most practical among the alternatives.

2. Even in cities with a large sample of vacant land sales, these transactions are likely not randomly distributed but are clustered in space. In a small but growing city, for example, most vacant land sales will be at or near the urban fringe. If transactions are geographically clustered, the observed and unobserved characteristics of vacant land sales may differ from the observed and unobserved characteristics of land that is not sold. This induces a form of sample-selection bias (Albouy, Ehrlich, and Shin, 2018).

Given these problems, researchers have relied on two alternatives:

1. Rely on transactions data for improved sites to estimate hedonic models that decompose total improved property value into observable structure characteristics and observable land characteristics. Wentland et al. (2020) rely on a variation of this approach to estimate land values for locations across the US. A problem with this approach is that it may fail to capture the combined influence of land and improvement characteristics on improved land values. As a result, the estimated coefficients in a hedonic model may differ between improved and vacant land models (Bourassa and Hoesli, 2022).

   Land characteristics in such models typically include:

   - Lot size
   - Zoned land use
   - Distance to CBD
   - Distance to public transit stop
   - Whether in a flood-prone zone
   - Presence or absence of property improvements such as paved street access, grading, platting and engineering, etc.
   - Whether designated as a brownfield
   - Features of adjacent properties (whether developed, whether blighted, etc.)
   - Features of the transaction (eminent domain, foreclosure, etc.)
   - Neighborhood quality (school quality, neighborhood crime rates, etc.)

   Structure characteristics in such models typically include:

   - Structure square footage
   - Building age
   - Structure type
   - Building construction material
   - # of bedrooms and bathrooms, in models estimated for residential land sales only

2. Estimate a hedonic regression using a sample of teardowns (properties slated for demolition), based on the rationale that teardowns occur when a property's value has depreciated to the point where its sales value equals the value of vacant land, assuming that demolition costs are trivial (Rosenthal and Helsley, 1994). If demolition costs are nontrivial, estimates must be deflated by the cost of demolition. Examples of this approach include Gedal and Ellen (2018), Dye and McMillen (2007), and Bourassa and Hoesli (2022). There are two problems with this approach. First, the sample size of teardowns may be small. Second, the assumption of zero demolition costs is likely unrealistic, and accurate estimates of demolition costs may be difficult to obtain.

Another issue affecting all methods described above is that the value of vacant land at any given point in time is determined by its current use value and the option value of that land in alternative future uses. The time path of a parcel's option values implies that development timing and land prices are jointly determined. Cunningham (2006) develops an approach to jointly estimate vacant land values and the timing of development, and Rosenthal and Helsley (1994) and Dye and McMillen (2007) estimate vacant land values conditional on the probability of a teardown.

A related issue is that the value of vacant urban land is endogenously determined by the timing and intensity of development on adjacent parcels. For example, assume that an edge city emerges at a given exurban location in time t. At any time prior to t, the value of vacant land in that location would be primarily determined by the productivity of land for agricultural uses. However, after time t, the parcel's use would be determined by its proximity to other development within the edge city, even if the parcel's characteristics have remained unchanged. Estimation of such a model would seem to require estimating the value of vacant land conditional on an adjacent property being previously developed, perhaps using a modified Heckman-style two stage model such as the ones employed by Rosenthal and Helsley (1994) and Dye and McMillen (2007), but to our knowledge, no studies have relied on this approach.

Summary of Approach

Data Collection

We compiled property-level data from several public sources to build a dataset of property sales within Baltimore City between 2013 and 2022 and a parallel dataset of all properties within the city in 2022. Both datasets included the same variables except that the former included sale prices and years. This sales dataset was used to estimate and validate our regression model. The latter dataset was fed into the estimated model to predict a land value surface for 2022.

Both datasets were based on property records from MDProperty View, a database of all properties throughout Maryland periodically updated and made publicly available by the Maryland Department of Planning (MDP). MDProperty View includes numerous property-level variables describing location, structure characteristics, condition, ownership, and sale information. NCSG has worked with MDP to archive annual versions of MDProperty View. We used this archive to develop our own database that combined sale records with the most contemporaneous available property records to form annual inventories of sales linked to annual snapshots of property characteristics. Annual inventories of sales between 2013 and 2022 (N=97,414), and properties in 2022 (N=134,996), formed the basis of our modeling dataset for this project.

We augmented sale and property records from MDProperty View with contextual data on zoning, schools, crime, demographics, floodplains, tree canopy, and access to jobs, and proximity to transportation infrastructure, parks, the central business district, and the City's waterfront. As part of the spatially-lagged model specification, property-level variables were additionally averaged within quarter-mile buffers around each property to reflect neighborhood characteristics. When possible, data used to augment MdProperty View were gathered from sources that were consistent for Baltimore City and neighboring counties—Baltimore County and Anne Arundel County—to avoid edge biases in neighborhood summaries at the city boundary. Data available at the statewide level were preferred in order to enable future expansion of modeling across Maryland. Collected variables and their sources are summarized in Table 1.

Despite the extensive scope of data available from MDProperty View, many of its fields contained null values for a large number of records, limiting sample sizes used for model estimation. We assumed that null values were randomly distributed and estimated models using all complete records for a given model specification; our modeling software automatically excluded records containing one or more null values for required variables. For this reason, the sample size used for estimation changed as we iterated model specifications, with models that relied on more inputs tending to be based on moderately smaller samples.

Table 1. Summary of data sources and variables assembled for modeling. Dates indicate data vintage, not publication date.

Data Source	**Variables**
MdProperty View Maryland Department of Planning 2013-2022	● Sale price ● Sale date ● Land area ● Development/vacancy status ● Land use ● Structure type/style ● Structure condition ● Structure age ● Structure square footage ● Structure stories ● Units in property ● Properties on parcel
Maryland Generalized Zoning Maryland Department of Planning 2021 Not publicly released. Obtained directly from MDP.	● High, medium, low, very low density residential ● Mixed use ● Commercial ● Industrial
Floodplain FEMA 2018	● 100-year floodplain

Data Source	Variables
Tree Canopy Cary Institute of Ecosystem Studies 2015	● Proportion of area within ¼ mile of property covered by tree canopy
Smart Location Database EPA 2021	● Residential, employment, activity density ● Land use entropy (diversity) ● Job access by auto and transit ● Walkability index
Longitudinal Employer Household Dynamics (LEHD) US Census 2022	● Total job accessibility
EJScreen EPA 2024	● Block group-level health outcome risk ● Environmental exposure risk
Centrality to CBD N Calvert St @ E Fayette St	● Distance to CBD
Transit Access Light Rail, Metro SubwayLink, and MARC 2017	● Distance to closest fixed guideway transit station
Park Access Open Baltimore 2024	● Distance to closest park
Waterfront Open Baltimore 2023	● Distance to harbor
School Performance Stanford Education Data Archive 2016	● Grade-cohort-standardized achievement score
Crime Open Baltimore 2022	● Arrests within ¼ mile of property
American Community Survey US Census 2021	● Block group-level age, income, race/ethnicity, education, population density, household density, owner/renter status

Modeling Methods

The hedonic approach was conceived as a structural model, designed to uncover the determinants of housing demand based on principles of economic theory (Koopmans, 1949; Nevo & Whinston, 2010). Rosen (1974) postulated that under equilibrium conditions, the bundle of housing goods can be decomposed into its constituent parts. The structural components rely on two critical assumptions: 1. that the housing market is in short-run equilibrium (i.e. a Tieboutian process is at work) and (relatedly) 2. that each consumer individually maximizes their utility function. Under those two assumptions, the market clears and we can observe revealed preferences and willingness to pay for different housing attributes. In practice, many scholars want to estimate hedonic models to understand the effect of some policy change on home prices, but in these cases identification issues are rife, especially if home prices can affect one another.

Spatial data can help solve some inherent difficulties in identifying the hedonic model. First, as Won Kim et al. (2003) describe, the traditional hedonic model is biased in the presence of spatial processes such as land use regulations or environmental quality. "The traditional hedonic property value model does not capture these induced [spatial] effects. It is customary in traditional models to include exogenous variables in the neighborhood characteristics category that try to explain why some neighborhoods tend to have higher or lower housing prices than other neighborhoods. Such a specification cannot capture spatial price effects generated by a change in a neighborhood's housing characteristics (such as a shift in environmental quality). Therefore, a traditional hedonic property value model may lead to a biased or at least imprecise estimate of the benefits of a housing characteristic change if these induced effects are present" (Won Kim et al., 2003).

Furthermore, the housing market is a clear example where the concept of spatial autoregressive processes are both conceptually intuitive and substantively interesting. Houses are sold at different times and their value is often unclear until the point of sale. As such, each sale in the market sends a signal about the value of homes in that neighborhood; there is a clear process of spatial spillover, as other nearby homes are revalued according to the nearby sale.

In this specification of housing price function, the value of a house at any location is dependent on its counterparts at nearby locations in addition to its structural and neighborhood attributes. A spatial weights matrix W determines the hypothesized spatial dependence among residential structures, and is specified *a priori*. The coefficient p measures the absolute price impact of nearby houses on the price of a particular house. As argued in Can (1992), this conceptualization corresponds more closely with the actual workings of the real estate institution in urban housing markets. "A realtor will appraise a house given the price history of houses in the immediate vicinity in addition to other substantive characteristics. At the same time, home owners will initiate or forego certain improvements based on the anticipated return on their investment considering housing prices in the immediate area" (Can, 1992).

Spatial econometric methods have been developed explicitly for use in contexts such as hedonic price modeling because they provide for efficient and unbiased estimation of coefficients when property sales are spatially interactive (Anselin & Lozano-Gracia, 2008; Can, 1992; Comber & Arribas-Bel, 2017; Diao, 2015; Diao et al., 2017; Dubé et al., 2014; Dubin, 1992; Kim et al., 2020; Steimetz, 2010; Won Kim et al., 2003). For example spatial econometric models provide an avenue for studying the relationship between housing characteristics and sales prices, even when nearby home sales have an endogenous influence on prices. These same properties provide an ideal opportunity for understanding how to value land and its features in a hedonic modeling framework when land and improvement characteristics would otherwise be difficult to separate.

Spatial econometrics offer important solutions to some identification problems, but also occupy a nebulous space in the econometric modeling world. Technically, spatial econometrics models are structural because they require a priori specification of a spatial weights matrix that defines the feedback structure among observations, and they should be guided by a strong theoretical basis for which the data generating process is at play (Gibbons & Overman, 2012; McMillen, 2012). Spatial models still require the parameterization of a spatial weights matrix (W) (Corrado & Fingleton, 2012), and it is often hard to specify with certainty whether residual autocorrelation is caused by unobserved variables or processes of spatial spillover. The spillover term can never be known in practice, and differentiating spillover from spatially autocorrelated errors is exceedingly difficult in practice, so the inclusion of a global spillover parameter needs to be governed by theory (LeSage, 2014b; LeSage & Pace, 2014).

Model Specification & Estimation

We began by fitting an ordinary least squares (OLS) regression model using the variables described above and the form

$$y=\alpha+\beta X+\epsilon$$

where coefficients in βX were each attributed to either improvement or land characteristics of each property.

The model specification was developed through an iterative approach in which all available variables were initially entered and were removed if they had consistently non-significant effects across all variables of a similar type. Several non-significant dummy variables were maintained in order to provide statistical controls for categorical references.

When refining the specification, we aimed to maximize model fit while also ensuring (a) logical interpretation of coefficients (e.g., were their signs and magnitudes appropriate given theoretical understandings of relationships between predictors and property values), (b) stability of coefficients (e.g., did adding a predictor dramatically change coefficients for other predictors, an indicator or colinearity or other confounding relationships), and (c) logical decomposition of land and improvement values. We did not include a constant term in the model because it was unclear whether this term would be attributed to land or improvement value.

Notably, we found that including predictors compiled within Census geographies (e.g., block groups), such as ACS demographics and EJScreen data, unrealistically impacted land valuations for many properties. This, in turn, caused the model to unrealistically adjust improvement values to compensate. We believe this was due to, in essence, to the modifiable areal unit problem: the potential to misrepresent conditions at a specific location by generalizing over an arbitrary areal unit. In our Baltimore case, block groups were likely insufficient for representing characteristics of neighborhood context that varied substantially and meaningful within them. To more realistically predict decomposed land and improvement values, we excluded demographic and environmental predictors compiled at the block group level, which did not substantially increase overall model fit and dramatically impacted balances between land and improvement values. Elementary school scores exhibited this issue to a lesser extent, potentially because the spatial units in which they were tabulated, school attendance boundaries, were less arbitrary, impacting all the properties within them in similar ways. Nonetheless, they unrealistically inflated land values, so they were excluded from the final model. Notably, block group-level job accessibility metrics from the EPA Smart Location Database (SLD) did not exhibit this issue, potentially due to lower variability in these metrics between adjacent units.

We initially modeled all property types but found that improvement-related variables available in MdProperty View were insufficient for representing the variability and complexity of non-residential and apartment-style residential properties. They also poorly predicted sale prices of high-value properties, likely also because high-value improvements were poorly represented. This resulted in artificially low estimates of improvement value and, therefore, high estimates of land value. We determined that the most consistent and reliable approach for estimating land value was to model only non-apartment residential properties with sale prices less than $800,000, which constituted approximately 72% of 2013–2023 sales and 57% of 2022 properties. Because non-apartment residential properties were located throughout most parts of the city, restricting analysis to these properties still allowed for widespread coverage of land value predictions. It did, however, result in lower densities of observations in heavily commercial and industrial areas, and no predictions in the central business district.

The full list of terms and parameter estimates for the OLS model of sales across all study years is included in Appendix 1. The OLS model had a strong fit with total sales prices given its relatively few exogenous variables. The R^2 value was 0.84 and residuals followed a nicely normal distribution centered on zero. Nullifying improvement and land characteristics and summing the remaining βX yielded predicted land and improvement values, respectively. The distributions of these values, alongside the distribution of predicted total values, are reported in Table 2. The OLS model was estimated with the *statsmodels* Python package.

Table 2. Distributions of Predicted 2013-2023 Land, Improvement, and Total Sales Values from the OLS Model

	Decomposed and Total Predicted Values		
	Land	**Improvement**	**Total**
Mean	$83,380	$90,952	$174,331
STD	$35,609	$90,087	$105,730
Min	$-34,153	$-91,156	$-15,223
25th Percentile	$57,713	$31,591	$97,133
50th Percentile	$79,979	$43,924	$139,911
75th Percentile	$106,861	$125,051	$224,464
Max	$531,092	$944,796	$1,111,554
N	97,414	97,414	97,414
N(<$0)	546	115	9

Spatial Modeling

Many hedonic models use a non-spatial form. Spatial model diagnostics (Table 3), however, indicated that the OLS model was biased or inefficient due to lag effects occurring across space. To address this, we investigated whether a model that accounted for spatial lags could more effectively predict values.

Table 3. Spatial Diagnostics for the OLS Model

	DF	Value	Prob
Regression Diagnostics			
Multicollinearity Condition Number		65.6	
Test On Normality Of Errors			
Jarque-Bera	2	153,318.2	0.000
Diagnostics For Heteroskedasticity			
Random Coefficients			
Breusch-Pagan Test	32	28,594.2	0.000
Koenker-Bassett Test	32	7,300.3	0.000
Diagnostics For Spatial Dependence			
SAR			
Moran's I (Error)		796.0	0.000
Lagrange Multiplier (Lag)	1	39,079.6	0.000
Robust Lm (Lag)	1	5,709.0	0.000
Lagrange Multiplier (Error)	1	608,598.6	0.000
Robust Lm (Error)	1	575,228.0	0.000
Lagrange Multiplier (Sarma)	2	614,307.6	0.000
Spatial Durbin			
Lm Test For Wx	32	14,514.8	0.000
Robust Lm Wx Test	32	564,453.1	0.000
Lagrange Multiplier (Lag)	1	39,079.6	0.000
Robust Lm Lag - Sdm	1	589,018.0	0.000
Joint Test For Sdm	33	603,532.7	0.000

We tested several types of spatially weighted models that account for spatial lags following Knaap's (2025) methodology for spatial econometric modeling. Our preferred model took the form of a spatial Durbin error model (SDEM), which allowed for local spillovers and a global error diffusion process:

$$y=\alpha+\beta X+\theta WX+u,$$

$$u=\lambda Wu+\epsilon.$$

This model included “spatially lagged” regressors (WX), which represented the average values of X variables from nearby properties (where the spatial weights matrix W encoded “nearby” as properties within a quarter mile of each parcel), as well as a spatially-correlated residual term (u). Model diagnostics, including classical and robust forms of the Lagrange Multiplier test, suggested that the SDEM specification was preferable to other well-known spatial models such as the spatial error model (SEM), which includes no exogenous spillover effects, the spatial lag of X model (SLX), which includes only exogenous spillover, and the and the simultaneous autoregressive (SAR) model, which includes only endogenous spillover. The SDEM combines the SLX and SEM to capture both exogenous spillover and spatial correlation in residuals.

The SDEM was also attractive from an interpretive perspective because it did not require computation of marginal effects; there was no Y variable on the right-hand side, so the model was still linear and coefficients could be interpreted in ways that were conventional for regression models. While the SDEM may be unfamiliar to audiences outside spatial econometrics, its parameters are straightforward compared to SAR or a spatial Durbin model (SDM), which includes a spatially lagged Y term on the right-hand side. As LeSage & Pace (2009, p. 1541) describe, “For the SDEM … models, the coefficients in the vector θ represent local spillovers, since there is an impact only on immediately neighboring observations. The dispersion estimates for SDEM could be readily estimated with standard spatial econometric software.”

The full list of terms and parameter estimates for the SDEM model is included in Appendix 2. SDEM models were estimated with the *spreg* Python package.

Table 4 shows the distributions of predicted land, improvement, and total sale values from the SDEM model. When predicting land and improvement values with the SDEM, we nullified respective improvement and land characteristics specific to each property, but not the lagged characteristics of other properties. This assumed, for example, that improvements from neighboring properties would impact the land value of property even if its own improvements were removed.

While the distribution of total predicted values from SDEM model was approximately similar to that from OLS—the SDEM mean while slightly higher, at $185,797 compared with $174,331 from OLS—decomposed land and improvement values exhibited the same issue with land value inflation that impacted the OLS when block group-level variables were introduced. Predicted land values skewed high, while nearly a third of predicted improvement values were negative. The similarity of this outcome to the block group variable issue from the OLS model makes sense given the SDEM’s heavy reliance on averaging predictors within arbitrary areal units—in the SDEM case, quarter-mile buffers around each property.

The SDEM model automatically included a constant term. When decomposing predictions into land and improvement values, we attributed this constant—$82,539—to land value. Had we instead attributed it to improvement value, the distribution of improvement values would have shifted more positive, but the distribution of land values would have coordinatingly shifted more negative, resulting in similarly unrealistic predictions.

Ultimately, while the SDEM may have better accounted for the existence of spatial lags in valuation, its inability to effectively account for variability in spatial structure of those lags—it assumed lags were equal in all directions from each property—and theoretical difficulty of attributing a substantial constant to either the land or improvement component, made it less effective for predicting decomposed land and improvement values.

Table 4. Distributions of Predicted Land, Improvement, and Total Sales Values from the SDEM Model.

	Decomposed and Total Predicted Values		
	Land	**Improvement**	**Total**
Mean	$189,453	$3,656	$185,797
STD	$59,867	$65,771	$107,948
Min	$5,362	$-855,778	$-12,334
25th Percentile	$144,806	$-17,408	$108,382
50th Percentile	$176,123	$27,289	$153,636
75th Percentile	$223,870	$44,975	$236,171
Max	$494,733	$163,381	$1,102,206
N	97,414	97,414	97,414
N(<$0)	0	31,218	45

Validation

Validating land value models is difficult because pure valuations of land without any degree of improvement, either positive or negative, are rarely observed in transactions. As a form of validation, we compared residuals from our model—the differences between observed sale prices and our modeled sale prices—to the differences between observed sale prices and the fair market values predicted by the internal model used by the Maryland State Department of Assessments and Taxation (SDAT), which were included in MdProperty View (Table 5). We interpreted these latter differences between observed and modeled values as the residuals of the SDAT model.

The mean residual from the SDEM model was more than an order of magnitude smaller than that from SDAT, and the standard deviation of SDEM residuals was about 63% that of the SDAT model. This suggested that our model provided a highly reasonable estimate of overall sales values, in many ways better than the SDAT comparison.

Table 5. Comparison of residuals from SDEM model and SDAT fair market valuation for sales between 2013-2022.

	Residuals for Sales 2013-2022		
	OLS Model	**SDEM Model**	**SDAT**
Mean	$873	$-10,592	$47,220
STD	$87,706	$84,857	$127,517
Min	$-858,551	$-938,394	$-673,975

25th Percentile	$-52,180	$-63,499	$-35,200
50th Percentile	$-7,088	$-16,791	$15,900
75th Percentile	$49,181	$37,817	$89,495
Max	$753,019	$763,903	$800,000
N	97,414	97,414	97,414

Prediction

To produce a citywide surface of predicted land values, we applied the SDEM model trained on 2013-2022 sales to all non-apartment residential properties with full land characteristics available in 2022 (N=134,996). Because these land value predictions were dependent on the sizes of parcels on which properties were located, we standardized land value by parcel acreage (Figure 1). Both the raw and standardized surfaces demonstrated monocentric tendencies reflecting the value of proximity to downtown. Pockets of higher value also emerged around Johns Hopkins University to the north, Patterson Park to the east, and Federal Hill to the south. Distributions for raw and standardized land value predictions are reported in Table 6.

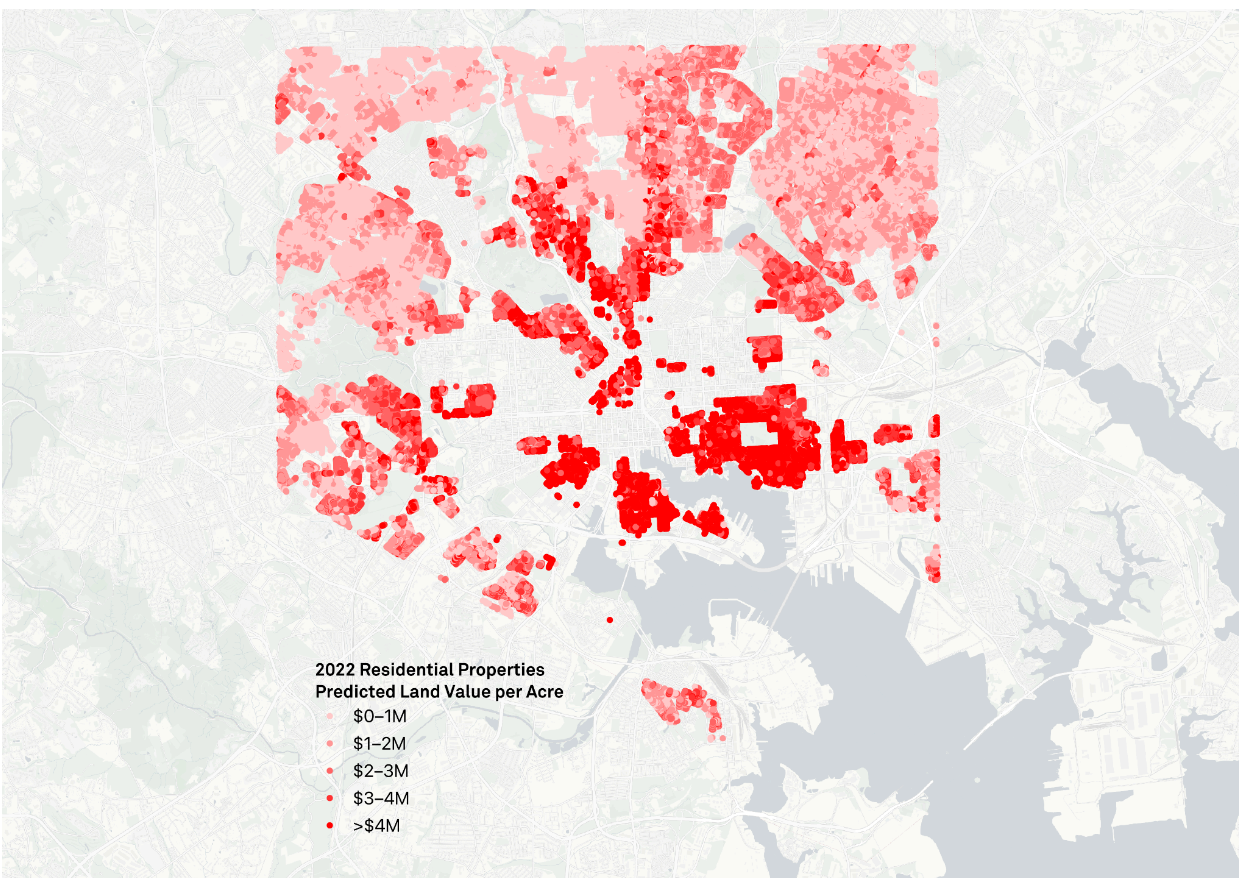

Figure 1. Predicted land value per unit acre for non-apartment residential properties in 2022.

Table 6. Distributions for OLS predicted land value and land value per unit acre for all properties in 2022.

	Predicted Land Value, All 2022 Properties	
	Land Value	**Land Value per Unit Acre**
Mean	$136,105	$3,147,656
STD	$26,318	$2,401,367
Min	$51,147	$17,253
25th Percentile	$118,386	$1,212,189
50th Percentile	$131,854	$2,620,178
75th Percentile	$154,040	$4,260,865
Max	$1,652,401	$26,590,674
N	134,996	134,996

Conclusion

Results from this proof-of-concept project demonstrated the feasibility of predicting land values across a city with heterogeneous built environmental characteristics in ways that broadly conformed with expectations about higher values in central and high-amenity districts. Our models used a relatively small number of input variables gathered from public sources to yield reasonably strong model fit (R^2=0.84). As a proof of concept for scalability, we successfully used input data that were available in consistent forms across Baltimore and neighboring counties and are often available in other U.S. metropolitan areas. We investigated the potential to account for spatial lags in property valuation using an SDEM model, but this framework unrealistically inflated land values, forcing many improvement values to be negative. The simpler and more commonly-used OLS model framework yielded more realistic predictions of property values when decomposed into separate land and improvement components.

A key finding from our model specification process was that predictor variables compiled within areal units, such as census geographies, as opposed to variables that were specific to individual properties or locations, unrealistically inflated land values and diminished improvement values, forcing many improvement values to be negative. While including these variables did not substantially modify total predicted values because lower improvement values compensated for higher land values, they made the model ineffective for predicting separate land and improvement components, which was a key objective of this project. This points to the importance of developing data inputs for land valuation that are as specific to individual properties as possible rather than being joined from arbitrary areal units. Notably, summarizing spatial lags within buffers around each property produced a similar issue in the SDEM model. This indicated that calculating customized areal summaries for each point location would not necessarily be a viable strategy for reducing land value inflation from areal summaries. Further research will be needed to better understand how neighborhood context variables can be represented in land value models while minimizing this effect.

This project benefited greatly from detailed property and sales data from the MdProperty View database, which we used as the units of analysis for modeling and a key source for both improvement and land characteristics. Our heavy use of this database demonstrated the degree to which this type of modeling is contingent on such a data source, but also limited by the variables and data quality provided. MdProperty View did not, ultimately, provide data on improvements with sufficient detail to reliably predict prices for non-residential or high-value properties, likely because these types of properties had characteristics that were too highly specialized or variable to document reliably through the same measures used to represent conventional residential properties. Indeed, even these standard measures, such as square footage, building type, structure quality, year built, were impacted by data inconsistencies. We determined, for example, that properties with vacant or condemned structures, about 1% of properties sold from 2013–2022, tended to be coded as having a square footage of zero despite other characteristics indicating the presence of a structure. This inconsistency still allowed for basic modeling of improvement characteristics with some creative recoding—we developed a vacancy dummy variable to indicate uninhabitable structures—but it revealed structural limitations in accounting for the value of non-standard properties. Even sources such as MdProperty View, which offered relatively consistent property records across broad spatial and temporal extents, had substantial limitations in data specificity and quality.

Despite these limitations, our modeling yielded results that were measurably more consistent with actual sales prices than those provided by SDAT. The average residual on our model of sales between 2013 and 2022 was more than an order of magnitude smaller than those from the SDAT model. Meanwhile, the breakdowns of land and improvement values provided by SDAT were notably different from ours. On average, our land valuations were two times higher than those from SDAT, amounting to a median difference in land value of approximately $28,000. This suggested that the SDAT model was overly sensitive to improvement characteristics, undervaluing the land component.

Future directions for this research could include expansion into more suburban and rural settings, which may pose different opportunities and challenges for representing land characteristics relative to improvements. Suburban settings dominated by single family homes and standardized commercial structures may allow for simplified estimates of improvement values. Rural areas with minimal or agricultural improvements may also streamline improvement estimates, but require different considerations for land value related to natural resources. There are also opportunities to examine urban land value by more directly comparing improved properties to vacant or teardown properties, either cross-sectionally or longitudinally, though this would require more precise data about the status and characteristics of vacancy over time. Because urban properties are only rarely pure greenfields, there will nearly always be a degree of statistical decomposition needed to distinguish between land and improvement values.

References

Ahlfeldt, G., & McMillen, D. (2018). Tall Buildings and Land Values: Height and Construction Cost Elasticities in Chicago, 1870-2010. *The Review of Economics and Statistics*, *100*(5), 861–875. https://doi.org/10.1162/rest_a_00734

Albouy, D., Ehrlich, G., & Shin, M. (2018). Metropolitan Land Values. *The Review of Economics and Statistics*, *100*(3), 454–466. https://doi.org/10.1162/rest_a_00710

Anselin, L., & Lozano-Gracia, N. (2008). Errors in variables and spatial effects in hedonic house price models of ambient air quality. *Empirical Economics*, 34(1), 5–34. https://doi.org/10.1007/s00181-007-0152-3

Bourassa, S. C., & Hoesli, M. (2022). Hedonic, residual, and matching methods for residential land valuation. *Journal of Housing Economics*, *58*, 101870. https://doi.org/10.1016/j.jhe.2022.101870

Can, A. (1992). Specification and estimation of hedonic housing price models. *Regional Science and Urban Economics*, 22(3), 453–474. https://doi.org/10.1016/0166-0462(92)90039-4

Comber, S., & Arribas-Bel, D. (2017). "Waiting on the train": The anticipatory (causal) effects of Crossrail in Ealing. *Journal of Transport Geography*, 64, 13–22. https://doi.org/10.1016/j.jtrangeo.2017.08.004

Corrado, L., & Fingleton, B. (2012). WHERE IS THE ECONOMICS IN SPATIAL ECONOMETRICS?. *Journal of Regional Science*, 52(2), 210–239. https://doi.org/10.1111/j.1467-9787.2011.00726.x

Cunningham, C. (2006). House price uncertainty, timing of development, and vacant land prices: Evidence for real options in Seattle. Journal of Urban Economics, 59(1), 1-31. https://doi.org/10.1016/j.jue.2005.08.003

Davis, M., & Palumbo, M. (2008). The price of residential land in large US cities. Journal of Urban Economics, 63(1), 352-384. https://doi.org/10.1016/j.jue.2007.02.003

Davis, M. A., Larson, W. D., Oliner, S. D., Shui, J., (2021). The price of residential land for counties, zip codes, and census tracts in the united states. Journal of Monetary Economics 118, 413-431. https://doi.org/10.1016/j.jmoneco.2020.12.005

Davis, M. A., Oliner, S. D., Pinto, E. J., & Bokka, S. (2017). Residential land values in the Washington, DC metro area: New insights from big data. Regional Science and Urban Economics, 66, 224–246. https://doi.org/10.1016/j.regsciurbeco.2017.06.006

Diao, M., Leonard, D., & Sing, T. F. (2017). Spatial-difference-in-differences models for impact of new mass rapid transit line on private housing values. *Regional Science and Urban Economics*, 67(May), 64–77. https://doi.org/10.1016/j.regsciurbeco.2017.08.006

Diao, M. (2015). Selectivity, spatial autocorrelation and the valuation of transit accessibility. *Urban Studies*, 52(1), 159–177. https://doi.org/10.1177/0042098014523686

Dubé, J., Legros, D., Thériault, M., & Des Rosiers, F. (2014). A spatial Difference-in-Differences estimator to evaluate the effect of change in public mass transit systems on house prices. *Transportation Research Part B: Methodological*, 64, 24–40. https://doi.org/10.1016/j.trb.2014.02.007

Dubin, R. A. (1992). Spatial autocorrelation and neighborhood quality. *Regional Science and Urban Economics*, 22(3), 433–452. https://doi.org/10.1016/0166-0462(92)90038-3

Dye, R. F., & McMillen, D. P. (2007). Teardowns and land values in the Chicago metropolitan area. *Journal of Urban Economics*, *61*(1), 45–63. https://doi.org/10.1016/j.jue.2006.06.003

Gedal, M., & Ellen, I. G. (2018). Valuing urban land: Comparing the use of teardown and vacant land sales. *Regional Science and Urban Economics*, *70*, 190–203. https://doi.org/10.1016/j.regsciurbeco.2018.03.006

Gibbons, S., & Overman, H. G. (2012). Mostly Pointless Spatial Econometrics? *Journal of Regional Science*, 52(2), 172–191. https://doi.org/10.1111/j.1467-9787.2012.00760.x

Kim, G., Schieffer, J., & Mark, T. (2020). Do superfund sites affect local property values? Evidence from a spatial hedonic approach. *Economic Analysis and Policy*, 67, 15–28. https://doi.org/10.1016/j.eap.2020.05.007

Knaap, E. (2025). *Urban Analysis & Spatial Science*. https://knaaptime.com/urban_analysis/

Koopmans, T. C. (1949). Identification Problems in Economic Model Construction. *Econometrica*, 17(2), 125. https://doi.org/10.2307/1905689

Kuminoff, N. V., & Pope, J. C. (2013). The Value of Residential Land and Structures during the Great Housing Boom and Bust. *Land Economics*, *89*(1), 1–29. https://doi.org/10.3368/le.89.1.1

LeSage, J. P., & Pace, R. K. (2014). Interpreting Spatial Econometric Models. In Handbook of Regional Science (pp. 1535–1552). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-23430-9_91

LeSage, J. P. (2014). What Regional Scientists Need to Know About Spatial Econometrics. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2420725

McMillen, D. P. (2012). Perspectives on Spatial Econometrics: Linear Smoothing with Structured Models. *Journal of Regional Science*, 52(2), 192–209. https://doi.org/10.1111/j.1467-9787.2011.00746.x

Nevo, A., & Whinston, M. D. (2010). Taking the Dogma out of Econometrics: Structural Modeling and Credible Inference. *Journal of Economic Perspectives*, 24(2), 69–82. https://doi.org/10.1257/jep.24.2.69

Rosen, S. (1974). Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *Journal of Political Economy*, 82(1), 34–55. https://doi.org/10.1086/260169

Rosenthal, S. S., & Helsley, R. W. (1994). Redevelopment and the Urban Land Price Gradient. *Journal of Urban Economics*, *35*(2), 182–200. https://doi.org/10.1006/juec.1994.1012

Steimetz, S. S. C. (2010). Spatial Multipliers In Hedonic Analysis: A Comment on "Spatial Hedonic Models of Airport Noise, Proximity, and Housing Prices". *Journal of Regional Science*, 50(5), 995–998. https://doi.org/10.1111/j.1467-9787.2010.00679.x

Thorsnes, P. (1997). Consistent Estimates of the Elasticity of Substitution between Land and Non-Land Inputs in the Production of Housing. *Journal of Urban Economics*, *42*(1), 98–108. https://doi.org/10.1006/juec.1996.2015

Wentland, S. A., Ancona, Z. H., Bagstad, K. J., Boyd, J., Hass, J. L., Gindelsky, M., & Moulton, J. G. (2020). Accounting for land in the United States: Integrating physical land cover, land use, and monetary valuation. *Ecosystem Services*, *46*, 101178. https://doi.org/10.1016/j.ecoser.2020.101178

Won Kim, C., Phipps, T. T., & Anselin, L. (2003). Measuring the benefits of air quality improvement: A spatial hedonic approach. *Journal of Environmental Economics and Management*, 45(1), 24–39. https://doi.org/10.1016/S0095-0696(02)00013-X

Appendix 1 - OLS Model Summary

See Appendix 3 for variable descriptions.

* Improvement variables marked with asterisks. These variables were nullified for prediction of land value.

OLS Regression Results

Dep. Variable:	considr1	R-squared (uncentered):	0.844
Model:	OLS	Adj. R-squared (uncentered):	0.844
Method:	Least Squares	F-statistic:	1.644e+04
Date:	Mon, 23 Mar 2026	Prob (F-statistic):	0.00
Time:	16:58:58	Log-Likelihood:	-1.2470e+06
No. Observations:	97414	AIC:	2.494e+06
Df Residuals:	97382	BIC:	2.494e+06
Df Model:	32		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
*sqftstr	52.3989	0.552	94.957	0.000	51.317	53.480
*age	-759.0649	34.581	-21.950	0.000	-826.843	-691.287
*age2	3.6730	0.202	18.163	0.000	3.277	4.069
*ctrunit	9118.9288	1304.429	6.991	0.000	6562.264	1.17e+04
*endunit	1.033e+04	1302.571	7.933	0.000	7779.701	1.29e+04
*strgrd1	1.709e+04	2.93e+04	0.584	0.559	-4.03e+04	7.44e+04
*strgrd2	-7.831e+04	3.1e+04	-2.524	0.012	-1.39e+05	-1.75e+04
*strgrd4	8.581e+04	831.188	103.239	0.000	8.42e+04	8.74e+04
*strgrd5	1.712e+05	1020.097	167.860	0.000	1.69e+05	1.73e+05
*strgrd6	2.524e+05	1758.581	143.514	0.000	2.49e+05	2.56e+05
*strgrd7	3.359e+05	4903.041	68.512	0.000	3.26e+05	3.46e+05
*strgrd8	3.412e+05	9460.324	36.070	0.000	3.23e+05	3.6e+05
comm	1.497e+04	2607.162	5.741	0.000	9858.766	2.01e+04
mixed	1.755e+04	3192.104	5.498	0.000	1.13e+04	2.38e+04
medres	4.231e+04	1312.221	32.240	0.000	3.97e+04	4.49e+04
lowres	1.084e+05	4140.944	26.182	0.000	1e+05	1.17e+05
vlowres	1.33e+05	6.21e+04	2.140	0.032	1.12e+04	2.55e+05
2014	1.103e+04	1349.922	8.169	0.000	8382.211	1.37e+04
2015	9663.2500	1301.357	7.426	0.000	7112.605	1.22e+04
2016	1.402e+04	1291.832	10.854	0.000	1.15e+04	1.66e+04
2017	2.319e+04	1278.927	18.134	0.000	2.07e+04	2.57e+04
2018	3.579e+04	1306.406	27.396	0.000	3.32e+04	3.84e+04
2019	3.415e+04	1312.668	26.014	0.000	3.16e+04	3.67e+04
2020	3.668e+04	1483.353	24.727	0.000	3.38e+04	3.96e+04
2021	6.46e+04	1240.770	52.063	0.000	6.22e+04	6.7e+04
2022	8.555e+04	1300.100	65.806	0.000	8.3e+04	8.81e+04
sldauto	0.1737	0.014	12.686	0.000	0.147	0.200
sldtran	0.0957	0.003	32.969	0.000	0.090	0.101
hrbrdst	-0.8593	0.174	-4.940	0.000	-1.200	-0.518
sqftland	0.3585	0.038	9.514	0.000	0.285	0.432
arstvlnt	-121.1851	2.593	-46.741	0.000	-126.267	-116.103
treeqtmi	8149.1632	3308.467	2.463	0.014	1664.606	1.46e+04

Omnibus:	20923.098	Durbin-Watson:	1.394
Prob(Omnibus):	0.000	Jarque-Bera (JB):	142532.042
Skew:	0.864	Prob(JB):	0.00
Kurtosis:	8.668	Cond. No.	8.73e+07

Appendix 2 - SDEM Model Summary

See Appendix 3 for variable descriptions.

* Improvement variables marked with asterisks. These variables were nullified for prediction of land value.

```
SUMMARY OF OUTPUT: GM SPATIALLY WEIGHTED LEAST SQUARES (HET) WITH SLX (SLX-Error)
------------------------------------------------------------------------
Data set            :     unknown
Weights matrix      :     unknown
Dependent Variable  :     dep_var   Number of Observations:       97414
Mean dependent var  : 175205.3278   Number of Variables   :          65
S.D. dependent var  : 136259.5541   Degrees of Freedom    :       97349
Pseudo R-squared    :      0.6123
N. of iterations    :           1   Step1c computed       :          No
```

Variable	Coefficient	Std.Error	z-Statistic	Probability
CONSTANT	82539.67991	43926.09195	1.87906	0.06024
*sqftstr	55.58914	1.49876	37.09007	0.00000
*age	-1477.48475	49.53971	-29.82425	0.00000
*age2	5.20116	0.29762	17.47577	0.00000
*ctrunit	-24778.98666	1455.83015	-17.02052	0.00000
*endunit	-22361.09857	1361.47583	-16.42416	0.00000
*strgrd1	-14028.67226	32207.57582	-0.43557	0.66315
*strgrd2	-61439.82288	17183.77960	-3.57545	0.00035
*strgrd4	40410.05803	964.14302	41.91293	0.00000
*strgrd5	86774.35267	1335.02267	64.99841	0.00000
*strgrd6	152913.09857	2426.29978	63.02317	0.00000
*strgrd7	222477.46852	7646.25879	29.09625	0.00000
*strgrd8	199979.99481	19834.25591	10.08256	0.00000
comml	865.51135	2872.03301	0.30136	0.76314
mixed	17643.80364	4224.85406	4.17619	0.00003
medres	17979.93660	1431.02182	12.56440	0.00000
lowres	46615.91941	6611.01205	7.05125	0.00000
vlowres	115409.50784	35692.91291	3.23340	0.00122
2014	4403.60540	1166.11245	3.77631	0.00016
2015	2331.62167	1099.55006	2.12052	0.03396
2016	7089.85579	1087.49662	6.51943	0.00000
2017	16490.32246	1072.03934	15.38220	0.00000
2018	29878.79876	1165.51736	25.63565	0.00000
2019	30079.36222	1126.14016	26.71014	0.00000
2020	34143.48111	1287.40080	26.52125	0.00000
2021	61703.93363	1114.42399	55.36845	0.00000
2022	83915.81471	1194.87656	70.22969	0.00000
sldauto	0.03766	0.08079	0.46612	0.64113
sldtran	0.00331	0.00773	0.42893	0.66797
hrbrdst	-9.95730	5.51117	-1.80675	0.07080
sqftland	0.09831	0.04836	2.03273	0.04208
arstvlnt	-97.02122	11.69116	-8.29868	0.00000
treeqtmi	-9871.73751	10267.90510	-0.96142	0.33634
w_sqftstr	1.99953	6.98856	0.28611	0.77479
w_age	1994.73597	389.62311	5.11966	0.00000
w_age**2	-9.82483	2.31596	-4.24223	0.00002
w_ctrunit	31526.70550	12342.96891	2.55422	0.01064
w_endunit	22989.62374	13084.02400	1.75708	0.07890
w_strgrd1	196042.58966	1052409.19487	0.18628	0.85223
w_strgrd2	116435.16810	1120070.00885	0.10395	0.91721
w_strgrd4	47571.34646	10777.70853	4.41386	0.00001
w_strgrd5	143267.90373	16411.69208	8.72962	0.00000
w_strgrd6	161526.77875	36932.10581	4.37361	0.00001
w_strgrd7	258445.98413	105480.72999	2.45017	0.01428
w_strgrd8	327374.00462	171119.12933	1.91314	0.05573
w_comml	-35507.91322	35939.27563	-0.98800	0.32315
w_mixed	87096.79868	30909.86676	2.81777	0.00484
w_medres	17960.67142	10305.01853	1.74291	0.08135
w_lowres	-80521.09375	41501.21294	-1.94021	0.05235
w_vlowres	528113.84587	229639.96303	2.29975	0.02146
w_2014	-9400.22525	40132.95991	-0.23423	0.81481
w_2015	20707.66798	36216.49120	0.57177	0.56747
w_2016	-18677.66282	33194.05187	-0.56268	0.57365
w_2017	2798.24631	37880.60459	0.07387	0.94111
w_2018	39400.81016	35311.86074	1.11580	0.26451
w_2019	16260.29497	39479.20605	0.41187	0.68043
w_2020	-45656.56088	47562.35426	-0.95993	0.33709
w_2021	17428.26965	35866.19769	0.48592	0.62702
w_2022	-20603.30318	35182.12532	-0.58562	0.55813
w_sldauto	-0.24787	0.27000	-0.91803	0.35860
w_sldtran	-0.00002	0.02855	-0.00073	0.99941
w_hrbrdst	-1.85702	8.07665	-0.22993	0.81815
w_sqftland	0.84396	0.51282	1.64571	0.09982
w_arstvlnt	-72.88746	39.15622	-1.86145	0.06268
w_treeqtmi	104874.98161	29310.06929	3.57812	0.00035
lambda	0.98699	0.01186	83.22325	0.00000

Appendix 3 - Variable Descriptions

Variable	Description
Improvement	
sqftstrc	Structure square footage (0 indicates vacancy)
age	Structure age in years
age2	Structure age squared
Unit Type (Reference: detached unit)	
ctrunit	Center rowhouse unit (dummy)
endunit	End rowhouse unit (dummy)
Structure Grade (Reference: grade 3)	
strgrd1	Structure grade 1 (dummy)
strgrd2	Structure grade 2 (dummy)
strgrd4	Structure grade 4 (dummy)
strgrd5	Structure grade 5 (dummy)
strgrd6	Structure grade 6 (dummy)
strgrd7	Structure grade 7 (dummy)
Land	
Zoning (Reference: high density residential)	
comm	Commercial zoning (dummy)
mixed	Mixed use zoning (dummy)
medres	Medium density residential zoning (dummy)
lowres	Low density residential zoning (dummy)
vlowres	Very low density residential zoning (dummy)
Sale year (Reference: 2013)	
2014	2014 sale year (dummy)
2015	2015 sale year (dummy)
2016	2016 sale year (dummy)
2017	2017 sale year (dummy)
2018	2018 sale year (dummy)
2019	2019 sale year (dummy)
2020	2020 sale year (dummy)
2021	2021 sale year (dummy)
2022	2022 sale year (dummy)

sldauto	Auto job access, EPA Smart Location Database
sldtran	Transit job access, EPA Smart Location Database
hrbrdst	Straight line distance to harbor waterfront
sqftland	Parcel square footage
arstvlnt	Arrests for violent crimes within ¼ mile
treeqtmi	Tree canopy coverage within ¼ mile