Local Polynomial Regressions versus OLS for Generating Location Value Estimates: Which is More Efficient in Out-of-Sample Forecasts?

Local Polynomial Regressions versus OLS for Generating Location Value Estimates

Jeffrey P. Cohen, Cletus C. Coughlin, and John M. Clapp*

October 29, 2015

Abstract

We estimate location values for single family houses using a standard house price and characteristics dataset and local polynomial regressions (LPR), a procedure that allows for complex interactions between the values of structural characteristics and the value of land. We also compare LPR to additive OLS models in the Denver metropolitan area with out-of-sample methods. We determine that the LPR model is more efficient than OLS at predicting location values in counties with greater densities of sales. Also, LPR outperforms OLS in 2010 for all counties in our dataset. Our findings suggest that LPR is a preferable approach in areas with greater concentrations of sales and in periods of recovery following a financial crisis.

Keywords: Land Values, Location Values, Semi-Parametric Estimation, Local Polynomial Regressions

JEL Classification: C14, R51, R53, H41, H54

*University of Connecticut, Federal Reserve Bank of St. Louis, and University of Connecticut, respectively. The authors appreciate the assistance of Brett Fawley, David Lopez, Diana Cooke, and Lowell Ricketts. Participants in the UConn Center for Real Estate 50th Anniversary Symposium and at the NARSC 2014 Annual Meetings provided helpful comments on prior versions of the manuscript. Clapp and Cohen acknowledge support from the Center for Real Estate, University of Connecticut. The views expressed are those of the authors and do not necessarily reflect official positions of the Federal Reserve Bank of St. Louis, the Federal Reserve System, or the Board of Governors.

Introduction

Identically-sized lots and houses in distinct locations in a metropolitan area likely have different market values, a difference many researchers attribute to the value of location since the structure can be renovated or even rebuilt at a similar cost, regardless of its location. The relatively high variability in land value has been well-known by real estate professionals and researchers for many years.[1] However, finding and implementing a theoretically sound and practical method for separating the value of the land (i.e., location) from the value of the housing structure has remained a challenge.[2]

We investigate the separate valuation of urban residential land and structures using house price sales data. Perhaps preferably, sales of vacant land could be used to estimate the value of location. However, vacant land sales are scarce in urban areas and their characteristics (e.g., amount of buildable area, shape and topography) present challenges, implying the need to use sales of properties with structures to infer land value.

Boom and bust cycles in urban house prices provide one motivation for separating land value from structure value: the relative volatility over time of the land value component contributes to macroeconomic risks as suggested by Davis and Palumbo (2008) and by Bourassa *et al.* (2011). Stress testing of mortgage loans would benefit from determining the ratio of structure value to land value because lower ratios imply greater volatility of house prices as implied by the term "land leverage" (Bourassa *et al.* 2011). Moreover, the ratio of structure to land is used by investors to choose the time and intensity of redevelopment (Hendriks, 2005; Dye and McMillen, 2007; Clapp and Salavei, 2010; Ozdilek, 2012).

[1] For example, Diamond (1980) stressed that the price of urban residential land depended primarily on location features and amenities.

[2] Throughout the paper, we use the terms land prices and location values interchangeably. Location value highlights that, as suggested by theory, the right to build at a specific location commands a price.

Even without volatility over time, tax assessors recognize that accurate cross-sectional property valuation must address the very different determinants of location value versus reproducible structural characteristics.[3] Separate estimates of land and building value are used to adjust property tax assessments for structure depreciation and for changes over time in land value. Measuring the percentage of property value attributable to land is important to the literature on capitalization of property taxes, amenities and environmental hazards. Land use planning and regulation over space benefit from more information on areas with relatively high ratios of land value to structure value. Kok, Monkkonen and Quigley (2014) show how changes in land use regulation in the San Francisco Bay Area influence land values indirectly, adding substantially to variation in land prices over space (over 100 municipalities are compared) and time (land values are much more variable than construction costs).[4]

Titman (1985) developed a model of the valuation of vacant urban land when future prices of land plus structure are uncertain. The part of vacant land valuation he finds most relevant is the right but not the obligation to build, i.e. option value. The owner of vacant land has two important decisions: when to build (the decision to build competes with itself delayed) and how big to build (to what "intensity," the amount of structure on a given parcel of land). Irreversibility (it is costly to tear down and start over) gives value to the option to build because an easily reversible decision would imply that every parcel of land would be always near optimal intensity. If, counterfactually, redevelopment were costless, then there would be no reason to hold vacant urban land and all parcels at similar locations would be built to about the same intensity.

[3] Longhofer and Redfearn (2009) point out that property taxes are typically based on total property value. Nevertheless, tax assessors separately estimate and report the two components (Gloudemans, Handel and Warwa, 2002) because this improves the predictive accuracy of their valuations.

[4] They estimate that the average ratio of land value to total residential property value is 32%, a fraction that has been increasing over time.

Clapp, Jou and Lee (2012), Dye and McMillen (2007), Clapp and Salavei (2010) and others develop the importance of irreversibility for urban land with existing structures. These properties can be changed by renovation or by tearing down and rebuilding, and teardowns are observed on the most valuable land. But the cost (the option exercise price) is much more than the sum of demolition and construction cost because the value of the existing structure must be sacrificed in most cases. This contributes to substantial heterogeneity in older urban neighborhoods, where one can see large, recently built or renovated structures next to older, smaller houses.

This line of reasoning shows the fallacy of modeling the value of urban properties as the sum of construction costs (including demolition costs for teardowns) and land value. This sum gives the correct value once an irreversible decision has been implemented. For example, for a newly constructed property or an older structure after the wrecking ball has done irreversible damage, one can find land value by subtracting construction costs from the value of the new property. But before that point, irreversibility implies a complex interaction between structure value and land value.[5]

The implication of this reasoning is that the law of one price (LOP) does not apply within a metropolitan area to the implicit values (shadow prices) of structural characteristics such as interior size or bathrooms. Longhofer and Redfearn (2009) argue that the implicit prices of structural attributes vary within an urban area because many neighborhoods were developed with relatively homogeneous structural characteristics that adjust slowly as the land values in the neighborhood change. They propose an hedonic valuation model using locally weighted regressions with a smoothing function for spatial variation in implicit prices.

[5] Theory for this interaction is provided by Clapp, Jou and Lee (2012), a paper circulated in draft form in 2007. An early empirical demonstration that interaction is relevant is provided by Fik, Ling and Mulligan (2003).

The assumption of partial irreversibility (meaning it is very costly to renovate or teardown and rebuild) by Longhofer and Redfearn (2009), by Gloudemans, Handel and Warwa (2002) and by this study is a major departure from Davis and Polumbo (2008), Kok, Monkkonen and Quigley (2014) and Saiz (2010): they assume a simple additive model where property value is the sum of structure characteristics times their prices and land area times its price per unit.[6] Likewise, the land leverage literature is based on the additive model, despite the fact that the model is most applicable in the special case of recent construction.

We introduce an approach that considers the interaction between structure and land. Our approach combines local polynomial regressions (LPR) with a linear ordinary least squares model. The former provides estimates of the location values of each property at a given point in time, while the latter, using the characteristics of the structure, provides estimates of the value of each structure. A backfitting method ensures orthogonality between location and structure.

To generate location values, our empirical analysis requires only a standard hedonic dataset that includes sales price, location (latitude and longitude), and housing characteristics. This differs from Kok, Monkkonen and Quigley (2014) who have a proprietary dataset for vacant land sales and numerous characteristics for each sale. They must deal with the scarcity of vacant land sales in the most densely populated areas and with very heterogeneous vacant land characteristics:[7]

> "We classified the current condition of these parcels into four categories (i.e., raw, rough graded, fully improved, and previously developed land). The proposed use of these parcels is classified into eight categories (i.e., hold for development, single family, commercial, industrial, multifamily, mixed use, public space, and public facilities). These

[6] For example, Saiz (2010) summarizes the model common to all these papers. "Recall from the model that, on the supply side, average housing prices in a city are the sum of construction costs plus land values (themselves a function of the number of housing units) (p. 1266)."

[7] Likewise, Longhofer and Redfearn (2009) use vacant land sales and their heterogeneous characteristics to estimate the level of land values at various points within Witchita, Kansas. But unlike Kok, Monkkonen and Quigley (2014), a large part of the variation in their estimates of land values comes from variation in the implicit prices of structural characteristics.

> categories, current condition and anticipated use, are presumably important determinants of the cross-sectional variation in land prices (p. 138)."

They deal with the problem of few sales in the more densely populated centers by calculating average land values within each town and eliminating towns with few vacant land sales.

Our spatial smoothing LPR methods are most closely related to Gibbons, Machin and Silva (2013) who improve on boundary fixed effects models by using spatial smoothing for more slowly varying cross-boundary trends related to demographic sorting and other factors. Similarly, Brasington and Haurin (2006) use spatial statistics as part of their identification strategy. They use a nearest neighbor spatial weight matrix that "acts like a highly localized dummy variable," controlling influences such as a nearby abandoned property (p. 260).

The major contributions of our research include our application of a semi-parametric estimation technique using local polynomial regressions (LPR) to separate the value of location from the value of structures. Second, we calculate the out-of-sample root mean squared error (RMSE) for the LPR and ordinary least squares (OLS) approaches, and make comparisons across counties and years to determine when the LPR approach is more efficient than OLS.

Following this introduction, the paper consists of several sections. First is a literature review, followed by a brief summary of the data used in our analyses. Next, we summarize a semi-parametric approach developed by Clapp (2004) for separating land prices from improvements. Our extension of Clapp (2004) is to compare the predictive accuracy of the LPR approach against OLS and to analyze the density required for more efficient LPR performance. We complete the paper with a summary of our main findings.

Estimating Location Values: Existing Research

U.S. housing prices (i.e., the total price that includes land and structures) experienced a dramatic increase in the years leading up to 2006. This boom in housing prices was followed by

a major bust that began in 2006. When one takes a closer look at the U.S. boom and bust, one sees much heterogeneity across regions. Such heterogeneity is described in detail in Cohen, Coughlin and Lopez (2012). A related finding during the boom and bust is that land prices have been more volatile than structure prices.[8]

In this paper, our focus is on residential location values in Denver, a metropolitan area that did not experience the boom and bust extremes of many areas. Our empirical analysis examines three years – 2003, 2006, and 2010 – and five counties – Adams, Denver, Douglas, Arapahoe, and Jefferson – in the Denver metropolitan area. Quarterly housing prices in Denver in the years 2000 through 2014, which includes the boom, bust, and nascent recovery in the U.S. housing market, are shown in Figure 1. Additionally, and explained in detail later, Figure 1 depicts the levels of the time dummy variables in an OLS (hedonic) regression of Denver housing prices for the years of our dataset, when we pool the data and include year fixed effects. It is noteworthy that the general trend in housing prices tracks the trend in the time dummy variables.

[Insert Figure 1 here]

The separation of urban land value from structure value is made challenging by the scarcity of vacant land sales in an urban setting. Hendriks (2005) evaluates three methods used by appraisal professionals for this purpose: fractional apportionment (FAT), rent apportionment (RAT) and price apportionment (PAT) theories. He raises substantial questions about each, recommending that appraisers caution their clients about the unreliability of apportionment methods. Our local regression method (LRM) is most closely related to PAT since it uses sales

[8] Recently, Nichols, Oliner and Mulhall (2013) have found such a result, a finding that is consistent with prior research by Davis and Heathcote (2007), Davis and Polumbo (2008), and Sirmans and Slade (2012).

prices together with location and property characteristics to allocate value (i.e., predicted price from a hedonic model) between land and structure.

Longhofer and Redfearn (2009), who examine how in practice one might disentangle the value of land from the value of structures on the land, argue that land and structures are inseparable, as does Hendriks (2005). Both appeal to an argument that houses within a neighborhood are reasonably homogeneous, in terms of the general size of structure relative to lot size. The Longhofer and Redfearn approach requires data on vacant land sales, and they estimate land values city-wide using locally weighted regressions. In some applications, a lack of vacant land sales data may pose challenges to implementing this approach.[9]

Clapp and Salavei (2010) focus on a different approach than the one in Longhofer and Redfearn (2009). Specifically, they implement an option value approach where existing structure relative to optimal structure at any time will influence the value of the land. There are high adjustment costs, including foregone rents from the existing structure and construction costs, so reaching the redevelopment "trigger point" takes time. Therefore, a property with a given set of characteristics will also have covariant location value and implied prices for these characteristics.

Longhofer and Redfearn (2009) use a nonparametric approach, locally weighted regressions. They allow the valuation of location and structural characteristics to vary smoothly over space. The spatial smoothing method is similar to Clapp (2004) except that he holds the implicit prices of structural characteristics constant and requires orthogonality between structure prices and location values. However, Longhofer and Redfearn (2009) attribute all spatial variation in implicit structure prices to a second stage land valuation equation, so the difference between the two valuation methods may not be great.

[9] In the context of commercial real estate, Haughwout, Orr, and Bedoll (2008) estimate land prices using a dataset that includes purchases of vacant land as well as plots with unoccupied structures slated for demolition and subsequent replacement by new constructions.

The Clapp (2004) LPR approach separates the value of land and improvements with a semi-parametric method. We use LPR in the present paper. Using root mean squared errors, we compare the predictive accuracy of the LPR versus the OLS approaches.

Data Summary

Descriptive statistics for the housing data are presented in Table 1 for Denver. There were over 326,000 observations for single family residential homes that sold between 2003 and 2010 in Denver. The distribution of sales across years was fairly uniform through 2006, then, transactions declined by as much as 50 percent. Still, due to the large sample size, the smallest number of yearly sales, in 2010, was over 20,000. The distribution of sales across counties was reasonably uniform. There were approximately 3.1 bedrooms, with approximately 2.3 full baths and 0.33 half-baths in the typical house sold in Denver over this period. Well over half the houses sold had a garage, a basement and a fireplace. The average sale price was approximately $250,000.

[Insert Table 1 here]

Method for Separating Land and Structure Values[10]

The preceding housing data are used in our method to disentangle location values from structure prices. We follow the LRM and "option value" approach of Clapp (2004) and Clapp and Salavei (2010), respectively. Location value (i.e., the value of the right to build a single family residence at a given location) exhibits more variation across both time and space than

[10] The LRM description parallels the discussion in Cohen *et al*. (2013). Additional details of the LRM are explained in Cohen *et al.* (2014).

structural values, which can be reproduced at the current cost of construction once the redevelopment trigger point has been reached.[11]

First, a parametric method – the standard hedonic model – is used by Clapp (2004) for generating implicit prices for all housing characteristics (structure and location), and a price index independent of these characteristics. He regresses the log of sales price (*lnSP*) on a vector of house structure characteristics (Z), locational characteristics (S), and time (*t= 1…T)*) which is represented here in the form of annual time dummies, Q_t:

$$lnSP_{it} = \gamma_0 + Z_{it}\alpha + S_i\beta + \gamma_1 Q_1 \cdots + \gamma_T Q_T + \varepsilon_{it} \qquad (1)$$

where ε is assumed to be an iid, normally distributed (for the purposes of hypothesis testing) noise term.[12]

We begin by estimating an analogous model to (1), using one OLS regression for all five counties in the Denver area and over all the years 2003-2010. The cumulative log price index for a standard house in the area is measured by the parameters on the annual time dummies, γ. Using our analysis, we plot the price index in Figure 1 as the exponential of each of the time dummies, with 2010 as the base year (which has a value of 100 in Figure 1). In constructing the price index, we assume the structure and location parameters do not vary over time. But since they are not constant over time, over any time interval *T* we are measuring the average implicit prices, α and β. This forces any changes over time into the estimates of the γ parameters; they can be considered an approximation to a pure time component that shifts the constant of the regression, γ_0.

[11] Davis and Palumbo (2008) decompose property value into structure and land components, and find significant changes in land value over time and across metropolitan areas. They subtract the cost of construction from sales prices, while we use the implicit value of the structure.

[12] The natural log of sales price is the dependent variable because logarithms control for heteroscedasticity and some nonlinearity, and enhance degrees of freedom. Hastie and Tibshirani (1990), pp. 52-55, discuss degrees of freedom for smoothing models.

The LPR model differs from equation (1) primarily by estimating the equation at each point on a grid composed of equally-spaced latitude and longitude points that span the data in a given year. If we were to estimate a model based on (1), there would be 15 latitude and 15 longitude points, for a total of 225 "knots" (or target points) on the grid. The size of the bandwidth determines whether or not an observation will be used to estimate the function value at the knot. For the estimation of (1), the bandwidth would be chosen as {0.4σ(latitude), 0.4σ(longitude)} and the bandwidth would be adjusted upward at any target point where there are fewer than 20 observations within one bandwidth. The technical appendix contains a discussion of cross-validation bandwidth selection, methods for dealing with insufficient density of transactions at any target point, estimation of standard errors, and other details of the LPR model.

In this paper, we focus on the nonlinear space relationships for 2003, 2006, and 2010. The semi-parametric LRM model enters because of the "curse of dimensionality." As a practical matter, there would typically be five or six variables for structural characteristics (e.g., interior area, bathrooms) on the right hand side of equation (1). If all were represented by even a coarse grid, the data would be sparse near any point. The semi-parametric solution assumes linearity for the equation (1) parameters, α, on all the housing characteristics.[13] An LPR model is used in the LRM method to estimate these coefficients conditional on the location of the house. This approach addresses the concern of Longhofer and Redfearn (2009) by requiring statistical independence between the estimated coefficients on Z and the nonlinear part of the model.

To implement this logic, the LRM method from Clapp (2004) begins by using ordinary least squares to estimate a cross section version of equation (1) followed by LPR estimation to

[13] Of course, a nonlinear relationship (e.g., with building age) is typically modeled with a quadratic term.

revise the $\hat{\alpha}$'s to assure independence from the location value estimates: the coefficients are the "Robinson" coefficients, $\hat{\alpha}_R$. The Robinson coefficients are estimated for a given year after conditioning the SP and Z variables on latitude and longitude. Then, we subtract the estimated value of structural characteristics to obtain the partial residuals:

$$partres_{it} = lnSP_{it} - Z_i \hat{\alpha}_R \quad (2)$$

where *partres* is the partial residual after subtracting structure value estimated with LPR.
A nonparametric part of the LRM model is:

$$partres_{it} = q(S_i, t_i) + \varepsilon_{it} \quad (3)$$

where S_i is a vector consisting of the latitude and longitude and t_i is the year of sale for house *i*; the model will be separately estimated for any given year of sale. The "backfitting" method iterates between equations (2) and (3) until there is negligible change in $\hat{\alpha}_R$.

In our approach, we focus on the LRM for each county in each of several individual years. Our data set is much more broad than the data from Clapp (2004), as ours covers five counties over an 8 year period (opposed to several years for one town). Furthermore, there is tremendous volatility in sale prices in these counties over the period 2003-2010, so we estimate our LRM separately for each county, for 3 individual years (2003, 2006, 2010). This enables us to assess how the LPR approach performs in a "boom" period (2003), at the beginning of a financial crisis (2006), and in a recovery period (2010). Adapting equation (1) for this specific context leads to:

$$lnSP_i = \gamma + Z_i\alpha + S_i\beta + \varepsilon_i \quad (1')$$

We estimate (1') separately for each county in each of 3 years (2003, 2006, and 2010).

To summarize, Previous work by Clapp (2004) uses LRM estimates as a reasonable approximation to location value in year t_i, $q(\mathrm{S}_i, t_i)$. In that analysis, an average value of structural characteristics, $Z_{\mathbf{1}}\hat{\alpha}_R$, is subtracted from the log of sale price. The estimation method requires statistical independence between location value and improvement value.[14] Our approach is a special case of Clapp (2004). In our specific context, we use LPR to estimate $q(\mathrm{S}_i, t_i)$ at each of 225 target points (15 latitude and 15 longitude) (or "knots") on a grid that spans the data for all sales in any given year t_i.

Results and Performance of the LPR Approach

Table 2 presents coefficients from a standard hedonic regression using the entire dataset.[15] All the OLS coefficients have plausible signs and magnitudes. The time dummy coefficients display a pattern consistent with the Denver house price index (Figure 1). Structural characteristics have

[14] Some, such as Davis and Palumbo (2008), have suggested that location value should be estimated as property value less construction costs. To get to this quantity, one would add back $Z_{\mathbf{1}}\hat{\alpha}_R$ and then subtract construction costs. An approximation to construction costs can be obtained by assuming that they are invariant within the metropolitan area and that they change slowly over time as the costs of material and labor change, and therefore the level of construction costs at time zero is the same for all properties in the city. The Marshall Valuation Service (MVS) is one approach to approximation of this level. Then percentage changes over time can be approximated by using a construction cost indexes such as those published by Engineering News-Record (ENR, http://enr.construction.com/economics/). With these adjustments, location value is estimated by:

$$\hat{q}(\mathrm{S}_i, t_i) + Z_i\hat{\alpha}_R - C_{it}$$

where C_{it} is an estimate of construction costs for house *i* in year *t*. This procedure may be considered as a robustness check.

[15] We subsequently estimate a separate hedonic equation, alongside separate Robinson Coefficients, for each of the five counties in each of 2003, 2006, and 2010. This full set of estimation results is available from the authors upon request.

magnitudes consistent with the literature. In particular, value decreases with structure age at a decreasing rate, a typical result for the housing market. Conversion of the age coefficients to an index equal to 100 for a new house show depreciation of about 1.5% per year declining to near zero at age 30, when the house is worth 80% of its initial value. After that values rise back to 100% at about age 60; this is likely due to renovations of older houses and to restrictions imposed by the quadratic functional form.

[Insert Table 2 here]

Next, we estimate the hedonic model and the Robinson coefficients for each of the five counties in each of 3 years (2003, 2006, 2010), and subsequently obtain the location value estimates for each county in each year.[16]

The Robinson coefficients handle location value (a function of latitude and longitude) in the nonparametric part of the model and they require orthogonality between the two parts of the model. The backfitting method dramatically changes the way location is modeled. The highly constrained hedonic specification for location – the quadratic in latitude and longitude – is replaced by the nonparametric part of the LRM model, equation (3) without the time dimension.

Table 3 contains a sample of the OLS and Robinson coefficient estimates. The results illustrate a number of points. First, while the OLS and Robinson estimates are frequently similar, the differences can be substantial. For example, for Denver County in 2003, the coefficient estimate for land area based on OLS is more than double the Robinson estimate (0.098 vs. 0.046) and for Denver County in 2006 the coefficient for number of bedrooms based on OLS is less than two-thirds the Robinson estimate (0.022 vs. 0.036).[17] Second, both the OLS

[16] The tables of these results for each county in each year are available from the authors upon request.
[17] Lower Robinson coefficients for land area are plausibly related to the LPR model of location value. An extra square foot of land is an amenity which should not be highly priced given that permission has been granted to build

and Robinson estimates vary over time. For example, the OLS estimates for Denver County for number of fireplaces increases from 0.145 in 2003 to 0.156 in 2006 to 0.210 in 2010 and the Robinson estimates for the stories dummy increases from 0.164 in 2003 to 0.199 in 2006 to 0.212 in 2010. Third, both the OLS and Robinson estimates vary across space. For example, a comparison for 2010 shows an OLS estimate for the stories dummy of 0.228 in Denver County and -0.022 in Arapahoe County and a Robinson estimate for the basement dummy of 0.185 in Denver County and 0.049 in Arapahoe County. We conclude that our strategies of estimating separate models for each year, and of comparing OLS and LPR predictions, are likely to produce informative results.

[Insert Table 3 here]

We conduct a set of exercises to compare the predictive accuracy of the two models - OLS and LPR - in estimating location values in the five counties in and around Denver for the years 2003, 2006, and 2010. We conduct two sets of experiments, to compare the efficiency gains of LPR relative to OLS, when we estimate separately for each year (2003, 2006, and 2010, as in (1')). In each of the two scenarios for both models, we run multiple simulations of an out-of-sample forecast to produce estimates of location values, which are then added to the respective structural values to produce an estimated sales price. We compare the estimated sales price to the actual sales price and use root mean squared error to determine which model, OLS or LPR, most accurately estimates location value. The exact steps taken are outlined below.

OLS

To forecast the location values using OLS, we omit 20% of the observations for each individual county in 2003, 2006, and 2010. We use the remaining 80% in each county in each

---

at that location. Small lots that constrain building size are an exception to this rule; evidence supporting this exception is presented in Clapp and Salavei (2010).

year to run the hedonic regression, and forecast the log of sales price for the omitted 20% using the coefficients from the 80% hedonic regression. We then use RMSE to compare the actual 20% to the forecasted 20% for the approach in (1').We repeat this procedure 30 times for each county in each of the 3 years, to account for sample bias.

LPR

To forecast the location values using the LPR technique we omit a random 20% of the sample of observations in each county in 2003, 2006, and 2010. We then obtain the Robinson coefficients with a regression using the remaining 80% of each sample. We forecast the structural values of the remaining 20% using the coefficients from the 80% Robinson coefficients regression. To obtain the partial residuals of the 80%, we subtract the fitted structural values of the 80% from the actual log of the sales price of the 80%.

The 20% subset must be completely contained within the 80% subset: the maximum longitude and latitude of the 20% must be less than the maximum longitude, and latitude of the 80%. Similarly, the minimum longitude and latitude of the 20% must be greater than the minimum longitude and latitude of the 80%. For each county in each year we remove the observations that fail to meet this requirement from the 20% subsets.

Using the LPR technique, we estimate the location values of the 80% from the partial residuals calculated earlier. Then, using bi-linear interpolation, we forecast the location values of the 20%. We add the forecasted location values to the previously forecasted structural values to get an estimate of the log of sales price of the 20% in each year. Finally, we compare the estimated log of sales price against the actual log of sales price of the 20% using the RMSE. We repeat this procedure 30 times to account for sample bias.

Location Values and Simulation Results

Figure 2 is a map showing the counties used in our analysis. We display maps of the location value estimates for two counties – Adams and Denver– in Figures 3 through 8.[18] The initial location value estimates are in natural logs, and we convert these estimates into dollars for ease of interpretation. In Adams County in 2003 (Figure 3), the highest valued locations are in the West, where the county approaches the foothills of the Rocky Mountains. The lowest values are associated with interstate I-25 running north-south and I-76 running from southwest to northeast. In 2006 (Figure 4) the areas of Adams County with relatively high land value have extended north and south and the bands of high value have expanded. In 2010 (Figure 4), after the global financial crisis, the areas of relatively high land value appear to have contracted back towards their 2003 boundaries.[19]

[Insert Figure 2, 3, 4, and 5 here]

Denver County (Figures 6, 7 and 8) shows a roughly monocentric pattern of location value consistent with its coverage of the central business district. Areas of relatively lower valued housing are found due west of the downtown and in the Northeast where I-70 takes traffic towards the airport. The relatively low valued area to the West expanded over time, as did higher values in the Northwest corner of the county.

[Insert Figure 6, 7, and 8 here]

[18] Figures containing the location values for the other three counties are available from the authors upon request. We use Jenks natural breaks classification method, which does not require the same number of observations in each value interval. The objective of the Jenks natural breaks classification method (as described on the ESRI website: http://www.esri.com/industries/k-12/education/~/media/files/pdfs/industries/k-12/pdfs/intrcart.pdf) is to reduce variance within groups and maximize variance between groups. More generally, this is done by seeking to minimize each interval's average deviation from the interval mean, while maximizing each interval's deviation from the means of the other intervals. We found that conclusions using Jenks are not dramatically different than the quintile method, which does require an equal number of observations in each value interval.

[19] Note that each figure reveals relative land values over space in a given year using the Jenks natural breaks classification method. The levels of land values cannot be compared across years because we have not modeled time other than by separating the sample into annual cohorts.

Clapp (2004) describes how reasonably high sales density is crucial for LPR to work better than OLS in Lincoln, Massachusetts. By a large margin for our sample, Denver County is the most dense of the counties in terms of housing sales as well as population. As shown in Table 4, sales density in Denver County, adjusted by area, is more than five times the density in any other county in any of our sample years. This fact leads us to expect LPR to be more efficient than OLS in Denver County and possibly in other counties that are somewhat denser. On the other end, sales in Adams County are the least dense of the five counties. Concerning the remaining counties, Douglas County is always fourth, while Jefferson County (i.e. second in 2003 and third in 2006 and 2010) and Arapahoe County (i.e., third in 2003 and second in 2006 and 2010) switch positions between 2003 and the latter two years.

[Insert Table 4 here]

Next we examine the simulation results for the OLS and LPR models in each county in 2003, 2006, and 2010. These results are presented in Tables 5 through 7.

[Insert Table 5, 6, and 7 here]

The relative efficiency of LPR compared with OLS is quite dramatic in some counties and years, while fairly minimal in others. LPR for Denver County and Arapahoe County performs approximately 8% better than OLS for 2003. LPR for Jefferson County in 2003 performs slightly better than OLS, while LPR for Douglas and Adams Counties perform slightly worse in 2003 than OLS. Given that Denver and Arapahoe are two of the most dense counties, it is not completely surprising that LPR performs better than OLS in those locations.

For 2006, LPR in Denver County performs approximately 20% better than OLS. On the other hand, LPR in Douglas, Adams, and Arapahoe Counties performs only marginally better than OLS. LPR in Jefferson County performs slightly worse than OLS.

Finally, in 2010 LPR performs better than OLS in all counties. The difference in Denver County is again the most pronounced, with an approximate 12% difference. It is also noteworthy that in addition to performing best in the county with the greatest density of sales, LPR also performs better than OLS in all counties in a year that follows a financial crisis (which might be viewed as a time of recovery).

Conclusion

We present a theoretically sound, semi-parametric estimation procedure - local polynomial regressions - to estimate location values. In addition to being grounded in statistical theory, the estimation procedure can be implemented in a straightforward manner with datasets that are commonly used in studies of housing markets. All that is required is data on sale prices, sales dates, and on the associated structural and location characteristics of the properties. We compare the LPR and OLS models using an out-of-sample forecasting procedure. We determine through comparisons of the respective RMSE in each year for each county that in general, the LPR model is more efficient at predicting location values than OLS.

Our results indicate that the (relative) density of sales is a key factor in the performance of our LPR model *versus* a standard OLS model. For Denver County, the densest county in our sample, LPR outperforms OLS in each of three years, with especially large differences for 2006 and 2010. Also noteworthy is our LPR results are better, albeit only marginally in some counties, than the OLS results for 2010. This is a year that can be viewed as a year of recovery following the financial crisis and one that is characterized by fewer sales than in 2003 and 2006.

One potential extension of our analysis would be to include time in the LPR as a third dimension. Preliminary tests indicate that the out-of-sample performance of LPR is degraded when we model all years with 20 grid points, increasing the number of knots from 225 to 4,500.

However, we might model time in two year overlapping intervals to produce a chained price index over time. LPR results for two year intervals might be combined with trilinear interpolation within the same grid as for the LPR estimates so that values for year in which a property did not sell are interpolated in the same way as those in which it did, producing a balanced panel of land values. We can validate the accuracy of the interpolated estimates with an out-of-sample forecasting approach.

There are many potential applications of such an interpolation procedure, together with the LPR estimates, for estimating how various amenities or disamenities are capitalized into land values. After obtaining a balanced panel of location values over time and space, it would be possible to econometrically estimate the determinants of land values, such as school spending, on location values. It would also be possible to assess the impacts other types of public goods, such as parks, on location values. In addition, it is easy to envision the usefulness for other applications, such as house price dynamics driven mostly by changes in land value or taxation of land separately from structures. One potential complication, however, is the fact that these location values are the results of an estimation procedure, so using them again in another estimation procedure implies the resulting p-values may be inefficient.[20]

[20] This issue was pointed out to us by Kelley Pace.

References

Bourassa, S.C., Hoesli, M., Scognamiglio, D., Zhang, S. (2011) Land Leverage and House Prices. *Regional Science and Urban Economics,* 41: 134-144.

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

Clapp, J. M. (2004) A Semiparametric Method for Estimating Local House Price Indices. *Real Estate Economics,* 32: 127-160.

Clapp, J.M., Salavei, K. (2010) Hedonic Pricing with Redevelopment Options: A New Approach to Estimating Depreciation Effects. *Journal of Urban Economics,* 67: 362-377.

Clapp, J.M., J.B. Jou and T. Lee (2012) Hedonic Models with Redevelopment Options Under Uncertainty, *Real Estate Economics*, 40(2): 197-216.

Cohen, J.P., Coughlin, C.C. and Clapp, J.M. (2014) "Semi-Parametric Interpolations of Residential Location Values: Using Housing Price Data to Generate Balanced Panels," Working Papers 2014-50, Federal Reserve Bank of St. Louis.

Cohen, J.P., Coughlin, C.C., Lopez, D.A., Clapp, J.M. (2013) Estimation of Airport Infrastructure Capitalization for Land Value Capture Purposes. *Working Paper WP13JC2,* Lincoln Institute of Land Policy.

Cohen, J.P., Coughlin, C.C., Lopez, D.A. (2012) The Boom and Bust of U.S. Housing Prices from Various Geographic Perspectives. *Federal Reserve Bank of St. Louis Review,* 94: 341-368.

Davis, M.A., Heathcote, J. (2007) The Price and Quantity of Residential Land in the United States. *Journal of Monetary Economics,* 54: 2595-2620.

Davis, M.A., Palumbo, M.G. (2008) The Price of Residential Land in Large US Cities. *Journal of Urban Economics,* 63: 352-384. Data located at Land and Property Values in the U.S., Lincoln Institute of Land Policy. Available at: http://www.lincolninst.edu/resources/.

Diamond, D.B. (1980) The Relationship between Amenities and Urban Land Prices. *Land Economics,* 56: 21-32.

Dye, R.F., McMillen, D.P. (2007) Teardowns and Land Values in the Chicago Metropolitan Area. *Journal of Urban Economics,* 61: 45-64.

Fik, T.J., D.C. Ling and G.F. Mulligan (2003), Modelling Spatial Variation in Housing Prices: a Variables Interaction Approach. *Real Estate Economics*, 31(4): 623-646.

Gibbons, S., Machin, S., Silva, O. (2013) Valuing School Quality Using Boundary Discontinuities. *Journal of Urban Economics*, 75: 15-28.

Gloudemans, R. J., Handel, S., Warwa, M. (2002). An Empirical Evaluation of Alternative Land Valuation Models. Lincoln Institute of Land Policy.

Hastie, T. J., Tibshirani, R.J. (1990) Generalized Additive Models. Chapman & Hall/CRC Monographs on Statistics & Applied Probability.

Haughwout, A., Orr, J., Bedoll, D. (2008) The Price of Land in the New York Metropolitan Area. *Federal Reserve Bank of New York Current Issues in Economics and Finance,* 14: 1-7.

Hendriks, D. (2005) Apportionment in Property Valuation: Should We Separate the Inseparable? *Journal of Property Investment & Finance*, 23: 455-470.

Kok, N., Monkkonen, P., Quigley, J.M. (2014) Land Use Regulations and the Value of Land and Housing: An Intra-Metropolitan Analysis. *Journal of Urban Economics*, 81: 136–148.

Longhofer, S.D., Redfearn, C.L., (2009) Estimating Land Values Using Residential Sales Data. *Working Paper WP09SL1,* Lincoln Institute of Land Policy.

Nichols, J.B., Oliner, S.D., Mulhall, M.R. (2013) Swings in Commercial and Residential Land Prices in the United States. *Journal of Urban Economics*, 73: 57-76.

Ozdilek, U. (2012) An Overview of the Enquiries on the Issue of Apportionment of Value between Land and Improvements. *Journal of Property Research*, 29: 69-84.

Saiz, A., (2010) The Geographic Determinants of Housing Supply. *Quarterly Journal of Economics* 125: 1253–1296.

Sirmans, C.F., Slade B.A. (2012) National Transaction-based Land Price Indices. *Journal of Real Estate Finance and Economics,* 45:829-845.

Titman, S. (1985) Urban Land Prices Under Uncertainty. *The American Economic Review*, 75(3): 505-514.

Table 1: Descriptive Statistics, Denver Single Family Home Sales, 2003-2010

Variable	Mean	Std. Dev.	Variance	Minimum	Maximum
Sale Price (Log)	12.4315	0.5486	0.3010	6.9078	15.4249
Sale Yr 2003	0.1393	0.3463	0.1199	0	1
Sale Yr 2004	0.1530	0.3600	0.1296	0	1
Sale Yr 2005	0.1528	0.3598	0.1295	0	1
Sale Yr 2006	0.1384	0.3453	0.1192	0	1
Sale Yr 2007	0.1232	0.3286	0.1080	0	1
Sale Yr 2008	0.1112	0.3144	0.0989	0	1
Sale Yr 2009	0.0955	0.2940	0.0864	0	1
Sale Yr 2010	0.0865	0.2811	0.0790	0	1
No. of Bedrooms	3.1587	0.8403	0.7061	1	13
No. of Full Baths	2.2924	0.8853	0.7838	1	12
No. of Half Baths	0.3259	0.4938	0.2439	0	5
No. of Fireplaces	0.7669	0.7313	0.5348	0	10
Garage Dummy	0.9103	0.2858	0.0817	0	1
Basement Dummy	0.8031	0.3977	0.1581	0	1
Stories Dummy	0.4839	0.4997	0.2497	0	1
Adams County Sales	0.1900	0.3923	0.1539	0	1
Denver County Sales	0.2268	0.4188	0.1754	0	1
Douglas County Sales	0.1782	0.3827	0.1465	0	1
Arapahoe County Sales	0.2114	0.4083	0.1667	0	1
Jefferson County Sales	0.1935	0.3951	0.1561	0	1
Longitude	-104.9384	0.1441	0.0208	-105.4648	-103.765
Latitude	39.6936	0.1440	0.0207	39.1305	40.242
Longitude Squared	11012.0871	30.2435	914.6665	10767.1109	11122.82
Latitude Squared	1575.6048	11.4253	130.5373	1531.1990	1619.42
Lat*Lon	-4165.3879	16.7768	281.4625	-4206.4213	-4098.49
Age	34.0694	27.3488	747.9552	0	145
Age Squared	1908.6075	2870.4277	8238049	0	21025
Land Sq. Feet (Log)	9.0213	0.6916	0.4765	6.2146	18.1084

Observations = 326,744

Table 2: Hedonic Regression, Denver SFR Home Sales, all counties, 2003-2010

Valid cases:	326744	Dependent variable:	Sales Price (Log)
Missing cases:	0	Deletion method:	None
Total SS:	98336.134	Degrees of freedom:	326717
R-squared:	0.591	Rbar-squared:	0.591
Residual SS:	40228.422	Std error of est:	0.351
F(24,178706):	18150.942	Probability of F:	0

Variable	Coeff.	T-Value	P-Value
Constant	-4233.666469	-13.527658	0.00
2003 Dummy	0.056438	21.222047	0.00
2004 Dummy	0.085489	32.724294	0.00
2005 Dummy	0.121738	46.576639	0.00
2006 Dummy	0.111406	41.845352	0.00
2007 Dummy	0.065053	23.879809	0.00
2008 Dummy	-0.051856	-18.623067	0.00
2009 Dummy	-0.041102	-14.261565	0.00
No. of Bedrooms	0.014671	15.866329	0.00
No. of Full Baths	0.164688	152.781209	0.00
No. of Half Baths	0.151836	91.099069	0.00
No. of Fireplaces	0.170455	161.099640	0.00
Garage Dummy	0.170616	71.760975	0.00
Basement Dummy	0.138199	80.941398	0.00
Stories Dummy	0.014131	8.167266	0.00
Adams County Dummy	0.003398	1.021937	.307
Denver County Dummy	0.183828	61.887987	0.00
Douglas County Dummy	0.016086	3.821778	0.00
Arapahoe County Dummy	-0.021698	-6.430422	0.00
Longitude	-107.719767	-21.667117	0.00
Latitude	-72.155186	-15.752523	0.00
Longitude Squared	-0.799777	-37.149268	0.00
Latitude Squared	-1.115896	-32.965914	0.00
Lat*Lon	-1.529108	-41.419968	0.00
Age	-0.014724	-160.94147	0.00
Age Squared	0.000122	156.981420	0.00
Land Sq. Feet (Log)	0.183304	159.507617	0.00

Table 3 – Illustrative OLS and Robinson Coefficient Estimates

	Denver 2003		Denver 2006		Denver 2010		Arapahoe 2010	
	OLS	Robinson	OLS	Robinson	OLS	Robinson	OLS	Robinson
No. of Bedrooms	0.030	0.040	0.022	0.036	0.019	0.036	0.053	0.052
No. of Full Baths	0.161	0.161	0.171	0.169	0.204	0.198	0.158	0.131
No. of Half Baths	0.127	0.130	0.121	0.127	0.177	0.172	0.136	0.112
No. of Fireplaces	0.145	0.159	0.156	0.159	0.210	0.199	0.124	0.107
Garage Dummy	0.137	0.125	0.184	0.161	0.234	0.195	0.163	0.115
Basement Dummy	0.122	0.141	0.174	0.184	0.183	0.185	0.049	0.049
Stories Dummy	0.169	0.164	0.211	0.199	0.228	0.212	-0.022	0.005
Age	3.4E-04	0.001	-0.004	-0.002	-0.008	-0.006	-0.020	-0.021
Age Squared	2.0E-06	2.0E-06	3.3E-05	2.5E-05	6.0E-05	5.6E-05	1.6E-04	1.4E-04
Land Sq. Feet (Log)	0.098	0.046	0.037	0.032	-0.045	-0.040	0.232	0.202

Table 4 – Population and Sales Densities by County and Year

	Adams			Arapahoe			Denver			Douglas			Jefferson		
	2003	2006	2010	2003	2006	2010	2003	2006	2010	2003	2006	2010	2003	2006	2010
Population	377.5	406.6	443.7	513.7	531.6	574.6	552.6	556.9	603.4	220.4	256.1	286.9	524.9	521.7	535.6
Obs	8756	8166	5179	9476	9656	6038	9510	10118	6607	8681	8764	4768	9390	8511	5675
Area	1182	1182	1182	805	805	805	155	155	155	843	843	843	778	778	778
pop/ sq mi	319.3	344.0	375.4	638.1	660.4	713.8	3567.4	3595.2	3895.2	261.4	303.8	340.4	674.7	670.5	688.4
sales/ sq mi	7.4	6.9	4.4	11.8	12.0	7.5	61.4	65.3	42.7	10.3	10.4	5.7	12.1	10.9	7.3

Table 5 – LPR and OLS Simulation Results by County, 2003

	Adams County	Denver County	Douglas County	Arapahoe County	Jefferson County
OLS RMSE	**0.2071**	**0.2961**	**0.2642**	**0.2626**	**0.2848**
LPR RMSE	**0.2107**	**0.2774**	**0.2686**	**0.2399**	**0.2800**
average # of observations	**8756**	**9510**	**8831**	**9476**	**9390**

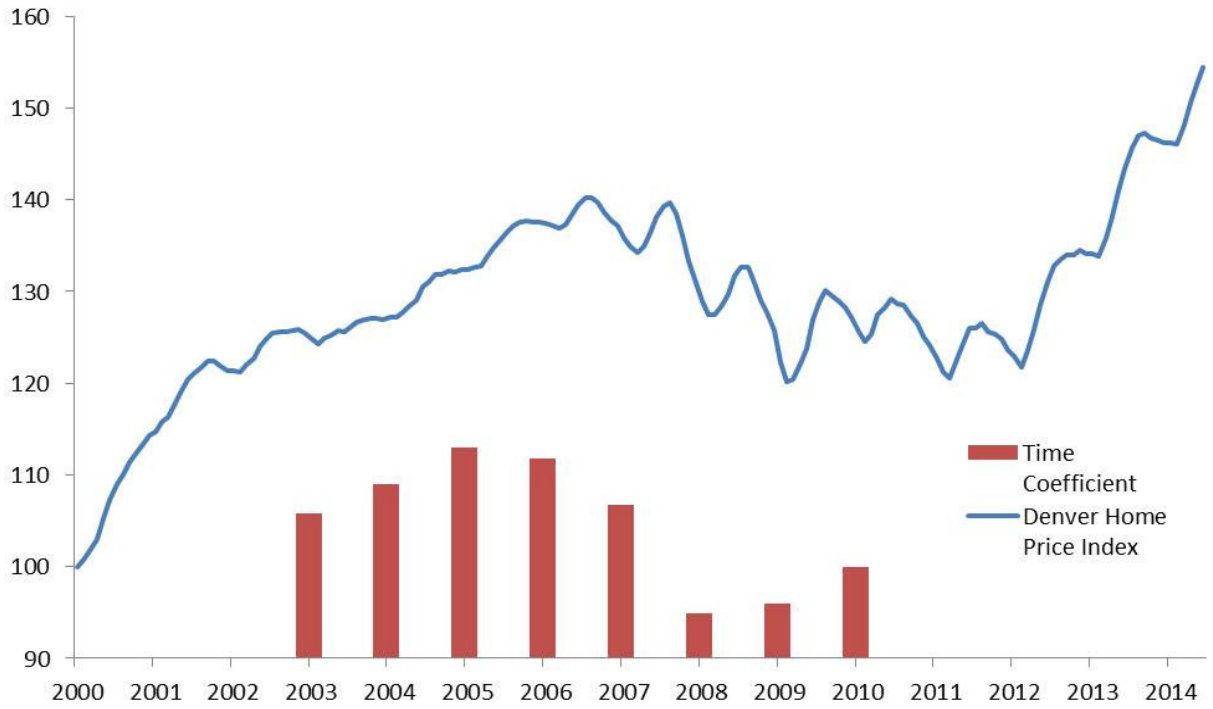
Table 6 – LPR and OLS Simulation Results by County, 2006

	Adams County	Denver County	Douglas County	Arapahoe County	Jefferson County
OLS RMSE	**0.2180**	**0.3554**	**0.4929**	**0.5792**	**0.2389**
LPR RMSE	**0.2105**	**0.2774**	**0.4912**	**0.5699**	**0.2397**
average # of observations	**8166**	**10118**	**8764**	**9656**	**8511**

Table 7 – LPR and OLS Simulation Results by County, 2010

	Adams County	Denver County	Douglas County	Arapahoe County	Jefferson County
OLS RMSE	**0.2459**	**0.4203**	**0.2372**	**0.2705**	**0.2583**
LPR RMSE	**0.2413**	**0.3616**	**0.2329**	**0.2470**	**0.2561**
average # of observations	**5179**	**6607**	**4768**	**6038**	**5675**

Figure 1 – Single Family Home Sale Prices, Denver



Source: Denver Home Price Index is from Federal Reserve Economic Data (FRED); Time dummy coefficient estimates are obtained from hedonic regression in Table 2, normalizing 2010 (the omitted year) to equal 100.

Figure 2 – Denver-area Counties in Analysis

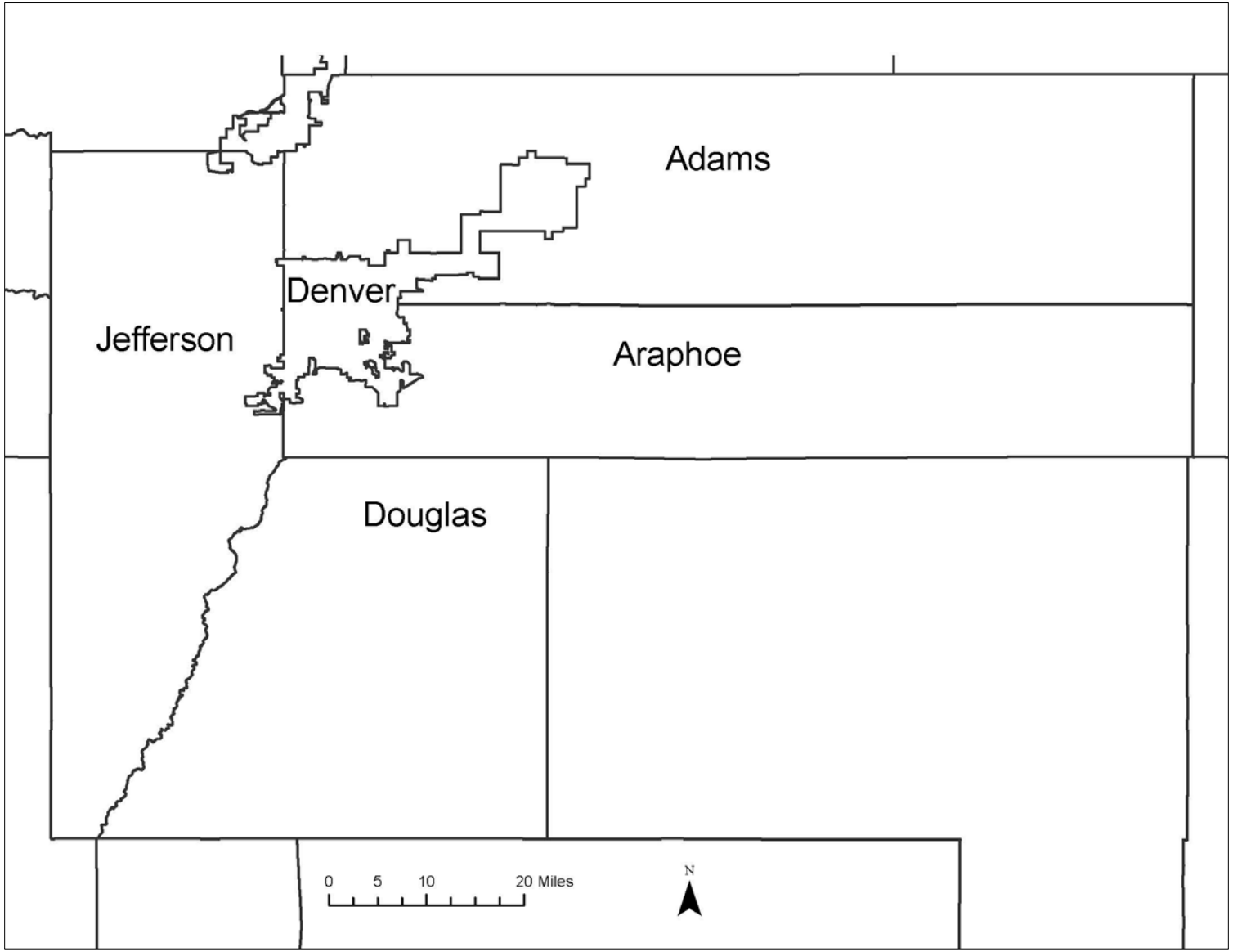


Figure 3 – Adams County Location Values (2003)

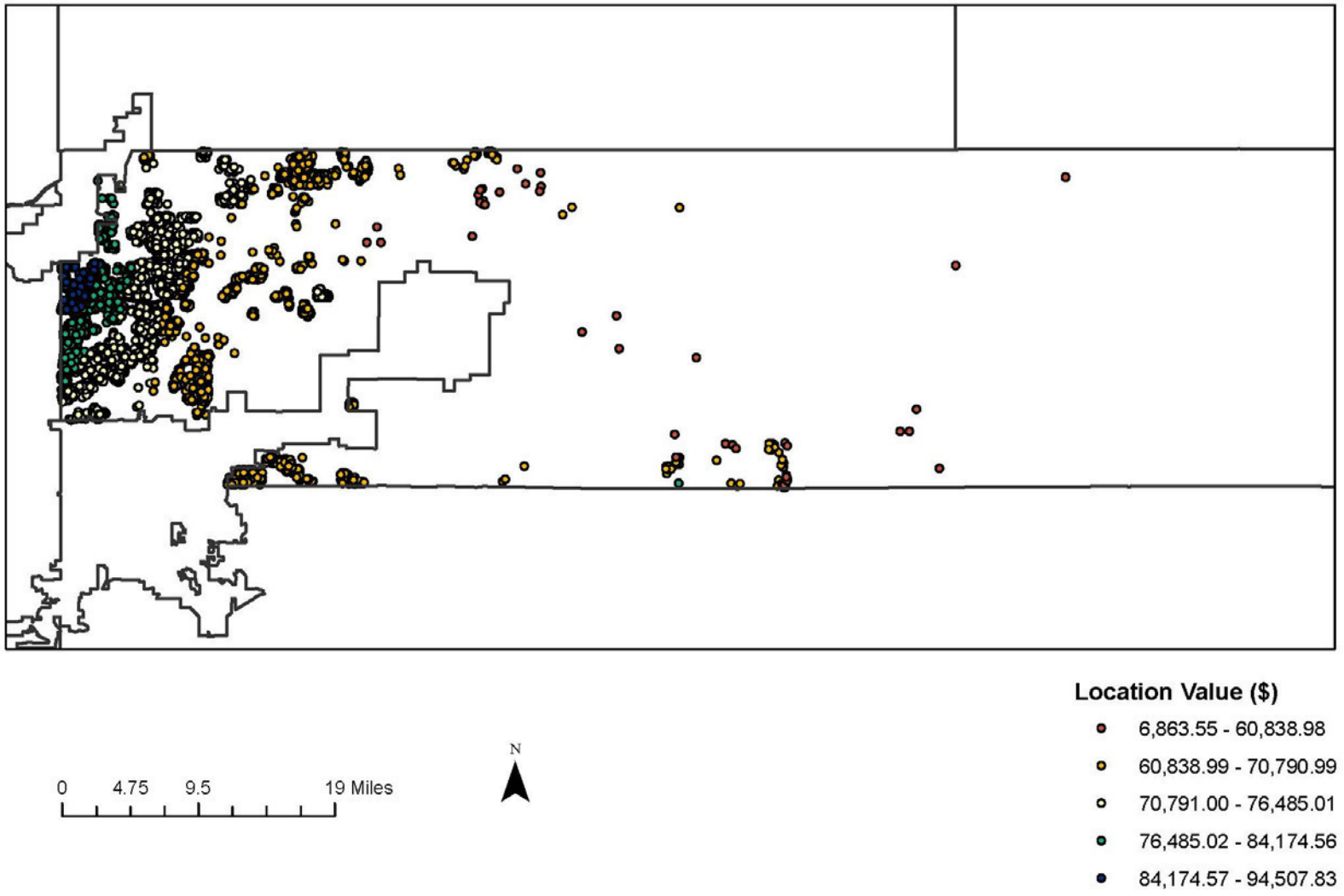


Figure 4 – Adams County Location Values (2006)

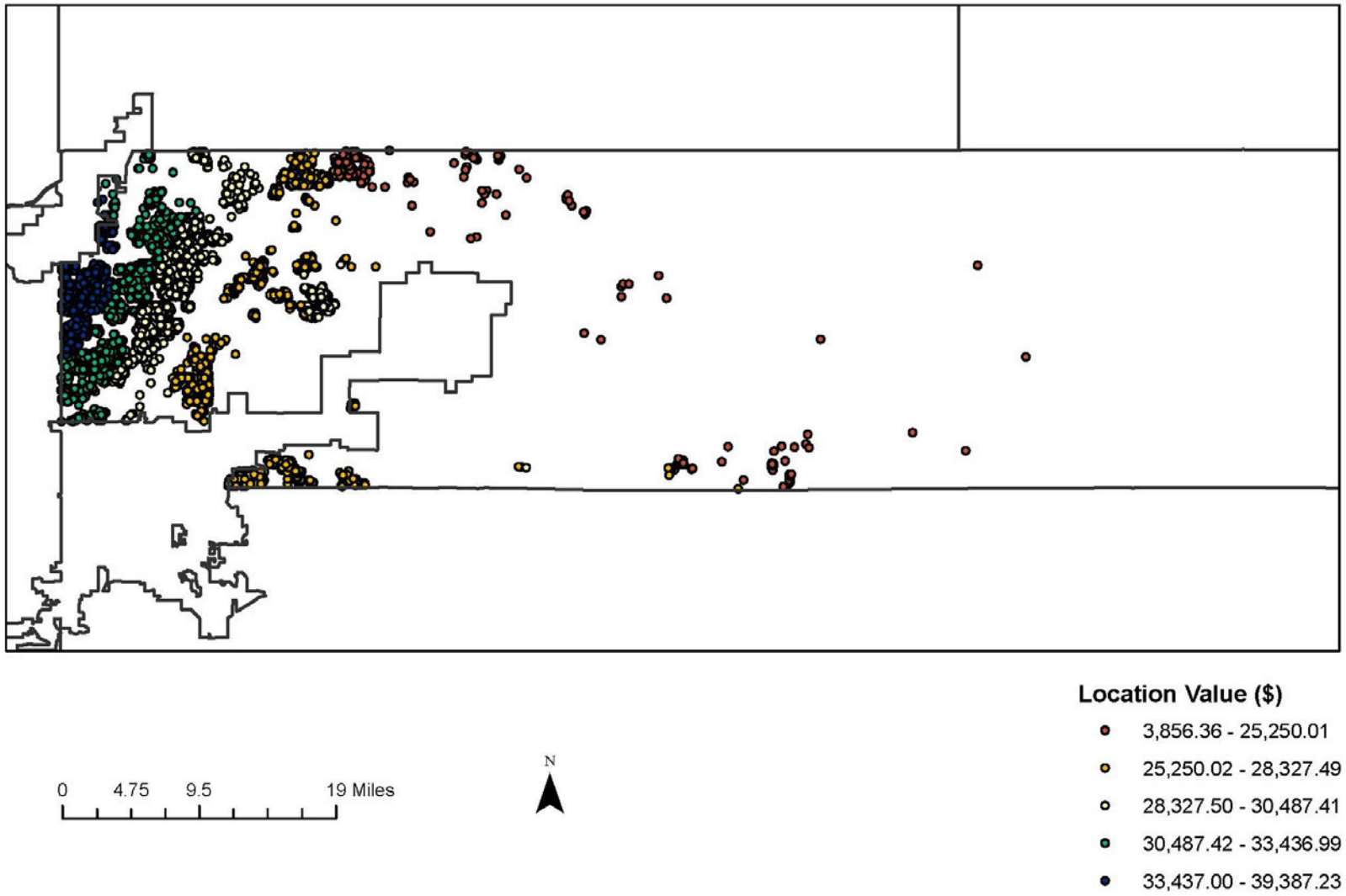


Figure 5 – Adams County Location Values (2010)

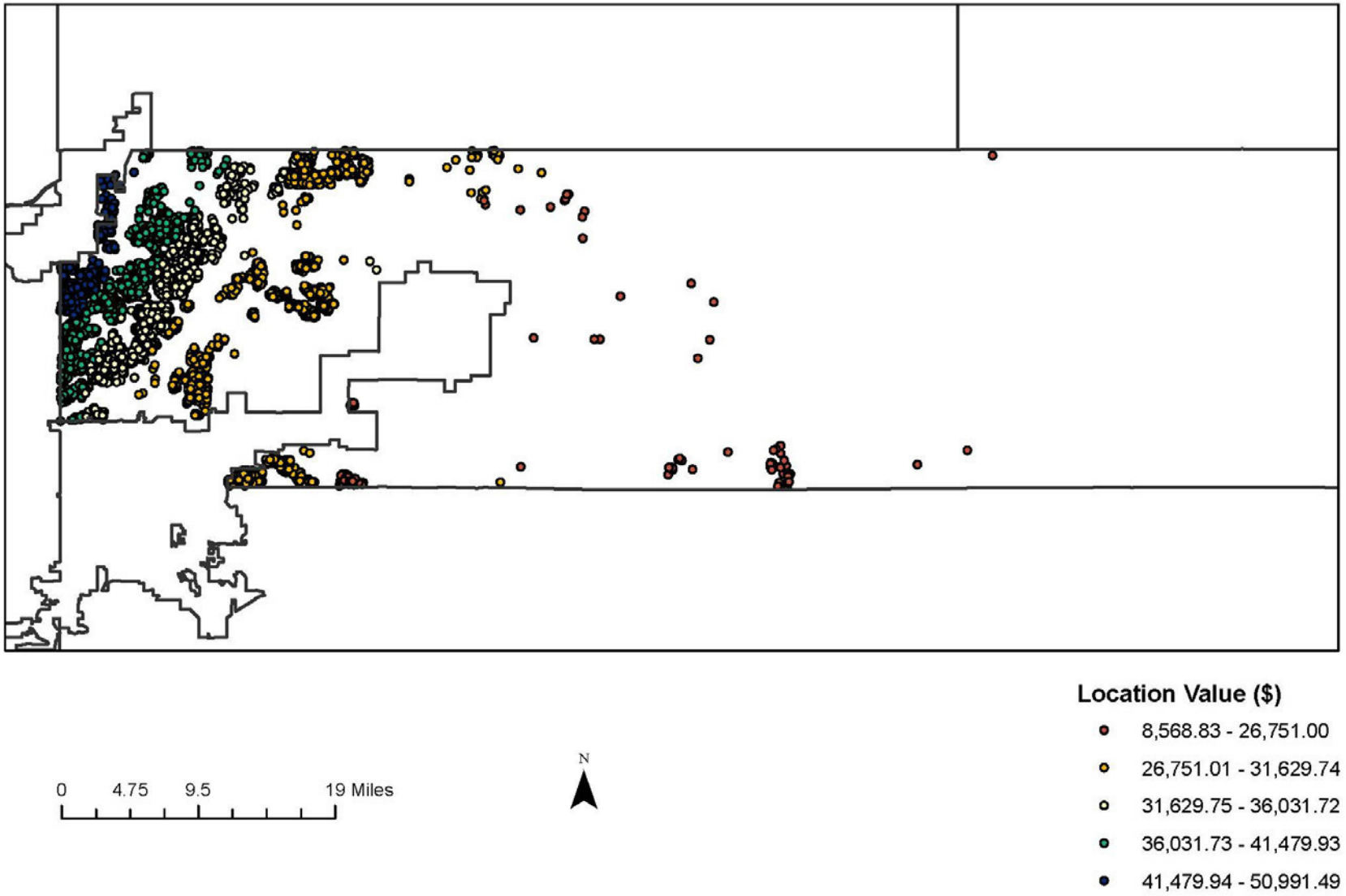


Figure 6 – Denver County Location Values (2003)

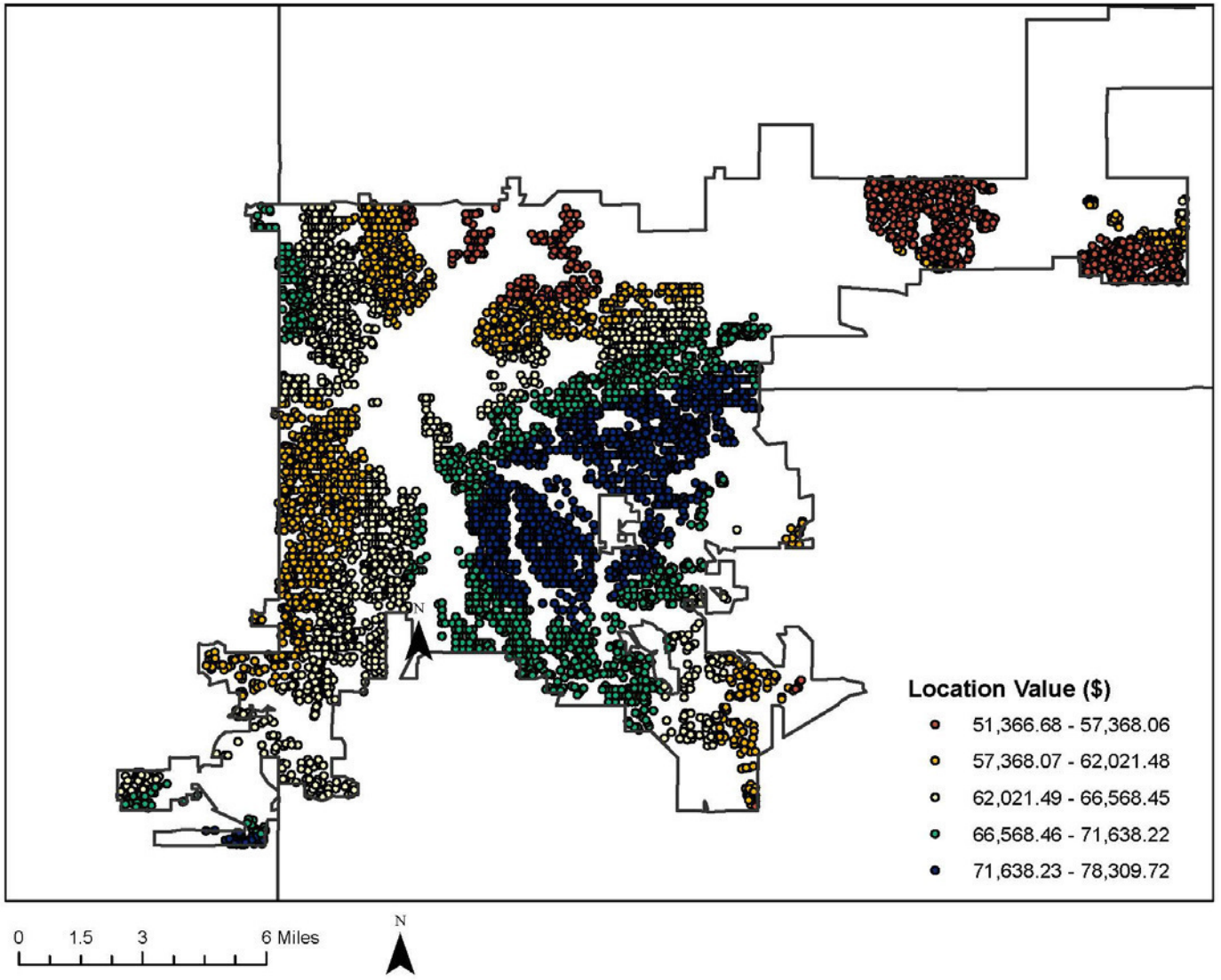


Figure 7 – Denver County Location Values (2006)

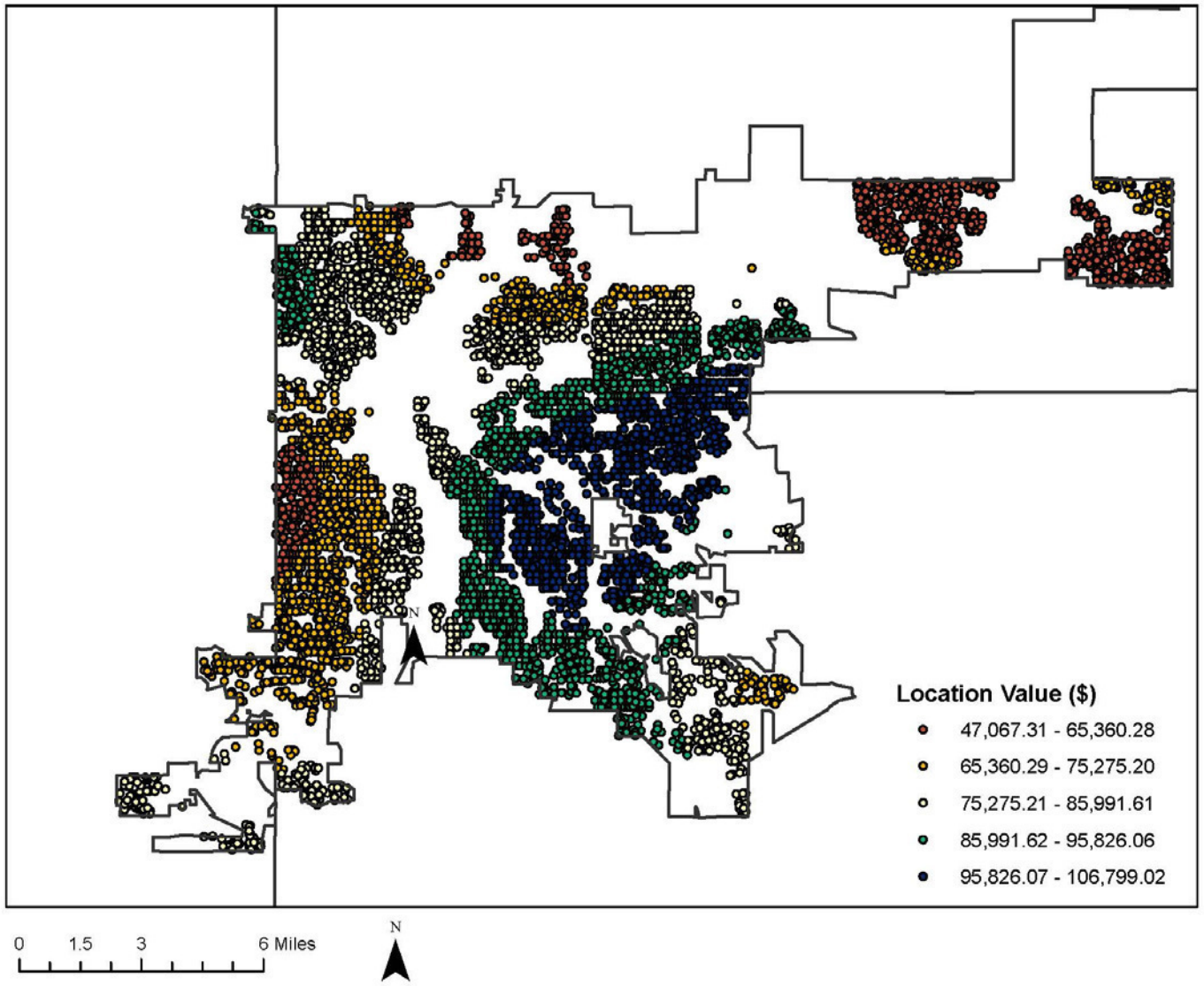


Figure 8 – Denver County Location Values (2010)