Price Indexes for Short Horizons, Thin Markets or Smaller Cities

Authors
Karl L. Guntermann, Crocker Liu, and Adam D. Nowak

Abstract
We propose a modification to the repeat sales procedure that incorporates pricing information from houses in close proximity that have sold only once. This nearest neighbor framework allows for both idiosyncratic property-specific effects and for common location effects in constructing sale pairs. The distinguishing feature of our model is that a significant percentage of discarded (non-repeat sale) observations are utilized. Like the repeat sales approach, the availability of housing attributes is not required to implement our model, in contrast to other hybrid models. The only additional calculation required is pairwise distances, which are used in lieu of property attributes in hybrid models. Our resulting model is not only parsimonious but also produces relatively tighter pricing interval estimates. A key advantage of our approach is that it can be used to obtain price estimates in thin markets and smaller cities that have relatively fewer transactions.

The purpose of our study is to propose and validate a pricing paradigm that retains the essence of the repeat sales regression (RSR) methodology but also incorporates pricing information from nearby houses that have sold only once. The latter information is discarded in implementing the RSR estimation process. The distinguishing feature of our approach is to use the sale price for the most recently sold house that is closest in distance (nearest neighbor) to the house being valued. Intuitively, since the properties in a given area typically have housing attributes that are of similar quality but are not necessarily identical, the difference between the subject property valued and the nearest neighbor’s house that sold will reflect (1) a change in the market price of housing over time, (2) a difference in location-specific flow of services over time, and (3) any difference in price due to property-specific factors. For the special case wherein the nearest neighbor is the subject property and therefore a repeat sale, the difference in price will only reflect a change in the market price of housing over time. One advantage of our proposed approach is that the researcher is now able to utilize nearly 100% of the observations. Another benefit of our approach is that we are able to produce...
tighter pricing interval estimates for the index values. Stated differently, our pricing estimate has relatively lower standard errors compared to the standard error of pricing estimates associated with either the traditional RSR and/or hybrid repeat sale approaches. As such, our method is applicable in thin markets and smaller cities with fewer transactions.

Two popular approaches to property valuation are the hedonic method and the grid method. In the hedonic method, the subject property price is expressed as a function of its property attributes. These property attributes can include both structural attributes such as square footage and locational attributes such as school district. One drawback to hedonic-based models is that the estimation process will always leave relevant attributes out of the pricing equation. In some instances, these attributes are difficult to measure or arbitrary in nature. Most relevant to this study, locational attributes can be difficult to quantify. Previous researchers have recognized the important contribution of spatial or locational attributes in pricing models. More recently, researchers have investigated the price impact due to views unique to the property. Wyman, Hutchison, and Tiwari (2014) find significant premiums of 178%–223% for oceanfront properties in South Carolina but require multiple definitions of ocean view that might not be applicable for other data sets. Gordon, Winkler, Barrett, and Zumpano (2013) examine condominium units in Alabama and find premiums of 3%–12% attributed to both elevation and view. Shultz and Schmitz (2008) find modest premiums of 8% for properties located along a non-recreational lake in Oklahoma. The smaller premium for location near a non-recreational lake calls into question whether or not the premium is due to the view or rather access to water recreation. In all of these studies, the definition of view is at the discretion of the researcher.

An alternative to explicitly specifying the functional form of the view or other locational attributes is to leave the price impact of these attributes in the error term and assume that these attributes are correlated based on location. In this approach, the researcher is interested in estimating either a measure of spatial spillover or would like to control for spatial correlation in the unobserved effects in order to produce the appropriate standard errors. Most relevant to this study, Gelfand, Ecker, Knight, and Sirmans (2004) model micro-level spatio-temporal effects alongside broader market-wide price effects. Conway, Li, Wolch, and Jerrett (2008) examine the effects of greenspace on property values where there are spillover effects due to location. Osland (2010) applies spatial errors to a hedonic model and notes that spatial effects are almost always present in data. McMillen (2010) finds that misspecification produces spatial errors and suggests non-parametric methods as a curative measure. Pace, Barry, Willey, and Sirmans (2000) use spatial effects to forecast real estate prices and Pavlov (2000) uses spatially-varying coefficients in a hedonic regression.

The grid method attempts to select comparable properties that can be used to price a subject property. Frequently, the comparable properties are located in close proximity to the subject property. Comparable properties can be defined using
various criteria including properties that border the subject property, properties located within a certain distance from the subject property or a number of nearest neighbors to the subject property. Implicit in the grid method is a recognition that flows of services due to locational attributes are shared in common with nearby properties. Having specified the comparable criteria, the grid method has been shown to be predictive of the value of comparable properties. Isakson (1986) finds that a nearest-neighbor criterion for selecting comparable properties outperforms a hedonic model. Lai and Wang (1996) compare the hedonic model and the grid method and also find that the grid method produces a smaller prediction error for the value of a subject property. The replication method that Lai, Vandell, Wang, and Welke (2008) propose also uses a nearest neighbor framework, although the authors define nearest neighbors in terms of attribute vectors, whereas we use spatial distance in lieu of attribute vectors since we are more interested in accurately estimating time coefficients by incorporating more differenced sale prices into the RSR.

The method we propose combines features of both the hedonic method and the grid method. We estimate a house price index using regression methods. However, the grid method is used in order to pair properties. The implicit recognition of location in the grid method is used to match properties using a nearest-neighbor approach; the resulting differences in sale prices are then used to augment the observations from the RSR. For a given transaction with no previous sale history, we define the set of potential comparable properties to be all previous sales in the data. Intuitively, we are using comparable properties that sold in lieu of the same properties that have sold multiple times when limited repeat sales exist. The nearest neighbor is then chosen as a comparable property, and the difference in sale prices is formed. This new observation is then used to augment the RSR in a manner similar to other hybrid RSR methods that have been proposed in the literature. Our nearest-neighbor framework implicitly places greater weight on nearby properties (nearest neighbors). In other words, our weighting scheme places all of the weight on the nearest neighbor. We argue that this nearest-neighbor framework results in a parsimonious pricing model in the spirit of the RSR that generates relatively efficient price estimates without the need to rely on nonparametric methods.

Like the repeat sales method, we only require the address and sale price to form sale pairs. However, our method requires calculating pairwise distances while other hybrid approaches require collecting property attributes. These pairwise distance calculations are used to construct paired sales of houses. The pairing of sales is based on the flow of services common to a location (e.g., a sale pair can consist of two distinct homes that share a common flow of locational amenities such as school district or distance to commercial districts). A specialized case of these sale pairs occurs when the same house is resold over time.
Methodology

Models for House Prices and Changes in House Prices

We first describe a model for house prices, the resulting model for differenced house prices, and our proposed method. The model for house price follows Quigley (1995). Log sale price for house $i$ in location $j$ at time $t$, $p_{ijt}$, is represented as the sum of a price index, $\{\delta_t\}_{t=1}^{T}$, and a flow of services, $\beta'x_{ijt}$, where $\beta'x_{ijt}$ is a linear combination of the $K$ observed house attributes. We could also generalize this model and assume that the vector $x_{ijt}$ contains house attributes as well as transformations of the house attributes such as quadratics for square footage, indicator functions for bedrooms, etc. Sale prices are observed over time periods $t = 1, 2, \ldots, T$.

\begin{equation}
 p_{ijt} = \delta_t + \beta'x_{ijt} + e_{ijt}. \tag{1}
 \end{equation}

The term $e_{ijt}$ is a composite error term given by:

\begin{equation}
 e_{ijt} = \mu_i + \lambda_j + \epsilon_{ijt}. \tag{2}
 \end{equation}

Equation (1) is a hedonic model that can be used to estimate a price index. The hedonic model in (1) expresses house prices as a linear function of a time index and observable attributes. This equation for price levels is used to develop a model for price changes. The hedonic model in (1) also plays a direct role in other hybrid methods, such as Case and Quigley (1991) and Hill, Knight, and Sirmans (1997).

Unmeasured attributes are property and location-specific. These attributes can be either quantitative or qualitative. Location attributes include school quality, subdivision amenities, distance to commercial areas, etc. The distinction we make between property-specific and location-specific attributes is that property-specific attributes provide a flow of services only to property $i$ while location attributes provide a flow of services to all properties in location $j$. For example, two neighboring houses might have identical commutes to downtown but different backyard views. In this setting, the backyard view is captured by $\mu_i$, while the commute is captured by $\lambda_j$.

We assume that unobserved attributes do not change over time. However, our model can be modified to accommodate dynamic, unobserved attributes such as unobserved property and/or location-specific attributes that have a temporal correlation. Unreported results show that our conclusions do not change when we incorporate an autoregressive structure in $e_{ijt}$. 
The RSR uses the difference in sale prices. Given a set of \( N \) observations, there are a total of \( N_R \) repeat sales. A repeat sale is defined as a sale for which there exists a previous sale of the same house. For each of these \( N_R \) observations, we can difference the log of sale prices. Using the preceding equations, the difference in sale prices can be expressed as:

\[
\Delta p_{ijt}^R = p_{ijt} - p_{ijt'} = \delta_t - \delta_{t'} + \beta'(x_{ijt} - x_{ijt'}) + \epsilon_{ijt} - \epsilon_{ijt'}.
\]  
(3a)

\[
\Delta p_{ijt}^R = p_{ijt} - p_{ijt'} = \delta_t - \delta_{t'} + \omega_{ijt}^R + \epsilon_{ijt} - \epsilon_{ijt'}.
\]  
(3b)

A common assumption is that the attributes expected change over time is random with \( E[x_{ijt} - x_{ijt'}] = 0 \), which results in (3b). The term \( \omega_{ijt}^R \) captures any unobserved change in attributes over time. Frequently, the researcher only observes sale prices and location in the repeat sales setting, hence the assumption of randomness on the attribute changes.

It is important to note that if the vector \( x_{ijt} \) also includes transformations of the observed attributes, the effect of these transformations of the attributes on the difference in price are eliminated so long as they enter additively into the price as in (1). For example, the correct functional form for price might include a variable with a quadratic term. This quadratic term will be differenced away when the RSR is used if both the variable of interest and the coefficient on the quadratic term are both time-invariant. In this way, the RSR mitigates any problems associated with model misspecification.

The RSR uses only the \( N_R \) differenced sale prices from (3b). While the assumption that hedonic variables do not change precludes the need to collect property attributes, a serious drawback is that the researcher must discard \( N_U = N - N_R \) observations. With this in mind, Case and Quigley (1991) suggest a hybrid approach. This method simultaneously estimates equation (1) for the \( N_U \) unpaired observations and equation (3a) for the \( N_R \) repeat sales with the constraint that \( \delta \) be equal across equations.

In the spirit of the Case and Quigley (1991) hybrid model, we develop an estimation approach that incorporates all observations and controls for unobserved factors \( \lambda_j \). This is important for several reasons. First, early studies did not have access to large data sets and instead used data sets with sales from single condominium buildings or subdivisions that implicitly controlled for location. With recent advances in data collection, storage, and availability, data sets can now include hundreds of thousands of transactions from a single metro area with heterogeneous locations. Given that the heterogeneity is spatial in nature, it is possible to improve on these regression estimates.

Second, attributes tend to vary from one data set to the next. We use data from the respective assessor’s offices in King County, Washington and Maricopa...
While the number of bedrooms is included in the King County data, the number of bedrooms is missing in the Maricopa County data. Further, while both data sets include the square footage for the building structure, only the Maricopa County data includes the square footage of the lot. For this reason, we seek a model like the RSR that mitigates the reliance on observable property attributes.

Although the information provided in real estate data sets do vary, each data set will typically include both the sale price and some form of property identification. In most cases, properties are identified by either address or latitude and longitude. Using a parsimonious data set that includes only sale price and address, a house price index can be estimated using a RSR.

In the subsequent discussion, we introduce an alternative to the RSR that is not only parsimonious but also does not discard $N_U$ observations as the RSR does. We then compare our model to the RSR and other hybrid models that have been proposed in the literature.

**Proposed Nearest-Neighbor Estimator**

For given house $i$ sold at time $t$, we first create the set of all unpaired observations sold at any time period $\tau < t$. Next, we calculate the distance between house $i$ and each of the other houses. We define $n(i, t)$ as the house in the set that is nearest in terms of spatial distance to house $i$. In other words, house $n(i, t)$ is the nearest, previously sold, unpaired neighbor to house $i$. The price and observable attributes for house $n(i, t)$ are used to model the differences in sale prices as:

$$
\Delta p_{ijt} = p_{ijt} - p_{n(i,t)\tau} = \delta_t - \delta_\tau + \beta'(x_{ijt} - x_{n(i,t)\tau}) + e_{ijt} - e_{n(i,t)\tau} + \mu_i - \mu_{n(i,t)},
$$

(4)

This is a one-to-one matching procedure; each unpaired observation is matched to its nearest neighbor. We define the model in equation (4) as the nearest neighbor with attributes (NNA) model. The $N_U$ pairs from equation (4) can be used to augment the $N_R$ repeat sales used in the RSR. This augmentation is similar to the hybrid methods that augment repeat sales with unpaired sale prices in level form. Comparing equation (3b) and equation (4), the additional $N_U$ pairs contain an extra error term and are therefore noisier estimates than the $N_R$ pairs in the RSR. However, the RSR covariance matrix is proportional to $N_R^{-1}$ while the estimator we propose is proportional to $(N_U + N_R)^{-1}$. In the Appendix, we show that if $N_U$ is large enough, the standard errors will decrease when adding noisy observations.

As previously mentioned, we use sales from period $\tau < t$. As such, we cannot match any unpaired sales from $t = 1$. To estimate non-paired sales from $t = 1$ one can use hedonic methods using equation (1).
Although the estimator in (4) requires observable property attributes, we also specify a parsimonious estimator that does not require any observable property attributes. As in the repeat sales methods, we assume that $E[x_{ijt} - x_{n(i),jt}] = 0$ or in vector form that $E[\Delta X_{ijt}] = 0$. With this assumption, the differenced sale prices for nearest neighbors are given by:

$$\Delta p_{ijt}^{(i)} = \delta_t - \delta_r + \omega_{ijt} + e_{ijt} - e_{n(i),jt} + \mu_t - \mu_{n(i),t}$$

We define the model in equation (5) as the parsimonious nearest neighbor (PNN) model. The term $\omega_{ijt}$ represents an error associated with the assumption about the differences in attributes across properties. Equation (5), like the RSR, only requires the time of sale and the address. Thus, the researcher is not required to collect any more data than the RSR requires.

The term $\omega_{ijt}$ represents the differences between the attribute vectors for the nearest neighbor pairs. We do not require that $E[\Delta X_{ijt}] = 0$. For conceptual purposes, suppose that the only attribute vector is square footage. $E[\Delta X_{ijt}] = 0$ implies that we are not forming pairs if the first house in the pair systematically has more square footage than the second house. If this were the case, we would have $E[\Delta X_{ijt}] > 0$. We further investigate our assumption below.

In repeat sales methods, the term $\omega_{ijt}$ represents changes in the attribute vector over time for the same property. Previous studies of repeat sales methods have also assumed that $E[\omega_{ijt}] = 0$. The studies that emphasize generalized least squares (GLS) have also focused on the second moment of the change, $\text{Var}[\omega_{ijt}]$. For example, Bailey, Muth, and Nourse (1963) assume that a constant variance exists: $\text{Var}[\omega_{ijt}] = \sigma^2$. In contrast, Case and Shiller (1989) assume a random walk: $\text{Var}[\omega_{ijt}] = A + B(t - \tau)$ for a house sold at times $t$ and $\tau$. Hill, Knight, and Sirmans (1999) conclude that $\omega_{ijt}$ is stationary. Hill, Knight, and Sirmans (1997) use an AR(1) model and assume that $\text{Var}[\omega_{ijt}] = \sigma_\omega^2 \rho^{t-\tau}$, where $\sigma_\omega^2$ is the per period variance of the innovation. None of these assumptions on the second moment of the error process invalidate the assumption on the first moment of the error process: $E[\omega_{ijt}] = 0$. Table 2 in Cheng, He, Lin, and Liu (2015) provides evidence that the error term follows a random walk for Phoenix and Seattle, the cities in our sample. Based on their result, we assume that the variance increases linearly with time, as in Case and Shiller (1989). We use a generalized (weighted) least-squares procedure for the RSR where the variance is $\text{Var}[\omega_{ijt}] = A + B(t - \tau)$. The parameters $A$ and $B$ are estimated from the squared residuals in a first-stage, repeat sales, ordinary least squares regression.

Equations (3b) and (5) highlight the fact that the RSR and PNN models have the same explanatory variables, but the errors terms are determined by different factors. The error term for the RSR is determined by both a random noise component, as well as changes to the property attributes over time. The error term
in the PNN model also includes a random noise component. However, the error term in the PNN model is determined by (1) attribute differences between nearest neighbors and (2) idiosyncratic location factors as described above. We describe in the Appendix how in some instances, it is possible for the PNN error term to have a smaller variance than the RSR error term. The relative variance of the PNN and RSR error terms plays a direct role in decreasing the standard errors for the resulting least squares estimates.

Competing Models

As previously mentioned, there are many proposed methods for estimating a housing price index. In this section, we discuss several competing models each of which we compare to our own. The first model is a variant of the hybrid model. We simultaneously estimate equation (1) for the unpaired observations and (3a) for the repeat sales:

\begin{align}
p_{ijt} &= \delta_t + \beta' x_{ijt} + e_{ijt}, \\
\Delta p_{ijt} &= \delta_t - \delta_t + \beta' (x_{ijt}^R - x_{ijt}^E) + \epsilon_{ijt}^R - \epsilon_{ijt}^E.
\end{align}

Because our estimator is based on the flow of services associated with location, we also consider a fixed effects model that controls for location using explicit location groups. Using Zip Code fixed effects, we estimate the following:

\begin{align}
p_{ijt} &= \delta_t + \beta' x_{ijt} + \zeta_j + e_{ijt},
\end{align}

where \( \zeta_j \) is the ZIP Code fixed effect for each ZIP Code \( j = 1, \ldots, J \); the location fixed-effects are normalized so that they sum to 0. Although (7) might initially appear to effectively control for unobserved location effects, there are two issues that the researcher must keep in mind. First, the choice of location grouping is debatable as properties can be categorized according to location using ZIP Code, school district or census block. This is important because several school districts each having a different level of education quality might be located within a single ZIP Code.

Second, the large sample properties of a fixed effects estimator typically rely on \( N \to \infty \) while holding the number of regressors, \( K + T + J \), constant. However, in practice \( J \) can be quite large. As such, the resulting assumption of a fixed number of regressors is not valid. This is especially true if the researcher defines location at a more granular level, necessitating a large value of \( J \).

Since our proposed model depends on spatial correlation, a competing model of spatial correlation in errors is also considered.\(^5\) We parameterize the correlation
in the \( e_{ijt} \) of equation (1) using spatial weighting. In particular, we use a single nearest-neighbor spatial error model (SEM) [see Pace, Barry, and Sirmans (1998)] and assume prices are generated by (1) and \( e_{ijt} \) as follows:

\[
e_{ijt} = \lambda e_{n(i,j,t)} + \tilde{e}_{ijt},
\]

(8)

where \( 0 < \lambda < 1 \) is a spatial correlation term, \( e_{n(i,j,t)} \) is the error of the nearest neighbor to house \( i \) sold at time \( t \), and \( \tilde{e}_{ijt} \) is a normally distributed variable with mean 0 and variance \( \sigma_{\tilde{e}}^2 \). The model in (8) is a single nearest-neighbor spatial error model because only the error term for the nearest neighbor appears on the right-hand side of equation (8).

Lastly, we account for potential serial correlation in \( e_{ijt} \). We assume that \( e_{ijt} \) is dynamic and follows an AR(1) process where:

\[
e_{ijt} = \rho e_{ijt-1} + \eta_{ijt}.
\]

(9)

Here \( -1 < \rho < 1 \), \( \eta_{ijt} \) is an iid (independent and identically distributed) random variable with a mean 0 and a variance \( \sigma_{\eta}^2 \). Using the error process in (9), we estimate (1) using a two-step feasible GLS procedure. We first estimate (1) and then calculate the estimated residuals, \( \tilde{e}_{ijt} \). Next, we use the residuals from the paired observations and create differenced, \( \tilde{e}_{ijt} - \tilde{e}_{ij,t-1} \). Our estimator for \( \rho \) is found by first taking expectations of (9), rearranging the terms, and then scaling the result to arrive at:

\[
E[(e_{ijt} - e_{ij,t-1})^2] = 1 - \rho^{t-t-1}.
\]

(10)

This scaling method is similar to that in Quigley (1995). The autoregressive (AR) parameter \( \rho \) is estimated using GMM on (10) where an estimate of \( E[e_{ijt}^2] \) is the mean-squared error of the residuals from (1).

**Competing Matching Procedure**

We also consider an alternative matching procedure. In equations (4) and (5), the nearest neighbor is matched and a differenced sale price is created. We would like to compare our proposed, parsimonious nearest-neighbor method to a method where both distance and property attributes are incorporated into the matching procedure. The resulting index produced by this alternative matching procedure is defined as the nearest neighbor alternative matching (NNAM) estimator.
**Exhibit 1** | Descriptive Statistics of the King County Data Set

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Median</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>SALEPRICE</td>
<td>402.97</td>
<td>330</td>
<td>302.85</td>
<td>1.08</td>
<td>4,980</td>
</tr>
<tr>
<td>SQFT</td>
<td>2.04</td>
<td>1.89</td>
<td>0.90</td>
<td>0.18</td>
<td>14.03</td>
</tr>
<tr>
<td>LOT SQFT</td>
<td>37.94</td>
<td>35</td>
<td>28.78</td>
<td>0</td>
<td>113</td>
</tr>
<tr>
<td>SALE DATE</td>
<td>3.35</td>
<td>3</td>
<td>0.91</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>BATH FIXTURES</td>
<td>1967.57</td>
<td>1969</td>
<td>28.52</td>
<td>1900</td>
<td>2013</td>
</tr>
<tr>
<td>CONSTRUCTION YEAR</td>
<td>0.48</td>
<td>0</td>
<td>0.50</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>GARAGE</td>
<td>0.54</td>
<td>1</td>
<td>0.50</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>CARPORT</td>
<td>0.46</td>
<td>0</td>
<td>0.50</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>PATIOS</td>
<td>0.86</td>
<td>1</td>
<td>0.73</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>POOL</td>
<td>1.53</td>
<td>1</td>
<td>0.65</td>
<td>0</td>
<td>8</td>
</tr>
</tbody>
</table>

Notes: This table displays the descriptive statistics for the Seattle data. SALEPRICE is the sale price measured in $1,000s. SQFT is the square footage measured in 1,000s. AGE is the number of years between the first recorded construction date for the property and the sale year. SALE DATE is the year of sale. BEDROOMS is the number of bedrooms. CONSTRUCTION YEAR is the year of construction. MULTISTORY is an indicator equal to 1 if the property has two or more stories. GARAGE is an indicator variable equal to 1 if the property has a garage. BASEMENT is an indicator variable equal to 1 if the property has a basement. FULLBATHS is the number of half baths. FULLBATHS is the number of full baths.

To construct the NNAM estimator, we define the vector \( z_{ijt} = (x_{ijt}^*, \ Lat_{ijt}^*, Lon_{ijt}^*) \)' as the vector of property attributes, as well as the latitude and longitude coordinates. Each element of \( x_{ijt}^* \), that is not an indicator variable is normalized to have unit variance by dividing by the standard deviations given in Exhibits 1 and 2.\(^6\) For each unpaired house \( i \), we collect the set of all previously unsold houses as before. Next, we calculate the Euclidean distance between \( z_{ijt} \) and the \( z_{kjt} \) for all houses in the set. We define \( n^*(i, t) \) as the house with the smallest Euclidean distance from \( z_{ijt} \). We now create differenced sale prices with this alternative matching procedure similar to equations (4) and (5).

The preceding nearest-neighbor matching procedure is very similar to the NNAM procedure. Both NNA and NNP pairs are formed by minimizing distance using the vector \((Lat_{ijt}, Lon_{ijt})'\). By augmenting the latitude and longitude coordinates with property attributes, houses that are paired using the nearest neighbors will no longer paired using the NNAM if the property attributes are significantly different. Alternatively, houses that have identical attributes but were not defined as nearest neighbors might now be paired. These new pairs created using the NNAM procedure could be miles apart or located on the same block. As an
example, consider three houses located in two suburban neighborhoods of Phoenix. The first house (House A) is a 3-bedroom house located in Fountain Hills. Our second house (House B) is a 3-bedroom house located in nearby Tempe while our third house (House C) is a 4-bedroom house located in Fountain Hills. We can either pair subject house (House A) with either House B or House C. If we pair the houses based on distance, we will pair houses A and C. If we pair the houses based on property attributes, we will pair houses A and B. In this scenario, pairing houses by distance attempts to pair houses with similar unobserved attributes (quality, school district, access to commercial and recreational areas), regardless of the size of the house.

Any difference in price due to observed attributes is explicitly controlled for in equation (4). For this reason, we are interested in comparing a parsimonious NNAM model, as in equation (5), to our PNN model. In other words, the question that we want to address is whether using the NNAM procedure that matches using both observable property attributes and location does better or worse than the
NNAM procedure that only uses location in the matching process. This procedure trades off a smaller error due to the difference in omitted, observable property attributes \( x_{it} - x_{n^*(i,t)} \) for a larger error associated with the differenced unobservable property attributes \( \mu_i - \mu_{n^*(i)} + \lambda_i - \lambda_{n^*(i)} \). This is in contrast to the PNN estimator that attempts to minimize the variance in the error due to an assumed locational correlation in the unobservable attributes.

Of course, the NNAM procedure could be improved on by weighting the attributes and location during the matching process. We do not doubt that an alternative weighting scheme might produce superior results compared to the un-weighted NNAM described above. However, any weighting scheme will add an extra layer of complexity when estimating the house price index. The nearest-neighbor procedure we describe is a special case of the NNAM procedure where all of the weight is placed on the location. This simple weighting procedure inherent in the nearest-neighbor estimator and the resulting performance demonstrate that a simple matching procedure can produce superior results.

Data

The data for this study comes from two sources. The first data source is the King County assessor’s office in the state of Washington. This data set contains all real property transactions for King County starting in 1982 and ending June, 2013. The second data set comes from the Maricopa County assessor’s office in Phoenix, Arizona. This data set contains all transactions for real property in Maricopa County beginning in 1982 and ending in December 2011. For computational reasons, we truncate both data sets prior to January 1, 2000. This truncation does not alter our findings.

The King County data set is chosen because it is available for download at no cost using the assessor’s website. The Maricopa County data were obtained vis-à-vis the private firm Ion Data Express during the time period that the authors were at Arizona State University. Due to financial considerations, it is not possible to compare our method using any other county-wide data sets. However, King County and Maricopa County include the cities of Seattle and Phoenix, respectively, so the results are most likely applicable to counties with large metropolitan areas.

Summary statistics for King County are displayed in Exhibit 1. The average sale price in King County was $413,370 over this 13-year period while the average house size is 2,060 square feet. The average age of a house sold is 36 years. The median sale date is 2005.67, indicating that half of the houses in the sample were sold before August of 2005.

The Maricopa County summary statistics are reported in Exhibit 2. The average house sold in Maricopa County had fewer square feet and sold for less than the average house in King County. However, the houses sold in Maricopa County
Exhibit 3 | Counts for King and Maricopa County

<table>
<thead>
<tr>
<th>Variable</th>
<th>Maricopa County</th>
<th>King County</th>
</tr>
</thead>
<tbody>
<tr>
<td>SALES</td>
<td>719,127</td>
<td>367,474</td>
</tr>
<tr>
<td>REPEAT SALES</td>
<td>312,781</td>
<td>134,681</td>
</tr>
<tr>
<td>ZIP CODES</td>
<td>140</td>
<td>78</td>
</tr>
<tr>
<td>PARCELS</td>
<td>406,346</td>
<td>232,793</td>
</tr>
<tr>
<td>REPEAT SALES / SALES</td>
<td>0.43</td>
<td>0.37</td>
</tr>
<tr>
<td>SALES / ZIP CODE</td>
<td>5,136.62</td>
<td>4,711.21</td>
</tr>
<tr>
<td>PARCELS / ZIP CODE</td>
<td>2,902.47</td>
<td>2,984.53</td>
</tr>
</tbody>
</table>

*Notes:* The table displays the counts by sale in the two data sets. SALES is the total number of sales. REPEAT SALES is the total number of sales where there exists a previous same-house sale. ZIP CODES is the total number of Zip Codes. PARCELS is the total number of unique parcels.

were much newer; the average age is only 15 years, while half of the houses sold are less than 9 years old.

Recall that the RSR discards $N_U$ observations. Exhibit 3 displays the total number of sales and repeat-sales for each county. Discarding the unpaired transactions would discard roughly 57% of the observations in Maricopa County and 63% of the observations in King County. This percentage of unpaired observations is significant considering that the success of our method requires that the number of additional observations must be large enough to dominate the noise from the additional $N_U$ pairs.

**Results**

**Main Findings**

Panel A of Exhibit 4 displays the time coefficients for King County for the NNA and PNN, as well as the hedonic model and the RSR. We display the time coefficients for these models only as the remaining models described above produce similar estimated values. Comparing the NNA to the RSR is our primary focus. Panel A shows that prices in King County increased from the beginning of the sample until they reached a peaked in the summer of 2007. Following the peak, prices started to decline until they reached a trough in the first quarter of 2012. The peak in King County prices was approximately 61% above prices in the 2000:Q1 base period. The trough was approximately 15% above the base period; thus, prices bottomed out at levels that were above 2000:Q1 prices.
Panel A: King County

House Price Index: King County

Panels A and B display regression estimates for the time coefficients.

- PNN: FGLS regression for the model in (5).
- RSR: Weighted repeatsales regression assuming a random walk error term as in equation (3b).
- HED: Hedonic model as in equation (1).
- NNA: FGLS regression for the model in (4).
Panel B of Exhibit 4 displays the time coefficients for Maricopa County. The price series for Maricopa County peaks in 2006:Q2. Following the peak, prices declined until a low is reached in 2009:Q1. In Maricopa County, the time coefficient index for the RSR peaked at roughly 83% above the 2000:Q1 base period and bottomed out at roughly 14% below base period prices. The price index from the hedonic model shows the same pattern but is shifted downwards. As mentioned in Lai, Vandell, Wang, and Welke (2008), the hedonic model can be misspecified or suffer from omitted variables resulting in biased time coefficients; for this reason, practitioners have not fully embraced hedonic methods. As previously mentioned, the RSR mitigates the bias that arises from functional form misspecification. In summary, house price movements in Maricopa County led house price movements in King County. However, the house price peak in King County was larger than the house price peak in Maricopa County. Conversely, the house price trough in Maricopa County was lower than the house price trough in King County.

The primary statistics of interest are the standard errors corresponding to the time coefficients, $\delta_t$. We graph these standard errors in Exhibits 5–8. Exhibits 5 and 6 (7 and 8) are for King County (Maricopa County). Exhibit 5 shows that the FGLS model for NNA that uses our proposed nearest-neighbor estimator has smaller standard errors relative to any competing model that we examine that requires property attributes. The FGLS procedure assumes that the variance in $\omega_{ijt}$ is linear in the time between sales as in Cheng, He, Lin, and Liu (2015). Exhibit 6 shows that the PNN also has smaller standard errors when compared to the traditional RSR, where the RSR is weighted assuming a random walk for the error term. This relative performance is important because both the RSR and the PNN require no additional data other than sale prices and property addresses. The increased performance in the PNN comes at no additional data collection cost to the researcher but instead arises from a more inclusive incorporation of data that is already on hand. Notice that the NNAM procedure has larger standard errors than either the RSR or PNN. This can arise either as the result of the weighting scheme used or the relative magnitude of the errors associated with the unobserved property attributes.

The results for Maricopa County are similar to the results for King County. Exhibit 7 and 8 (Maricopa County) display the same pattern as Exhibits 5 and 6 (King County). The standard errors are smaller for the NNA relative to the standard errors corresponding to any other competing estimator analyzed that requires attributes. In addition, the parsimonious estimators in Exhibit 8 show that the PNN that incorporates sales as in equation (5) produces smaller standard errors than either the RSR or NNAM estimator.

**Localized Property Effects**

In our proposed nearest-neighbor estimator, we assume that there are unobserved location effects. To show that the unobserved error term does include a location effect, we perform an F-test for the joint significance of fixed effects in equation
This figure displays the standard errors for the regression models that require property attributes.

AR: GLS estimation assuming an error process as in (9).
HED: Hedonic model as in equation (1).
NNA: FGILS regression for the model in (4).
HYB: Hybrid model using (6a) and (6b).
SPA: Spatial weighting model as in (8a).
ZIP: Fixed effects model for ZIP Codes as in (7).
Exhibit 6 | Parsimonious Index Standard Errors: King County

This figure displays the standard errors for the time coefficients for the regression models that do not require property attributes.
PNN: FGLS regression for the model in (5).
RSR: Weighted repeat-sales regression assuming a random walk error term as in equation (3b).
NNAM: FGLS regression for the model in [5] when sale prices are matched using distance and attributes in $x_{ijt}$. 
This figure displays the standard errors for the regression models that require property attributes.

AR: GLS estimation assuming an error process as in (9).

HED: Hedonic model as in equation (1).

NNA: FGLS regression for the model in (4).

HYB: Hybrid model using (6a) and (6b).

SPA: Spatial weighting model as in (8a).

ZIP: Fixed effects model for ZIP Codes as in (7).
This figure displays the standard errors for the time coefficients for the regression models that do not require property attributes.

PNN: FGLS regression for the model in (5).
RSR: Weighted repeat-sales regression assuming a random walk error term as in equation (3b).
NNAM: FGLS regression for the model in (5) when sale prices are matched using distance and attributes in x_{ijt}.
(7). Specifically, we test the null hypothesis that all of the ZIP Code coefficients are equal to 0, \( H_0 : \lambda = 0 \). Tests using F-statistic methods can be performed using the data in Exhibits 9 and 10. For Maricopa County, the F-statistic from the change in the sum of squared residuals is distributed as \( F(139,719,074) \). The value for the F-statistic is 1,521.1 while the critical value for the 0.1% level of significance is 1.3. Consequently, strong evidence exists that location effects are significant in pricing after controlling for hedonic attributes. For King County, the F-statistic is 20.16 and is also statistically significant.

In addition to the fixed effects model, we also test for spatial correlation using the spatial model in equation (8). We find that \( \hat{\lambda} = 0.19 \) in King County and \( \hat{\lambda} = 0.31 \) in Maricopa County. We strongly reject the null hypothesis that \( \lambda = 0 \) for both data sets using a likelihood ratio test. Consequently, the spatial correlation model also supports the idea that error terms for properties located near one another are positively spatially correlated.

**Performance in Subgroups**

We are also interested in how our proposed method performs when the sample is bifurcated into old and new properties, as well as small and big properties. A property is defined as an old property if its construction year is older than the median construction year for its respective data set. Exhibits 1 and 2 display the cutoffs used: the cutoff for old houses is 1969 for King County and 1996 for Maricopa County. A big house is defined as a property with square footage larger than the median square footage for all houses built in the same year. We create this group by first sorting houses into bins according to their year of construction. Next, we calculate the median and define any houses with a square footage larger than the median square footage in that year as big.

In unreported results, we find that the ratios of the standard errors for each estimator are constant over time. This is similar to the patterns in Exhibits 5–8. For this reason, we only report the average standard errors over time in Exhibit 11. Since the ratios are constant across time, the ratio in any given time period is equal to the ratio of the average standard errors across time periods. Panels A and B show that the nearest neighbors estimators—the NNA and the PNN—have smaller standard errors relative to either the hedonic or the RSR in general. Thus, our proposed estimator is able to provide uniformly tighter pricing intervals over the subgroups.

**Comparing Attribute Vectors**

In equation (5), we made the assumption that \( E[\Delta X_{ik}] = 0 \). Panel A in Exhibit 12 provides the average and the standard deviation for the difference in nearest neighbor attributes. Individually, it appears that the attribute differences are equal to 0. However, we want to test the joint hypothesis \( H_0 : E[\Delta X_{ik}] = 0, \forall k = 1, 2, \ldots \).
<table>
<thead>
<tr>
<th></th>
<th>AR</th>
<th>HED</th>
<th>NNA</th>
<th>HYB</th>
<th>SPA</th>
<th>ZIP</th>
</tr>
</thead>
<tbody>
<tr>
<td>SQFT</td>
<td>0.3824</td>
<td>0.3828</td>
<td>0.2528</td>
<td>0.4053</td>
<td>0.3618</td>
<td>0.3814</td>
</tr>
<tr>
<td>AGE</td>
<td>0.0014</td>
<td>0.0014</td>
<td>0.0017</td>
<td>0.0018</td>
<td>0.0015</td>
<td>0.0014</td>
</tr>
<tr>
<td>BASEMENT</td>
<td>0.0023</td>
<td>0.0023</td>
<td>-0.0021</td>
<td>0.0027</td>
<td>0.0016</td>
<td>0.0023</td>
</tr>
<tr>
<td>AGE</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>BASEMENT</td>
<td>0.0845</td>
<td>0.0840</td>
<td>0.0210</td>
<td>0.0925</td>
<td>0.0670</td>
<td>0.0841</td>
</tr>
<tr>
<td>BASEMENT</td>
<td>0.0019</td>
<td>0.0019</td>
<td>0.0022</td>
<td>0.0024</td>
<td>0.0020</td>
<td>0.0019</td>
</tr>
<tr>
<td>BEDROOMS</td>
<td>-0.0909</td>
<td>-0.0913</td>
<td>-0.0264</td>
<td>-0.1005</td>
<td>-0.0784</td>
<td>-0.0909</td>
</tr>
<tr>
<td>FULLBATHS</td>
<td>0.0011</td>
<td>0.0011</td>
<td>0.0012</td>
<td>0.0013</td>
<td>0.0011</td>
<td>0.0011</td>
</tr>
<tr>
<td>GARAGE</td>
<td>0.0754</td>
<td>0.0756</td>
<td>0.0148</td>
<td>0.0814</td>
<td>0.0637</td>
<td>0.0753</td>
</tr>
<tr>
<td>GARAGE</td>
<td>0.0018</td>
<td>0.0018</td>
<td>0.0020</td>
<td>0.0024</td>
<td>0.0019</td>
<td>0.0018</td>
</tr>
<tr>
<td>GARAGE</td>
<td>0.0290</td>
<td>0.0287</td>
<td>0.0365</td>
<td>0.0284</td>
<td>0.0279</td>
<td>0.0284</td>
</tr>
<tr>
<td>GARAGE</td>
<td>0.0020</td>
<td>0.0020</td>
<td>0.0022</td>
<td>0.0026</td>
<td>0.0021</td>
<td>0.0020</td>
</tr>
<tr>
<td>HALFBATHS</td>
<td>0.0783</td>
<td>0.0784</td>
<td>0.0185</td>
<td>0.0882</td>
<td>0.0641</td>
<td>0.0782</td>
</tr>
<tr>
<td>HALFBATHS</td>
<td>0.0015</td>
<td>0.0015</td>
<td>0.0016</td>
<td>0.0019</td>
<td>0.0015</td>
<td>0.0015</td>
</tr>
<tr>
<td>MULTISTORY</td>
<td>0.1265</td>
<td>0.1263</td>
<td>0.0747</td>
<td>0.1321</td>
<td>0.1117</td>
<td>0.1257</td>
</tr>
<tr>
<td>MULTISTORY</td>
<td>0.0018</td>
<td>0.0018</td>
<td>0.0022</td>
<td>0.0023</td>
<td>0.0020</td>
<td>0.0018</td>
</tr>
</tbody>
</table>
## Exhibit 9 | (continued)
Hedonic Estimates for King County

<table>
<thead>
<tr>
<th></th>
<th>AR</th>
<th>HED</th>
<th>NNA</th>
<th>HYB</th>
<th>SPA</th>
<th>ZIP</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSE</td>
<td>78,598</td>
<td>78,597</td>
<td>98,349</td>
<td>76,344</td>
<td>76,344</td>
<td>78,265</td>
</tr>
<tr>
<td>MSE</td>
<td>0.2139</td>
<td>0.2139</td>
<td>0.2677</td>
<td>0.2078</td>
<td>0.2289</td>
<td>0.2131</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.4522</td>
<td>0.4522</td>
<td>0.9071</td>
<td>0.9967</td>
<td>0.4105</td>
<td>0.4545</td>
</tr>
<tr>
<td>N</td>
<td>367,474</td>
<td>367,474</td>
<td>367,474</td>
<td>367,474</td>
<td>367,474</td>
<td>367,474</td>
</tr>
<tr>
<td>DF</td>
<td>367,412</td>
<td>367,412</td>
<td>367,412</td>
<td>367,412</td>
<td>367,412</td>
<td>367,335</td>
</tr>
</tbody>
</table>

Notes: This table displays the standard errors for the hedonic coefficients for the regression models. SQFT is the square footage measured in 1,000s. AGE is the number of years between the first recorded construction date for the property and the sale year. BASEMENT is an indicator equal to 1 if the property has a basement. BEDROOMS is the number of bedrooms. FULLBATHS is the number of full baths. GARAGE is an indicator variable equal to 1 if the property has a garage. HALF BATHS is the number of half baths. MULTISTORY is an indicator variable equal to 1 if the property has two or more stories. SSE is the total sum of squared errors. MSE is the mean squared error. DF is the total degrees of freedom. All variables are significant at the 1% confidence level.

AR: A GLS estimation assuming an error process as in (9).
HED: A baseline hedonic model as in equation (1).
NNA: A FGLS regression for the model in (4).
HYB: A hybrid model using (6a) and (6b).
SPA: A spatial weighting model as in (8).
ZIP: A fixed-effects model for ZIP Codes as in (7).
### Exhibit 10 | Regression Estimates for Maricopa County

<table>
<thead>
<tr>
<th></th>
<th>AR</th>
<th>HED</th>
<th>NNA</th>
<th>HYB</th>
<th>SPA</th>
<th>ZIP</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SQFT</strong></td>
<td>0.2504</td>
<td>0.2514</td>
<td>0.2294</td>
<td>0.2505</td>
<td>0.2613</td>
<td>0.2562</td>
</tr>
<tr>
<td></td>
<td>0.0013</td>
<td>0.0013</td>
<td>0.0012</td>
<td>0.0017</td>
<td>0.0016</td>
<td>0.0011</td>
</tr>
<tr>
<td><strong>LOT SQFT</strong></td>
<td>0.0306</td>
<td>0.0305</td>
<td>0.0186</td>
<td>0.0294</td>
<td>0.0283</td>
<td>0.0258</td>
</tr>
<tr>
<td></td>
<td>0.0002</td>
<td>0.0002</td>
<td>0.0002</td>
<td>0.0003</td>
<td>0.0003</td>
<td>0.0002</td>
</tr>
<tr>
<td><strong>AGE</strong></td>
<td>−0.0013</td>
<td>−0.0013</td>
<td>−0.0065</td>
<td>−0.0013</td>
<td>−0.0030</td>
<td>−0.0079</td>
</tr>
<tr>
<td></td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0001</td>
<td>0.0001</td>
<td>0.0001</td>
<td>0.0001</td>
</tr>
<tr>
<td><strong>BATH FIXTURES</strong></td>
<td>0.0448</td>
<td>0.0447</td>
<td>0.0093</td>
<td>0.0507</td>
<td>0.0293</td>
<td>0.0164</td>
</tr>
<tr>
<td></td>
<td>0.0004</td>
<td>0.0004</td>
<td>0.0003</td>
<td>0.0005</td>
<td>0.0004</td>
<td>0.0003</td>
</tr>
<tr>
<td><strong>CARPORT</strong></td>
<td>−0.0335</td>
<td>−0.0337</td>
<td>−0.0159</td>
<td>−0.0220</td>
<td>−0.0300</td>
<td>−0.0100</td>
</tr>
<tr>
<td></td>
<td>0.0024</td>
<td>0.0024</td>
<td>0.0019</td>
<td>0.0031</td>
<td>0.0028</td>
<td>0.0003</td>
</tr>
<tr>
<td><strong>GARAGE</strong></td>
<td>−0.1846</td>
<td>−0.1840</td>
<td>−0.0471</td>
<td>−0.1614</td>
<td>−0.1369</td>
<td>−0.0849</td>
</tr>
<tr>
<td></td>
<td>0.0023</td>
<td>0.0023</td>
<td>0.0019</td>
<td>0.0029</td>
<td>0.0027</td>
<td>0.0020</td>
</tr>
<tr>
<td><strong>POOL</strong></td>
<td>0.0002</td>
<td>0.0002</td>
<td>0.0001</td>
<td>0.0003</td>
<td>0.0002</td>
<td>0.0001</td>
</tr>
<tr>
<td></td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td><strong>PATIOS</strong></td>
<td>0.0289</td>
<td>0.0292</td>
<td>0.0120</td>
<td>0.0311</td>
<td>0.0232</td>
<td>0.0233</td>
</tr>
<tr>
<td></td>
<td>0.0007</td>
<td>0.0007</td>
<td>0.0006</td>
<td>0.0010</td>
<td>0.0009</td>
<td>0.0006</td>
</tr>
<tr>
<td><strong>AC</strong></td>
<td>0.2068</td>
<td>0.2077</td>
<td>0.0592</td>
<td>0.3362</td>
<td>0.1227</td>
<td>0.1231</td>
</tr>
<tr>
<td></td>
<td>0.0157</td>
<td>0.0158</td>
<td>0.0108</td>
<td>0.0198</td>
<td>0.0171</td>
<td>0.0134</td>
</tr>
</tbody>
</table>
## Exhibit 10

Regression Estimates for Maricopa County

<table>
<thead>
<tr>
<th></th>
<th>AR</th>
<th>HED</th>
<th>NNA</th>
<th>HYB</th>
<th>SPA</th>
<th>ZIP</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSE</td>
<td>104,255</td>
<td>104,253</td>
<td>92,948</td>
<td>96,730</td>
<td>96,730</td>
<td>74,445</td>
</tr>
<tr>
<td>S2</td>
<td>0.1450</td>
<td>0.1450</td>
<td>0.1293</td>
<td>0.1345</td>
<td>0.1754</td>
<td>0.1035</td>
</tr>
<tr>
<td>R²</td>
<td>0.6004</td>
<td>0.6004</td>
<td>0.9421</td>
<td>1.0048</td>
<td>0.4915</td>
<td>0.7146</td>
</tr>
<tr>
<td>N</td>
<td>719,127</td>
<td>719,127</td>
<td>719,127</td>
<td>719,127</td>
<td>719,127</td>
<td>719,127</td>
</tr>
<tr>
<td>DF</td>
<td>719,074</td>
<td>719,074</td>
<td>719,074</td>
<td>719,074</td>
<td>719,074</td>
<td>718,935</td>
</tr>
</tbody>
</table>

Notes: This table the standard errors for the hedonic coefficients for the regression models. **SQFT** is the square footage measured in 1,000s. **LOT SQFT** is the lot square footage measured in 1,000s. **AGE** is the number of years between the first recorded construction date for the property and the sale year. **BATH FIXTURES** is the number of fixtures in the house. A bathroom with a sink, shower, and bathtub would have three fixtures. **CARPORT** is an indicator variable equal to 1 if the property has a carport. **GARAGE** is an indicator variable equal to 1 if the property has a garage. **POOL** is an indicator variable equal to 1 if the property has a pool. **PATIOS** is the number of patios. **AC** is an indicator variable equal to 1 if the property has central air. Standard errors are below. All variables are significant at the 1% confidence level.

- **AR**: A GLS estimation assuming an error process as in (9).
- **HED**: A baseline hedonic model as in equation (1).
- **NNA**: A FGLS regression for the model in (4).
- **HYB**: A hybrid model using (6a) and (6b).
- **SPA**: A spatial weighting model as in (8).
- **ZIP**: A fixed-effects model for ZIP Codes as in (7).
Exhibit 11 | Average Standard Errors

<table>
<thead>
<tr>
<th>Panel A: King County subgroups</th>
<th>HED</th>
<th>NNA</th>
<th>RSR</th>
<th>PNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Observations</td>
<td>0.0083</td>
<td>0.0073</td>
<td>0.0107</td>
<td>0.0082</td>
</tr>
<tr>
<td>New Houses</td>
<td>0.0101</td>
<td>0.0092</td>
<td>0.0127</td>
<td>0.0106</td>
</tr>
<tr>
<td>Old Houses</td>
<td>0.0130</td>
<td>0.0114</td>
<td>0.0169</td>
<td>0.0120</td>
</tr>
<tr>
<td>Big Houses</td>
<td>0.0123</td>
<td>0.0105</td>
<td>0.0157</td>
<td>0.0115</td>
</tr>
<tr>
<td>Small Houses</td>
<td>0.0111</td>
<td>0.0100</td>
<td>0.0145</td>
<td>0.0106</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Panel B: Maricopa County subgroups</th>
<th>HED</th>
<th>NNA</th>
<th>RSR</th>
<th>PNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Observations</td>
<td>0.0050</td>
<td>0.0034</td>
<td>0.0063</td>
<td>0.0038</td>
</tr>
<tr>
<td>New Houses</td>
<td>0.0062</td>
<td>0.0047</td>
<td>0.0068</td>
<td>0.0047</td>
</tr>
<tr>
<td>Old Houses</td>
<td>0.0072</td>
<td>0.0052</td>
<td>0.0105</td>
<td>0.0057</td>
</tr>
<tr>
<td>Big Houses</td>
<td>0.0067</td>
<td>0.0043</td>
<td>0.0083</td>
<td>0.0048</td>
</tr>
<tr>
<td>Small Houses</td>
<td>0.0072</td>
<td>0.0054</td>
<td>0.0094</td>
<td>0.0058</td>
</tr>
</tbody>
</table>

Notes: Panels A and B display the average of the standard errors for the time coefficients within subgroups, across all time periods. Averages are calculated using the arithmetic average.

HED: Hedonic model as in equation (1).
NNA: FGLS regression for the model in (4).
NNP: FGLS regression for the model in (5).
RSR: Weighted repeat sales regression assuming a random walk error term as in equation (3b).

..., K variables. To test this hypothesis, we perform a Wald test. For each nearest-neighbor sale pair, we calculate the differenced vector of attributes $\Delta X_U$. Next, we estimate the covariance matrix of $\Delta X_U$, $\text{Var} [\Delta X_U] = \Sigma_U$. Assuming that $\Delta X_U$ satisfies the central limit theorem, the Wald test statistic is calculated as $W = \Delta X'_U \Sigma^{-1}_U \Delta X_U$. Under the null hypothesis, this statistic converges in distribution to a Chi-squared random variable with K degrees of freedom. Panel B in Exhibit 12 displays the Wald statistics for the Maricopa and King County data sets, as well as various subsets. The Wald statistics support the idea that $E[\Delta X_U] = 0$ (e.g., we cannot reject this null hypothesis for either city or in any subset).

Theoretically, we assume in our model that houses located near to each other in geography will also be similar to each other with respect to their attributes. The results in Panel B of Exhibit 12 tend to support this assumption. We are also interested in documenting the correlation between two vectors of nearest neighbors. We do this by writing the differenced nearest-neighbor attribute vectors as:
### Exhibit 12 | Nearest Neighbor Attribute Vectors

#### Panel A: Sample average and standard deviation

<table>
<thead>
<tr>
<th></th>
<th>King County</th>
<th>Maricopa County</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SQFT</strong></td>
<td>0.013</td>
<td>0.015</td>
</tr>
<tr>
<td></td>
<td>(0.314)</td>
<td>(1.033)</td>
</tr>
<tr>
<td><strong>LOT SQFT</strong></td>
<td>0.109</td>
<td>−0.142</td>
</tr>
<tr>
<td></td>
<td>(5.164)</td>
<td>(766.476)</td>
</tr>
<tr>
<td><strong>AGE</strong></td>
<td>0.186</td>
<td>0.004</td>
</tr>
<tr>
<td></td>
<td>(52.079)</td>
<td>(1.413)</td>
</tr>
<tr>
<td><strong>BATH FIXTURES</strong></td>
<td>0.025</td>
<td>0.007</td>
</tr>
<tr>
<td></td>
<td>(4.12)</td>
<td>(0.379)</td>
</tr>
<tr>
<td><strong>CARPORT</strong></td>
<td>−0.001</td>
<td>0.009</td>
</tr>
<tr>
<td></td>
<td>(0.106)</td>
<td>(0.322)</td>
</tr>
<tr>
<td><strong>GARAGE</strong></td>
<td>0.004</td>
<td>0.007</td>
</tr>
<tr>
<td></td>
<td>(0.108)</td>
<td>(0.868)</td>
</tr>
<tr>
<td><strong>POOL</strong></td>
<td>0.013</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>(5.981)</td>
<td>(0.696)</td>
</tr>
<tr>
<td><strong>PATIOS</strong></td>
<td>0.006</td>
<td>−0.005</td>
</tr>
<tr>
<td></td>
<td>(0.524)</td>
<td>(0.346)</td>
</tr>
<tr>
<td><strong>AC</strong></td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.002)</td>
<td></td>
</tr>
</tbody>
</table>

#### Panel B: Wald test

<table>
<thead>
<tr>
<th></th>
<th>King County</th>
<th>Maricopa County</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>All</strong></td>
<td>0.0007</td>
<td>0.0031</td>
</tr>
<tr>
<td><strong>Small</strong></td>
<td>0.0009</td>
<td>0.0028</td>
</tr>
<tr>
<td><strong>Big</strong></td>
<td>0.0014</td>
<td>0.0014</td>
</tr>
<tr>
<td><strong>Old</strong></td>
<td>0.0003</td>
<td>0.0026</td>
</tr>
<tr>
<td><strong>New</strong></td>
<td>0.0025</td>
<td>0.0043</td>
</tr>
</tbody>
</table>

**Notes:** Panel A displays the sample average and standard deviations for the nearest neighbor attribute vectors. The variables are those in Exhibits 9 and 10. Standard deviations are in parentheses. Panel B provides information about the attribute vectors for repeat sale pairs and nearest neighbor pairs; the columns display the test statistic for the asymptotic Wald test. The null hypothesis is $H_0: E(\Delta \mathbf{X}_u) = 0$, where $\Delta \mathbf{X}_u$ is the difference in attribute vectors for the nearest neighbor sale pairs. There are nine variables in the vector $\Delta \mathbf{X}_u$ for Phoenix and eight variables in the vector for Seattle. We cannot reject the null hypothesis for the entire sample or any subsample.
Exhibit 13  | Correlation between Attribute Vectors of Unpaired Sales and Nearest Neighbors

<table>
<thead>
<tr>
<th>Maricopa County</th>
<th>King County</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Correlation 2.5 Percentile 97.5 Percentile</td>
</tr>
<tr>
<td>All</td>
<td>0.930       0.427                   0.429</td>
</tr>
<tr>
<td>Small</td>
<td>0.222       0.115                   0.120</td>
</tr>
<tr>
<td>Big</td>
<td>0.502       0.360                   0.363</td>
</tr>
<tr>
<td>Old</td>
<td>0.454       0.283                   0.287</td>
</tr>
<tr>
<td>New</td>
<td>0.409       0.197                   0.203</td>
</tr>
</tbody>
</table>

Notes: This table displays the correlation between the attribute vectors for unpaired sales and nearest neighbors, \( \rho^{nn} \). Correlation is the correlation between the vectorized matrices \( \text{vec}(X_U) \) and \( \text{vec}(X^{nn}_U) \) given in equation (11). The table also provides these correlations when a random sample of houses is used instead of the nearest neighbor. The 2.5 Percentile and 97.5 Percentiles are calculated using non-parametric methods. For \( s = 1, 2, \ldots, 500 \), we select \( N_U \) random attribute vectors from the entire data set or subset without replacement and store them in a matrix \( X_i \). We then create \( \text{vec}(X_i) \), calculate the correlation between \( \text{vec}(X_U) \) and \( \text{vec}(X_i) \) called \( \rho^i \), and store each \( \rho^i \) in the vector \( \rho^{1:500} = (\rho^1, \rho^2, \ldots, \rho^{500}) \), where we use the correlations in \( \rho^{1:500} \) to construct confidence intervals for \( \rho^{nn} \). The columns 2.5 Percentile and 97.5 Percentile give the lower and upper bounds for the middle 95% samples in \( \rho^{1:500} \).

\[
\Delta X_U = X_U - X^{nn}_U, \quad (11)
\]

where \( X_U \) is the attribute vector of sales where there is no repeat sale and \( X^{nn}_U \) is the attribute vector of nearest neighbors. To measure the correlation between these two vectors, we first vectorize \( X_U \) and \( X^{nn}_U \) and then calculate the correlation between \( \text{vec}(X_U) \) and \( \text{vec}(X^{nn}_U) \) called \( \rho^{nn} \). Exhibit 13 displays the correlation between these matrices for the entire data set and all of the subsets we examine. The average correlation between these attributes is 0.93 for the Maricopa County data set and 0.80 for the King County data set, indicating that the attributes for any unpaired house and its nearest neighbor are highly correlated in general. Other correlation variations are also considered. For example, we also computed \( \rho^{nn'} \) as the average correlation of the diagonal elements in the correlation matrix between \( X_U \) and \( X^{nn'}_U \); neither the results nor the results in Exhibit 13 change substantially when this alternative measure is used.

We also compare these correlations when a random sample of houses is used instead of the nearest neighbor. For \( s = 1, 2, \ldots, 500 \), we select \( N_U \) random attribute vectors from the entire data set or subset without replacement and store them in a matrix \( X' \). We next create the vector \( \text{vec}(X') \), calculate the correlation between \( \text{vec}(X_U) \) and \( \text{vec}(X') \) denoted by \( \rho^i \), and store each \( \rho^i \) in the vector...
The correlations in $\rho^{1:500}$ are used to construct confidence intervals for $\rho^m$. Duranton and Overman (2005) select random subsamples in a similar manner when constructing confidence intervals for their agglomeration measure. Exhibit 13 displays the middle 95% range in correlations for these random simulations. For the entire data set and for all subsets, the correlation in attributes when using nearest neighbors lies above the 95% confidence interval. The data in Exhibits 12 and 13 provide evidence that is consistent with the assumption that $E[\Delta X_{ij}] = 0$ is appropriate and that attributes for houses in close proximity are highly correlated above and beyond what we would find when a random pairing process is used.

**Conclusion**

Existing variations of the repeat sale paradigm either exclude the price information on houses that trade only once during a given interval and/or omit information that is available on housing attributes. We propose a method that uses the same information as the traditional repeat sales method but also incorporates this discarded information. Our method also requires the computation of pairwise distance calculations since our method uses a nearest neighbor framework. One advantage of our approach is that almost all transactions are used. Another benefit of our method is that the standard errors associated with the price estimate are relatively smaller for our approach compared to the pricing errors for various alternative RSR that we study. This is especially an advantage in cities with fewer transactions or in thin markets. In other words, our proposed method produces relatively tighter prediction intervals for the price index. The greater precision of our price estimate arises in part because we difference away nuisance parameters without sacrificing sample size. We document this precision in performance using one private data set and one public data set; each contains sales from large metropolitan areas. In our data, the standard errors from our methods are almost two-thirds the standard errors for the repeat sales index.

**Appendix**

Using equations (3b) and (5), we write the error term for the RSR model as $u^{RSR} = \omega_{ijt}^R + \epsilon_{ijt} - \epsilon_{ijt}$ and the error for the NNP model as $u^{NNP} = \omega_{ijt}^U + \epsilon_{ijt} - \epsilon_{n(i,t)} + \mu_i - \mu_{n(i,t)}$. Define $T_{ijt}$ as the time between sales in the RSR. Let the $N_R$ pairs from the RSR be defined as the RSR pairs. Similarly, let the $N_U$ pairs we form using the nearest neighbor matching approach be defined as the PNN pairs. Assuming that $\omega_{ijt}^R$ follows a random walk, the variance of $u^{RSR}$ conditional on the time between sales is given by:

$$Var[u^{RSR}|T_{ijt}] = T_{ijt}\sigma_\omega^2 + 2\sigma_\epsilon^2,$$

(A1)
where \( \sigma^2_w \) is the variance in the per-period increments to \( \omega^i_j \), and \( \text{Var}[\epsilon_{ij}] = \sigma^2_\varepsilon \).

To calculate the unconditional variance, we decompose the unconditional variance into an expectation of the conditional variance and the variance in the conditional expectation as follows:

\[
\text{Var}[u^{\text{RSR}}] = E[\text{Var}[u^{\text{RSR}}|T_{ij}]] + \text{Var}[E[u^{\text{RSR}}|T_{ij}]]
\]

\[
= E[T_{ij}\sigma^2_w + 2\sigma^2_\varepsilon] = E[T]\sigma^2_w + 2\sigma^2_\varepsilon.
\]

(A2)

The first equality is an identity. The second equality follows from the fact that the error terms are mean independent of the time between sales, \( E[u^{\text{RSR}}|T_{ij}] = 0 \), \( \forall T_{ij} \), which implies \( \text{Var}[E[u^{\text{RSR}}|T_{ij}]] = 0 \). In the context of the RSR, mean independence of the error terms indicates that the change in price for a same-house sale is uncorrelated with the length of time between sales. The third equality defines the unconditional expected length of time between sales to be \( E[T] \).

The PNN method that we propose produces an error term with the following variance:

\[
\text{Var}[u^{\text{NNP}}] = \sigma^2_{\varepsilon,\mu} + 2\sigma^2_\varepsilon + 2\sigma^2_\mu.
\]

(A3)

Here, \( \sigma^2_\mu = \text{Var}[\mu_{ij}] \) measures the variance of the idiosyncratic time-invariant effect, and \( \sigma^2_{\varepsilon,\mu} = \text{Var}[\beta'(x_{ij} - x_{N(i),j})] \) measures the variance of the price difference due to differences in attributes for nearest neighbor properties.

The difference between the variance of the PNN pairs and that of the RSR pairs is given by:

\[
\text{Var}[u^{\text{RSR}}] - \text{Var}[u^{\text{NNP}}] = E[T]\sigma^2_w - \sigma^2_{\varepsilon,\mu} - 2\sigma^2_\mu.
\]

(A4)

If this value is positive, then the RSR pairs are noisier than the PNN pairs. As this value increases, the PNN pairs will provide a more precise signal relative to the RSR pairs. The value in (A4) is increasing in the average time between sales and the per-period variance in the change in price over time due to attribute changes. This implies the PNN pairs are noisier than the RSR when the time between sales for the RSR pairs is large. The value in (A4) is decreasing in both the variance in price due to attribute differences for nearest neighbors and the variance in price due to idiosyncratic features. Therefore, the PNN pairs will contain less noise when the nearest neighbors are more homogenous in nature.

The results in the preceding paragraph indicate that it is possible for the PNN pairs to have less noise than the RSR pairs. This will be true for areas where (1)
houses have similar bedrooms, bathrooms, square feet, etc. or (2) there are few idiosyncratic attributes that are property-specific, such as view, waterfront access, etc. The estimated values of the unconditional variances for the NNP and RSR pairs are presented in Exhibit A1.

### Exhibit A1

**Variance of the NNP and RSR Pairs**

<table>
<thead>
<tr>
<th></th>
<th>Seattle</th>
<th>Phoenix</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{Var}[u_{\text{NNP}}] )</td>
<td>0.3947</td>
<td>0.1161</td>
</tr>
<tr>
<td>( \text{Var}[u_{\text{RSR}}] )</td>
<td>0.2129</td>
<td>0.1434</td>
</tr>
</tbody>
</table>

The difference in the unconditional variances will manifest itself in the mean squared error of an OLS regression, similar to Bailey, Muth and Nourse (1963). If OLS is used in estimating the RSR, the mean squared error will converge to:

\[
s^2_{R,\text{OLS}} = \lim_{N_R - T + 1} \frac{1}{N_R - T + 1} \sum_{i=1}^{N_R} (\Delta p_{ijt}^R - \hat{\delta}_t + \hat{\delta}_r)^2
= E[T] \sigma_w^2 + 2 \sigma_\epsilon^2. \tag{A6}
\]

The mean squared error for an OLS regression that uses both the PNN pairs and RSR pairs will converge to:

\[
s^2 = \lim_{N_R + N_U - T + 1} \frac{1}{N_R + N_U - T + 1} \left\{ \sum_{i=1}^{N_R} (\Delta p_{ijt}^R - \hat{\delta}_t + \hat{\delta}_r)^2 + \sum_{l=1}^{N_U} (\Delta p_{ijt}^U - \hat{\delta}_t + \hat{\delta}_r)^2 \right\}
= \alpha^R[E[T] \sigma_w^2 + 2 \sigma_\epsilon^2] + \alpha^U[\sigma_w^2 + 2 \sigma_\mu^2 + 2 \sigma_\epsilon^2] = \alpha^R \text{Var}[u_{\text{RSR}}]
+ \alpha^U[\text{Var}[u_{\text{NNP}}] - \text{Var}[u_{\text{RSR}}]] \tag{A7}
\]

where \( \alpha^R = N_R / (N_R + N_U - T + 1) = 1 - \alpha^U \). We can combine this result with the formula for the standard errors for the OLS regression. Although we use a FGLS regression in our estimation procedure, the OLS example provides the intuition for the smaller standard errors when the PNN pairs are introduced into the regression. The OLS covariance matrices are given by:
equation (3b) and $X$ definition, we have Johnson (2012) to show that if $A$ given by:

$$\text{Cov}[\hat{\beta}_{\text{OLS}}^*] = s_{R,\text{OLS}}^2[X_R'X_R]^{-1}$$

$$\text{Cov}[\hat{\beta}_{\text{OLS}}^*] = s^2*[X^*X^*]^{-1} = s^2*[X_R'X_R + X_U'X_U]^{-1}, \quad (A8)$$

Where $X_R$ is the $N_R \times T - 1$ matrix of differenced time dummy variables from equation (3b) and $X^* = [X_R', X_U']'$ is the $N_R + N_U \times T - 1$ matrix that also includes the $N_U$ PNN pairs.

We can show that our proposed estimator will have smaller standard errors than the RSR if $\text{Cov}[\hat{\beta}_{R,\text{OLS}}^*] - \text{Cov}[\hat{\beta}_{\text{OLS}}^*]$ is a positive semi-definite (psd) matrix. A matrix $M$ is psd if $x'Mx \geq 0, \forall x$. We can use Corollary 7.7.4(a) in Horn and Johnson (2012) to show that if $A - B$ is psd, then $B^{-1} - A^{-1}$ is psd. Thus, we can also show that the OLS estimator will have standard errors if $\text{Cov}[\hat{\beta}_{R,\text{OLS}}^*] - \text{Cov}[\hat{\beta}_{\text{OLS}}^*]$ is psd. Applying this reasoning to (A8) shows that the OLS standard errors from our proposed methodology will be smaller if:

$$\frac{1}{s^2*}X^*X^* - \frac{1}{s_{R,\text{OLS}}^2}X_R'X_R = \frac{1}{s_{R,\text{OLS}}^2}X_U'X_U - \left[ \frac{1}{s_{R,\text{OLS}}^2} - \frac{1}{s^2*} \right]X_R'X_R \quad (A9)$$

is psd. A sufficient condition is that the PNN pairs have a smaller variance than the RSR pairs, which implies $s^2* < s_{R,\text{OLS}}^2 \Rightarrow 1/s_{R,\text{OLS}}^2 - 1/s^2* < 0$ and the second matrix on the right hand side of (A9) is psd. In this situation, where the PNN pairs are less noisy than the RSR pairs, both matrices on the right-hand side of (A9) will be psd and the OLS estimator using the nearest neighbor method that we propose will have smaller standard errors.

An alternative situation arises when $s^2* > s_{R,\text{OLS}}^2$ but $N_U$ is sufficiently large. For any $x \in R^N$ the jth element of the vector $X_Ux$ is given by $z_j = x_{t(j)} - x_{\tau(j)}$, where $t(j)$ is the time of the first sale and $\tau(j)$ is the time of the second sale. With this definition, we have $X_U'X_Ux = \Sigma_{j=1}^{N_U}(x_{t(j)} - x_{\tau(j)})^2 \geq 0$. For any given $K = -[1/s_{R,\text{OLS}}^2 - 1/s^2*]X_R'X_R$, we can choose $N_U$ sufficiently large so that $x'X_U'X_Ux \geq K = -[1/s_{R,\text{OLS}}^2 - 1/s^2*]X_R'X_R$. Thus, the difference in the OLS estimators will approach a psd matrix as $N_U$ grows large.

The results for the FGLS estimator that we propose follow a similar line of reasoning to the OLS case. The covariance matrices for the FGLS estimators are given by:

$$\text{Cov}[\hat{\beta}_{\text{FGLS}}] = s_{R,\text{FGLS}}^2[X_R'X_R]^{-1}$$

$$\text{Cov}[\hat{\beta}^*_{\text{FGLS}}] = s^2*[X^*X^*]^{-1} = s^2*[X_R'X_R + X_U'X_U]^{-1}, \quad (A10)$$

The results for the FGLS estimator that we propose follow a similar line of reasoning to the OLS case. The covariance matrices for the FGLS estimators are given by:
\[
\text{Cov}[\hat{\beta}_{FGLS}] = [X'_RW_R^{-1}X_R]^{-1}
\]
\[
\text{Cov}[\hat{\beta}^*_\text{FGLS}] = [X^*W^{*^{-1}}X^*]^{-1}
\]
\[
= [X'_RW^{*^{-1}}X_R + X'_UW^{*^{-1}}X_U]^{-1}, \quad (A10)
\]

where \( \hat{W}^* = \begin{bmatrix} W^*_R & 0 \\ 0 & W^*_U \end{bmatrix} \) is a diagonal matrix of weights determined by the variances in (A1) and (A2). \( \hat{W}_R \) is an \( N_R \times N_R \) diagonal weighting matrix with entry \( \hat{w}_{jj} = \hat{w}_j \). The value \( \hat{w}_j \) is the variance in the differenced sale prices for the \( j \)th repeat sale given by (A1). The matrix \( \hat{W}_R \) is a matrix of estimated variances and as such, it is a function of the data.

Applying the same reasoning to the matrices in (A10), we can show that \( \text{Cov}[\hat{\beta}_{FGLS}] - \text{Cov}[\hat{\beta}^*_\text{FGLS}] \) is positive semi-definite if:

\[
\begin{align*}
X^*W^{*^{-1}}X^* - X'_RW^{*^{-1}}X_R &= X'_R[W^{*^{-1}}X^* - W^{*^{-1}}]X_R + X'_UW^{*^{-1}}X_U \\
- X'_RW^{*^{-1}}X_R &= X'_R[W^{*^{-1}} - W^{*^{-1}}]X_R + X'_UW^{*^{-1}}X_U \quad (A11)
\end{align*}
\]

is positive semi-definite. The term \( X'_UW^{*^{-1}}X_U \) is a positive semi-definite matrix. The difference in the weighting matrices \( W^{*^{-1}} - W^{*^{-1}} \) is driven by the relationship between the estimated parameters of the random walk innovations using the squared residuals in the second-stage regression. If the time coefficients for the RSR and our proposed method have the same probability limits, then the residuals and resulting random walk parameters will also have the same probability limits.

Therefore, the difference \( W^{*^{-1}} - W^{*^{-1}} \) will be negligible, and the difference in the covariance matrices will be determined by the positive definite matrix, \( X'_UW^{*^{-1}}X_U \). Using the result from Horn and Johnson (2012), \( X^*W^{*^{-1}}X^* - X'_RW^{*^{-1}}X_R \) is positive semi-definite which implies that \( \text{Cov}[\hat{\beta}_{FGLS}] - \text{Cov}[\hat{\beta}^*_\text{FGLS}] = [X'_RW^{*^{-1}}X_R]^{-1} - [X^*W^{*^{-1}}X^*]^{-1} \) is positive semi-definite. Therefore, the standard errors of the estimator are smaller when using the proposed method.

**Endnotes**

1 The repeat sales regression (hereafter RSR) approach developed by Bailey, Muth, and Nourse (1963) and further refined by Case and Shiller (1989) estimates house price indices using repeat sales by regressing changes in same-house sale prices on dummy
variables corresponding to time periods. Unfortunately, sales without a previous same-house sale and the information they contain are discarded.

2 Meese and Wallace (1997) compare the performance of hedonic models to that of RSR models and note that drawbacks exist when observations are discarded using the RSR. Researchers have proposed alternative hybrid models that incorporate all transactions and find that these models reduce the standard errors corresponding to the price estimate. (e.g., Case and Quigley, 1991; Quigley, 1995; Hill, Knight, and Sirmans, 1997; Englund, Quigley, and Redfearn, 1998; Hwang and Quigley, 2004). Despite the advantages of these hybrid techniques, Bourassa, Cantoni, and Hoesli (2013) point out that the requisite hedonic variables are not always available to the researcher.

3 Lai, Vandell, Wang, and Welke (2008) propose a replication method that uses error correlations estimated on a large sample to improve the price estimation process wherein a smaller set of comparable properties having homogeneous attributes is used. More specifically, given a set of comparable properties, their method determines how to optimally weight attributes in order to minimize the mean squared error for the predicted value of the property. Implicit in this methodology is that the attributes of comparable properties can be used to price a target property. If the attributes of comparable properties include a spatial component, their method will place a greater weight on nearby properties.

4 Deng, McMillen, and Sing (2012), McMillen (2012), and Guo, Zheng, Geltner, and Liu (2014) are recent studies of a matching approach.

5 For a review of the nearest neighbors estimator and spatial errors, see Pace, Barry, and Sirmans (1998).

6 We also tried matching by cosine similarity as well as the un-normalized Euclidean distance, absolute value metric, max element metric, and taxicab metric; all metrics and the cosine similarity produced nearly identical results to results obtained when using the Euclidean distance. We chose the Euclidean distance because it had a significantly smaller run-time on the computer than the other methods.


References


*We would like to thank three anonymous referees and especially Ko Wang, the editor, for many useful comments. We also wish thank Charles Capone and participants at the 2010 AREUEA mid-year meeting for their helpful comments. We also acknowledge the Center for Real Estate Theory and Practice at Arizona State University for making the Ion Data Express database available to us.*