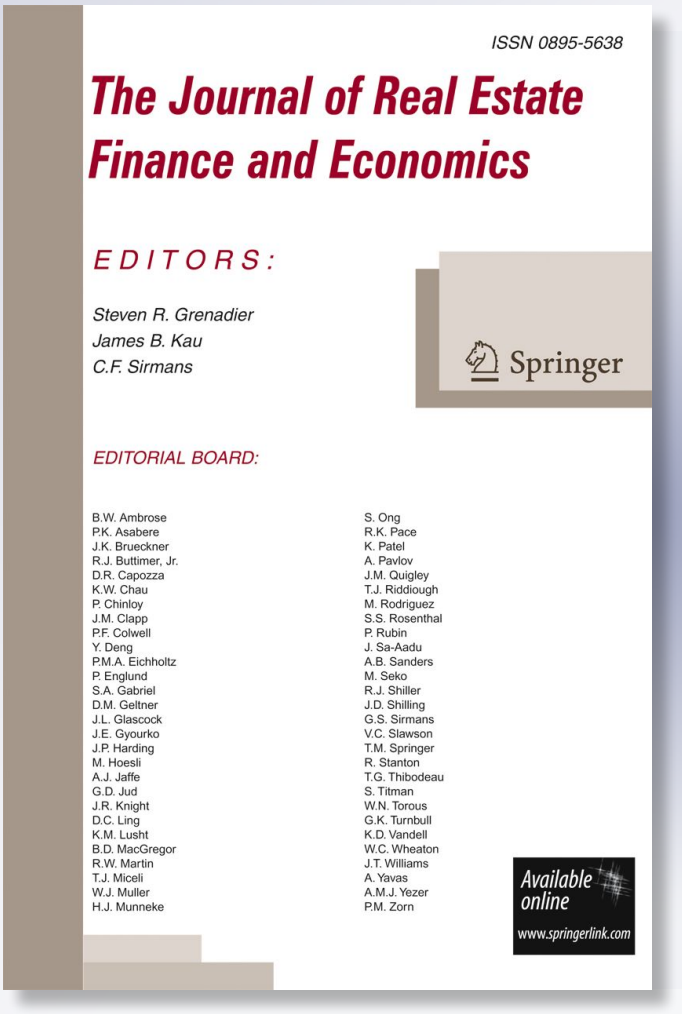
Inconsistency in Welfare Inferences from Distance Variables in Hedonic Regressions

Justin M. Ross, Michael C. Farmer & Clifford A. Lipscomb

The Journal of Real Estate Finance and Economics

ISSN 0895-5638
Volume 43
Number 3

J Real Estate Finan Econ
(2011) 43:385-400
DOI 10.1007/s11146-009-9221-z



Springer

J Real Estate Finan Econ (2011) 43:385–400
DOI 10.1007/s11146-009-9221-z

Inconsistency in Welfare Inferences from Distance Variables in Hedonic Regressions

Justin M. Ross · Michael C. Farmer · Clifford A. Lipscomb

Published online: 3 December 2009

Abstract In hedonic analysis, a common approach for eliciting information regarding the welfare significance of some landmark or (dis)amenity is to control for its distance from each observation. Unfortunately, the effects of distances to amenities on housing prices are generally not consistent indicators of the true price impact of that amenity. Instead these variables serve as proxies for the relative position of every observation in space. Whenever a household considers more than two landmarks in a housing purchase, distance variable parameter estimates are simply the best linear fitted weights for that multiple criteria location decision. Simulations illustrate extreme sensitivity in parameter estimates to the researcher's choice of landmarks. One strategy models the location of each observation directly instead of its distances to amenities. Using the quadratic controls of longitude and latitude controls for location effects on price to assure unbiased estimates of non-distance variable regressors.

Keywords Hedonic · Distance · Sensitivity analysis

JEL codes R0 · Q5 · C8

J. M. Ross (✉)
School of Public & Environmental Affairs, Indiana University, Bloomington, IN 47405, USA
e-mail: justross@indiana.edu

M. C. Farmer
Department of Agricultural and Applied Economics, Texas Tech University,
Agricultural Sciences Building—Box 42132, Lubbock, TX 79409, USA
e-mail: michael.farmer@ttu.edu

C. A. Lipscomb
Department of Marketing and Economics, Langdale College of Business Administration,
Valdosta State University, Valdosta, GA 31698-0075, USA
e-mail: calipscomb@valdosta.edu

Introduction

The advent of powerful GIS tools has enabled the increased application of distances variables in hedonic price analysis. Researchers commonly assign welfare significance to the housing price response of the distance between housing units and a landmark believed to be an amenity or disamenity. A classic example is Harrison and Rubinfeld (1978), who included a weighted distance to employment centers to examine the demand for environmental quality. More recently, Noonan et al. (2007) use the distance to the historic city center to explain variation in median home value across the United States. Cameron (2006), using polar coordinates rather than Cartesian coordinates, control for directional heterogeneity in hedonic property value models to improve the efficiency of distance variables. Brasington and Hite (2005) use the distance to an environmental hazard to estimate the value of clean-up. Ihlanfeldt and Taylor (2004) consider the distance from a Superfund site in Atlanta to value site clean-up.

However, equally efficient statistical specifications can be achieved using very different landmarks as independent variables. If very different hedonic price specifications contain similar statistical information, researchers may draw conclusions about the *meaning* of specific coefficient estimates on distances to various locations that may be unfounded. Several authors treat this as a specification problem; or they seek to solve the problem using the statistical information from very different specifications of the location effect on housing prices. Deaton and Hoehn (2004) add a second distance variable—the distance to the Central Business District—as a control to estimate the effect of the distance to an amenity as a mechanism to efficiently assess its value; and Fik et al. (2003) compare functional forms of the latitude and longitude coordinates of properties in Tucson to estimate hedonic prices. These works expose a robustness problem in deriving welfare implications from 'distance variables' in hedonic price analysis. Here we show the problem suffers from relatively irresolvable identification limitations that statistical specification tests and corrections alone generally cannot detect.

To demonstrate this point, we conduct a set of simulations to assess the extent of parameter sensitivity to approximated distance variables. Critically, these simulations use the popular longitude and latitude distances from an assumed key amenity or landmark (e.g. Fik et al. 2003). Our first simulations illustrate the fundamental complication of distance variables, which is that a distance to any hypothesized landmark in space carries a linear relationship with any other truly important landmark(s). The implication is that if the researcher wishes to determine welfare significance, they must be very accurate to properly identify the specific targeted amenities valued by the household, a modeling accuracy that exceeds the demands of other correlated independent variables. Education, income, and family size for instance are measured on different dimensions; and a regression can be fitted over that space. For distance variables, which are measured as some form of longitude and latitude (applicable to linear distances or driving distances), the difficulty is increased to isolate *the* effect of a single landmark among many that affect home price in hedonic regression models.

A wide diversity of hedonic specifications of housing price can show similarly high levels of significance and explanatory power to explain home price variation.

The first simulation shows that distance variables of interest to key landmarks often change coefficient values considerably and maintain high goodness-of-fit statistics even with slight misidentifications. Also, models that specify an incorrect number of landmarks that affect home price or specify incorrect landmarks of true interest to the household can show the highest levels of coefficient significance and the strongest explanatory power if the true model includes multiple landmarks of interest. This finding places a strong burden on the researcher to identify the correct landmark or amenity of interest and to flawlessly identify all other landmarks of interest to avoid spurious results.

The second set of simulations is a Monte Carlo experiment that not only confirms the findings of the first simulation, but also demonstrates that these unstable distance variable coefficients introduce inconsistency into the independent variable coefficients as well. Nevertheless, the simulations also reveal that, while we cannot learn anything definitive about an individual distance variable, a collection of distance variables can be used to generate very reliable estimates of the optimal location in space. More specifically, the use of two distance variables to *any* location does a good job of triangulating the observations' optimal position in space and of generating both reliable predicted values of the dependent variable and stable estimates of the correlation coefficients of the independent random variables. This last finding motivates the third simulation.

The first two simulations demonstrate that little, if anything, of welfare significance can be inferred from distance variables; yet these variables do predict the optimal location in space and therefore can serve partially to remedy the statistical problem of omitted variable bias. A regression model that formally addresses optimal locations performs considerably better. We demonstrate that a regression with a quadratic linear specification of the latitude and longitudinal positions in space performs as well as a model in which the researcher correctly identifies the true influential landmarks. This has the added benefit that it leads to more consistent estimates of other regressors correlated to location by isolating a robust overall (or bundled) location effect even as many locations that influence housing prices remain latent.

The rest of the paper follows. Next we describe the background of hedonic price theory and the nature of distance variables. Then, section "Illustrative Simulation" provides a simple simulation that illustrates the problems and challenges of using distance variables in a hedonic regression. Sections "Monte Carlo Simulation I" and "Monte Carlo Simulation II—Finding the Optimal Location in Space" then conduct the Monte Carlo experiments described above. Finally, section "Conclusion" offers the conclusions of our work.

Background

In hedonic price analysis a household pays a premium to live in a home due to its proximity to (or distance from) several distinct amenities (disamenities). By this act, they reveal a preference for a single location, holding fixed all other home and neighborhood attributes. In the true hedonic price function, then, the value of proximity to an optimally best point (i.e. the highest valued point) is implicitly

generated, typically, as some weighted effect among distances to multiple landmarks of interest. So, the actual revealed price effect on housing is a bundled good that is indexed by the weightings of distances to multiple amenities.

This work shows that the value of a given housing location vis-à-vis its optimal location may be estimated with a high degree of precision even if the true landmarks of interest to the household are not included in the regression specification. However, the robustness of the ability to estimate an overall location effect fails to allow the researcher to reliably decompose this indexed effect. When a given landmark that is considered by a household in a home purchase is included in a hedonic analysis, the coefficient value of living close to, or far from, that landmark often changes dramatically as other landmarks are added or removed from the hedonic specification. While this is a general result of regression analyses, we show this is particularly acute for distance variables. An emerging stylized fact about hedonic analysis is the propensity for individual coefficient values to be unstable across studies even though all results show high levels of coefficient significance, very high explanatory power and low p-values to indicate low multi-collinearity (Zietz et al. 2008).

The problem of coefficient instability among distance variables has been fashioned as a specification problem. Deaton and Hoehn (2004) demonstrate considerable efficiency gains from the addition of a second landmark (the Central Business District) into a hedonic regression equation as a critique of studies that had included only the distance to an environmental amenity. A more general diagnostic that tests successfully for spatial misspecification broadly within a given regression has been developed by Pace and LeSage (2008) in the form of an adapted Hausman test. These diagnostic reveals if the overall location effect on housing price has been well specified; but as we show, in themselves the diagnostics do not reveal whether the true landmarks have been identified or if the coefficients estimated correspond to a true value gradient. Palmquist (1984, 2004) also discusses the problem of distance variables as a mix of specification and identification problems. Preferring zonal or neighborhood effects to distance variables, he defines discrete indicators such as the presence or absence of a park or of a neighborhood nuisance, perhaps within a certain discrete distance or on one side of a barrier. The identification of the welfare effects of that landmark is impaired by unmodeled diversity of submarkets; and, in these works, specification problems arise from general collinearity among variables, much as Deaton and Hoehn (2004) considered.

Since distance variables are measured on identical dimensions (as multiple vectors) defined on the same two-dimensioned planar space, the correlation between two random locations will share a correlation that, say, income and education do not. Each location chosen for a distance variable will have an effect on the coefficient estimates of other distance variables, even if the landmarks themselves have no underlying behavioral connection to the home-owner choice. The statistical difficulties in choosing a particular landmark can be understood as a case of implementing proxy variables.

If we allow d to serve as a vector of distance measurements in n-space that proxy for the true measurements, D, then it is implicitly assumed that $E(y|x,D,d)=E(y|x,D)$. This simply states that once D is included in the estimation, a proper proxy variable d becomes either irrelevant or redundant in explaining variation in y. If the estimated

equation for E(y|x,D) takes an additive functional form, as in regression frameworks, a weaker condition set requires a) that d be irrelevant or redundant and b) that the proxy variable be uncorrelated with the error term (Wooldridge 2002). The problem is that the proxies embed mathematical relationships to the true measure. In the case of distance variables, the conditions are generally violated as they all are multiple lines measured along the same two dimensions.

For a given housing unit, D is a vector composed of a longitude (k^* $long$) and latitude (m^*lat) and d is a vector composed of a longitude (j^*long) and latitude (n^*lat). Regressing house prices (HP) on these two distance variables, we obtain a regression line that will be regressed through a mean point in HP, $long$, lat space: $HP=K+a_1[(k^*\ long)^2+(m^*lat^2)]^{1/2}+a_2[(j^*\ long)^2+(n^*lat^2)]^{1/2}$. In these conditions orthogonality cannot be assured: a_1 will not be independent of j and n; and a_2 will not be independent of k and m. So unless each element of d is at a right angle to each element of D (perfect orthogonality) for the preponderance of the observations, d continues to hold some relevant influence on the regression estimates of the true landmarks. This is much more pervasive than other related variable effects. In practice, this makes it very difficult to isolate the true effect, D. If a nearby location is chosen as the landmark (the proxy) instead of the true landmark, the difference $D-D(hat)=v$ from the equation $D=\theta_1+\theta_2 d+v$ will be systematically related to any other location, making the estimate on D in E(y|x,D,d) inconsistent. In two sets of simulations we demonstrate this effect; in one of them we also detail an inconsistency problem for other non-distance variable effects, which is potentially more correctable.

This resolves a conundrum in many applied works. Though seldom reported in the reviewed literature, published works using now routinely large data sets tend to report only two distance variables. Those two variables show almost no multi-collinearity. Yet, informally, researchers note commonly the inability to fit a third location variable to due to overwhelming multicollinearity, so much so it obviates convergence around a nearly singular matrix.

Two points in space, measured as weighted vectors to the optimal location (the location that would realize the highest home value) will triangulate that optimal position by fundamental geometry. There is no visible multicollinearity though because the regression coefficients—weights in the triangulation—are fitted to locate that 'best' position. Add a third, and no unique weights that are not multiplicative combinations of each other can be found (especially in large data sets). The weights themselves then are highly conditioned on the location of the other landmark.

The regression then fits a vector weighted estimate among these two points of the optimal point by the best triangulation of that point given the data and the specified landmarks. Models that pass specification tests and perform well generate an embedded prediction of the optimal point, or best location. That is what the regression estimates. Yet these tests are not enough by themselves to resolve the underlying problem of isolating a specific landmark's influence on home price.

The linear mathematical relationships above suggest that the true effect of a single landmark of interest may be captured efficiently by a hedonic specification that includes two, three, or more location regressors; but it may be difficult to partition that location effect into component parts to identify the independent effect of the landmark. Conversely, a true multi-landmark location effect may be captured by only

one centrally weighted single point that, it itself, is not valued by the household. We might expect these instabilities to be particularly acute in a target rich environment such as an urban area.

Illustrative Simulation

A simple illustration illuminates the linear relationships between distance variables and the corresponding trouble for regression analysis that is described in the previous section. Figure 1 is a Cartesian grid of 25 points, with D as the origin (0,0). Each of those 25 points labeled with capital letters represents an observed housing unit in space. Additionally, three points on the line segment between B and D are added at (−.5, .5), (−.75, .75), and (−.25, .25) as points a, b, and c respectively. A house price (HP) for each point is calculated using a linear equation of the distances from three landmarks: points A, B, and C on Fig. 1. To isolate the effects of multiple distance variables on each other in hedonic estimation, no other attributes are considered in this initial demonstration, nor is there a random error component in the true model. So, the true data generating process for the hedonic model is

$$HP_i = 14 + \delta_1 dist\_A_i + \delta_2 dist\_B_i + \delta_3 dist\_C_i$$

For simplicity of the illustration, $\delta_1=\delta_2=\delta_3=1$; so the value of a home located at point B is 10 (14–2–0–2). This simple generating process allows us to explicitly demonstrate the interdependent nature of the distance variables. Housing prices across the grid range from a low of 5.18, at points H and I, to 10.14 at an unlabeled position (−0.59175, 0.59175), which lies on the line between points a and b. This

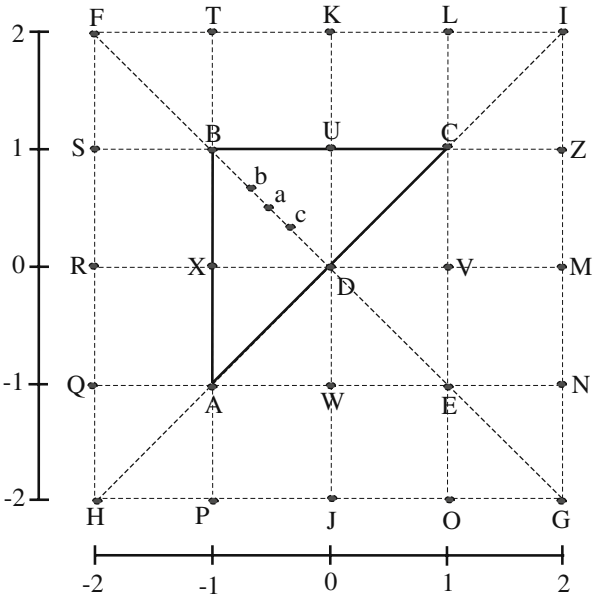


Fig. 1 Cartesian grid of optimal house price over landmarks

location is the optimal HP location, calculated as the kernel of triangle ABC (Schmeidler 1969), a method familiar in game theory.

A few representative hedonic specifications are presented in Table 1. In these models, the only control variables are the distances to select "hypothesized locations." This data generating process for housing price uses only an intercept and three distance variables. Therefore, any regression error is the result of misidentifying the true landmarks. So the first simulations isolate the impact of including or excluding distance variables on the capacity to generate consistent welfare inferences. In the subsequent section we expand this to include the effects on other hedonic attributes. Secondly, a regression model that specifies three distance variables often in practice has serious problems with near singular matrices, a problem discussed in the subsequent exercise. This limitation flows from the impacts of landmark choice inclusion/exclusion to place a heavy burden on the analyst to infer welfare impacts from distance variables. For this reason, Table 1 includes only regressions that use fewer than three distance variables.

The first model regresses home price against the distance to points U and X. Both coefficients equal −1.106 and are statistically significant at the 99% confidence level. Despite incorrectly identifying the landmarks that affect home owners in the regression, the model carries an R^2 of 0.956. No landmark of true impact is included in the hedonic model.

The researcher, however, might accurately observe several important landmarks; so the next model in Table 1 includes two landmarks, A and B, valued by households. Unfortunately, the adjusted R^2 is far lower at 0.819 than the model using U and X; and while the true value of both coefficients is −1, with absence of landmark C generates coefficient estimates of −0.460 and −1.373 for distance from points A and B, respectively. Therefore, even if strong theory leads the analyst to include points A and B, the absence of point C in the model has a profound effect on the remaining estimated coefficients. Though both landmarks exercise an equal effect on home values, one landmark is valued at 300% larger than the other. Another model with distances to points A and C, again both true landmarks, has the lowest R^2 statistic (0.742) among the regressions in Table 1. The regression coefficients on both distance variables are −1.531, instead of −1.0; or 50% higher

Table 1 Selected regression results for illustrative simulation

Hypothesized locations	Intercept	δ_1	δ_1	R^2	Implied optimal location
U, X	11.857***	−1.106***	−1.106***	0.957	(−0.5, 0.5)
	(0.182)	(0.093)	(0.093)		
A, B	11.457***	−0.460***	−1.373***	0.819	(−1, 1)
	(0.424)	(0.165)	(0.147)		
A,C	14.262***	−1.531***	−1.531***	0.742	(0, 0)
	(0.779)	(0.210)	(0.210)		
b	11.016***	−1.901***		0.930	(−0.5, 0.5)
	(0.195)	(0.099)			

Sample size is 28. Standard errors are reported in parentheses

than their actual influence on home price. Notably, the value for point A shifts from −0.460 to −1.531, an increase of 333%. Moreover, if U and X are plausible landmarks (e.g. nearest retail center; school), the analyst might easily reject a model with two correctly identified landmarks and attribute welfare meaning to spurious significance of the estimated distances to points U and X.

Several single point models using a landmark not valued by the household also outperform both models that include two *true* landmarks. The midpoint between B and D, (−0.5, 0.5), yields an adjusted R^2 of 0.93, a coefficient of −1.901 and a t-statistic equal to 19.258. Other specifications using a single distance variable near the true optimum point (−0.59, 0.59) return similar results.

Single point models underscore the difference between estimating an aggregate location effect versus estimating the underlying value of an individual landmark. In the model using points A and B, B arises as the estimated optimum; and it is near the true optimum. In the two models symmetric around the segment FG, the optimal point can lie along the segment U to X or A to C. Yet the highest valued point on these segments is the midpoint (−0.5, 0.5) and point B, respectively, which are identified as the estimated optimum in a two point model using a more common distance decay model, such as the natural log of distance to those same landmarks. The key here is not the decay of the distance effect, which we might generally expect, but the non-linearity of the distance variable as a means to predict the optimum position. Even in fairly naïve models, the two landmark hedonic models in these simulations tend to reasonably approximate the optimal location, at least as it compares to the inconsistent and at times spurious information about the value of home proximity to a specific landmark. Figure 2 helps us to understand why.

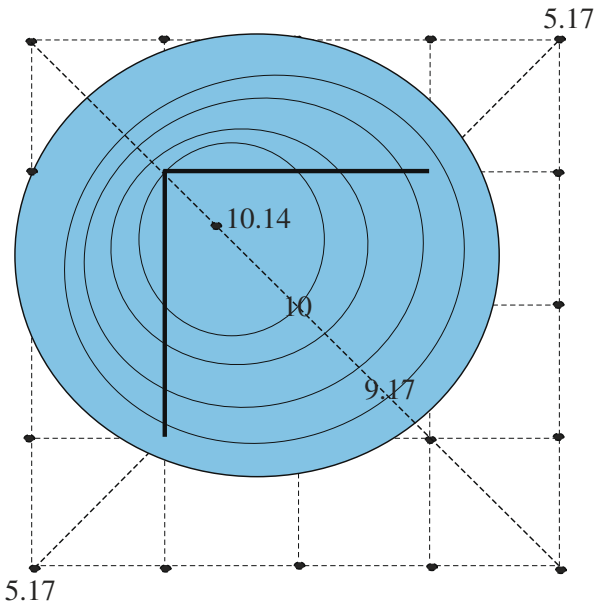


Fig. 2 Housing price gradient for illustrative simulation

Figure 2 is an iso-value graph for the grid presented in Fig. 1. The optimum lies, accounting for rounding error, at (−0.59, 0.59). In this case, positions of equal value away from (−0.59, 0.59) are not concentric around the optimal point. Iso-value positions for home prices of 10 and 9.17 on Fig. 2 illustrate the asymmetry of, in this case, the elliptical shape of iso-value positions along this value gradient. With more than three landmarks affecting home price there is no guarantee of an optimum, but a typological property of the kernel in a metric space graphing of a closed value set does assure a local optimum (Kasriel 1971). In the rest of the work, we take advantage of the ability for any two landmarks to triangulate an optimal position from a non-linear weighting of longitude and latitude from two points; that is, for example, to locate the point (−0.59, 0.59) from an estimate of the non-linear effect of value changes away from points L and E. Below we identify an optimal position using quadratic longitude and latitude distance effects.

These illustrative simulations set the stage for a full Monte Carlo simulation with a larger set of points and a more traditional data generating process for housing prices. It again shows that multiple specifications of an underlying true model perform well regardless of their direct relation to household preferences. Moreover, simulations show that these several different specifications of distances to or from amenities pass a spatial Hausman test, supporting the argument that the concern raised here centers on model identification, not specification. The overall location effect can be validly identified but the values of specific landmarks are not, in general, reliably estimated. While the coefficients on true landmarks vary considerably and landmarks not valued by the household routinely show spurious significance, the optimal point is approximated with a high level of accuracy and consistency.

Monte Carlo Simulation I

In each repetition of our Monte Carlo simulation, a new dataset is randomly generated, as well as a random set of "hypothesized" landmarks, and regressions are estimated over alternative specifications of our data set. In all cases, the generated housing price is a function of an intercept, a single random variable $x_i \sim N(0,1)$, and a Gaussian error term, $\varepsilon_i \sim N(0,2)$. Each observation will be assigned a latitude-longitudinal position in space randomly generated from a uniform distribution within the neighborhood depicted as the shaded region visualized in Fig. 2. Also demonstrated in Fig. 3, are two locations at A: (0,15) and B: (10,15) whose geographic proximity will be influential in the true data generating process, but are unknown to the researcher.[1] The distance to these locations will be designated as D^A and D^B, and two sets of housing prices are generated according to Eqs. 1 and 2:

$$HP_i^1 = \beta_0 + \beta_1 x_i + \gamma_1 D_i^A + \varepsilon_i \tag{1}$$

$$HP_i^2 = \beta_0 + \beta_1 x_i + \gamma_1 D_i^A + \gamma_2 D_i^B + \varepsilon_i \tag{2}$$

[1] Whether or not this point is within or outside the neighborhood is irrelevant for demonstrating the instability that may arise from picking an incorrect representative point.

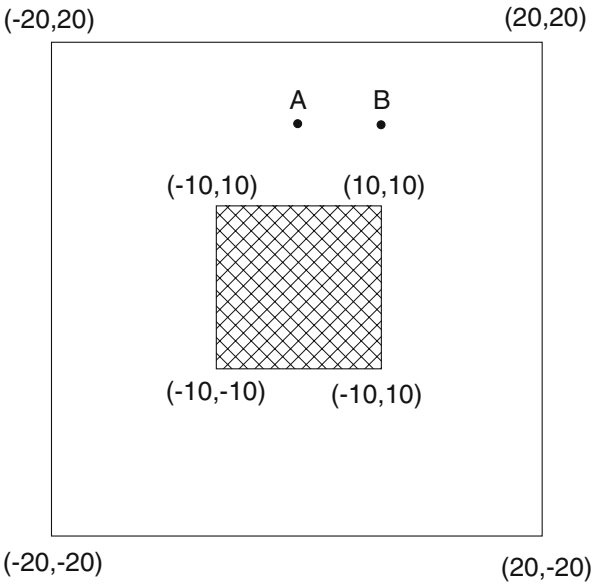


Fig. 3 The sampling space for the Monte Carlo data

Notice that Eq. 1 is a special case of Eq. 2 where $\gamma_2=0$. Otherwise, the parameters will be set to $\beta_0=1$, $\beta_1=2$, $\gamma_1=-0.25$, and $\gamma_2=-0.10$, and there will be 400 observations in each trial. Note that the distance coefficients are negative, so the implication is that locations A and B are treated as amenities. Since the distance to A and B is unknown to the researcher, in each trial two random points in the white space of Fig. 3 will be drawn to calculate a variable to represent the researcher's hypothesized or proxy distance variables, which will be designated d^A and d^B. Then, they will be used to estimate the following regression models:

$$HP_i^1 = \beta_0 + \beta_1 x_i + \gamma_1 d_i^A + e_i \tag{3}$$

$$HP_i^2 = \beta_0 + \beta_1 x_i + \gamma_1 d_i^A + \gamma_2 d_i^B + e_i \tag{4}$$

For each data generating process (DGP) in Eqs. 1 and 2, both regression model specifications in Eqs. 3 and 4 are also estimated; meaning four (2^2) regressions are estimated in each of the 1,000 trials of the simulation. Table 2 provides summary statistics for each regression, including the mean and standard deviation of the correlation coefficients, the average adjusted R^2, the sum of squared errors (σ^2), and the percentage of the trials where the specification failed the spatial Hausman Test at the 95% confidence level.

For each trial, the γ coefficients fluctuate considerably both in magnitude and sign, a finding which also can be inferred from their means and standard deviations in Table 2. For example, for the hedonic model with two true distance variables, Eq. 2, 95% of the Monte Carlo estimates fit between 1.48 and −1.50, with the mean

Table 2 Means and standard deviations of Monte Carlo I results

	HP^1	HP^1	HP^2	HP^2
Intercept	−54.27	−69.78	−86.16	−108.85
	(88.76)	(59.90)	(126.55)	(85.83)
x	2.06	2.20	2.10	2.27
	(0.71)	(1.26)	(1.01)	(1.78)
d^A	−0.04	0.00	−0.06	0.00
	(0.44)	(0.17)	(0.77)	(0.24)
d^B	0.00		0.01	
	(0.44)		(0.76)	
σ^2	303.11	928.27	617.69	1,853.91
Adj-R^2	0.85	0.54	0.85	0.54
% reject Spatial H	7.00	4.00	7.50	3.50

Simulation based on 1,000 trials on sample sizes of 400. Reported above are means, with standard deviations in parentheses. HP^1 has only one true distance variable in data generating process, while HP^2 has two

estimate of −0.06 that can be seen in the third column of Table 2. In most of these trials, the correlation coefficients are statistically significant, and they frequently produce very high model fit statistics. These observations are also consistent with the findings of the previous section that examined only one simulated data set.

The identification of the landmarks used in the distance variable calculation clearly had implications for the other regressors in the model. Of course the intercept varies wildly, but this is rarely of concern to researchers. More concerning are the consequences of identification of the coefficient on *x*. The most consistent estimate of β_1 is when there was one true distance variable, but two distance variable controls. In that case, β_1 had a mean of 2.06, and 95% of the estimates were between 3.48 and 0.64. The least consistent estimate of β_1 had 95% of the estimates between 5.83 and −1.29, which was in the case of two true distance variable regression that controlled for only one distance variable. The point to take away here is that incorrectly identifying the landmark has consequences for producing unbiased estimates of the non-distance variables in the regression.

Another observation is that, regardless of whether or not the true DGP is Eq. 1 or 2, using two distance variables as in Eq. 4 provides a more consistent set of estimates than using a single distance variable. The difference between the mean $\widehat{\beta}_1$ of the trials and its true value of two is statistically insignificant in both cases when the two distance variables are employed, and statistically significant when including only one distance variable. As discussed earlier, using any two distance variables essentially identifies an observation's relative location in space. This can be seen visually by plotting the predicted values over the latitude-longitude positions, as we did for one of the trials in Fig. 4.[2]

[2] This is the predicted value, holding the non-distance variables constant at the sample-mean value.

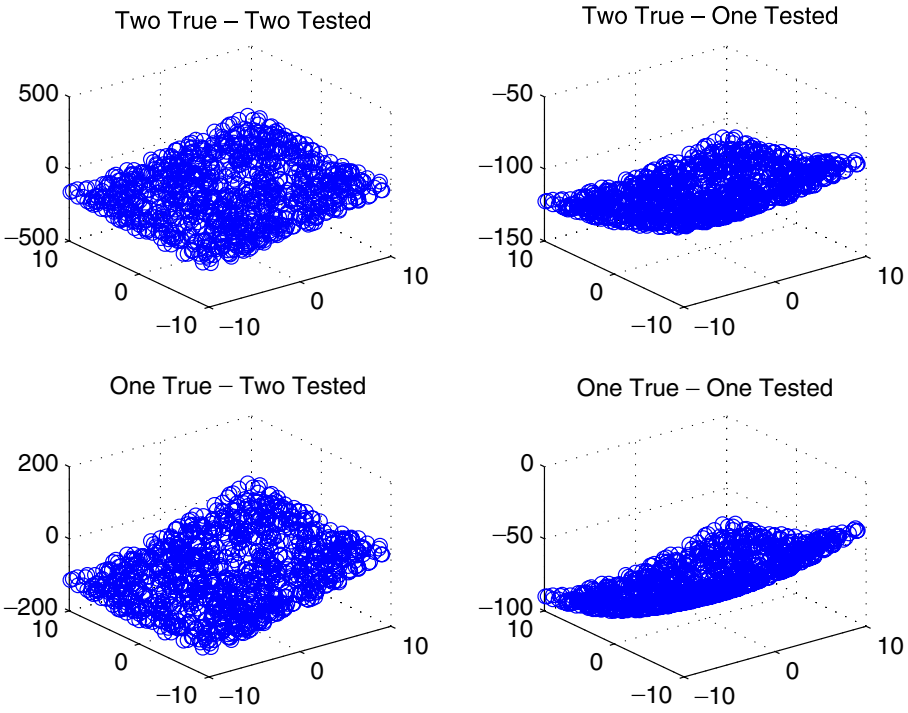


Fig. 4 Predicted values from a representative trial of Monte Carlo experiment

The graphs in Fig. 4 plot the predicted values on the z-axis, and the observation's latitude-longitude coordinate on the x-y axes. When two distance variables are tested, the predicted values increase as you move closer to the area with the amenities, locations A and B at (0, 15) and (10, 15) respectively. This is true regardless of whether or not the true DGP has one distance variable or two. When only one distance variable is tested [e.g. the regression model in Eq. 3] the model is less likely to put those greater predicted values in the correct quadrant. Indeed, in Fig. 4 the one distance variable model puts the highest predicted values on the correct side of town but winds up getting the wrong corner.

These findings support our claim that we can use distance variables only to provide information regarding optimal locations in space. Virtually *any* two distance variables will accomplish this without giving the researcher insight into what the causal factor may be driving this determination. Therefore, trying to draw welfare significance upon any distance variable without very strong theoretical foundations could provide misleading conclusions. Furthermore, on average using two distance variables seems to clear the omitted variable bias in $\widehat{\beta}_1$ that would result from choosing the incorrect landmark more than when choosing only one distance variable.

This ability of a two distance variable model to better find the optimal location in space is undoubtedly the reason it had both a higher adjusted R^2 and lower sum of

squared errors throughout the simulation.[3] The two distance variable specification also seems to be more likely to fail the spatial Hausman test in the single variable specifications. The spatial Hausman test carries the null hypothesis that the difference between the correlation coefficients of OLS and spatial error model is zero, and rejection of this is taken as evidence of model misspecification (Pace and LeSage 2008). In all trials, the regressions used the incorrect landmarks, so it seems to be the case that the real estate researcher is slightly more likely to fail this specification test. Nevertheless, in all cases the frequency of null hypothesis rejections is close to what asymptotic theory would predict the frequency of Type I errors to be at the 95% confidence level, so this is not much of an advantage.

Monte Carlo Simulation II—Finding the Optimal Location in Space

If welfare significance of distance variables cannot be inferred, then their salvaging point is for their statistical properties, namely clearing omitted variable bias in the non-distance coefficients and predicting optimal locations in space. In this case, a model that formally attempts to accomplish this, as opposed to justifying its inclusion as a proxy variable, might have better statistical properties. In the next Monte Carlo simulation, Eqs. 1 and 2 are once again estimated with the same parameter values as in the previous section; the difference this time is that locations serving as the true landmark(s) are allowed to vary with each trial. In particular, these true locations are randomly generated and can be in or outside the neighborhood. This dataset is generated according to the boundaries depicted in Fig. 3. In each trial, a regression is estimated for both HP^1 and HP^2 over the following model specification:

$$HP_i = \beta_0 + \beta_1 x_i + \gamma_1 lat_i + \gamma_2 lat_i^2 + \gamma_3 long_i + \gamma_4 long_i^2 + e_i, \tag{5}$$

where *lat* and *long* represent the latitude and longitude coordinate of each observation, respectively.

As established in the "Illustrative Simulation", an angular relationship between any two lines suggests that the correlation coefficient of any distance variable is a partial function of the underlying landmark of the second distance variable. Furthermore, in the previous Monte Carlo simulation we demonstrated that the misidentification of the important landmark(s) could result in bias in the other correlation coefficients. In particular, using a single distance variable resulted in greater bias and instability of the correlation coefficients, while the inclusion of a second distance variable considerably reduced these problems. Essentially, using two distance variables implicitly identifies each observation's location in space, but they have a tendency to interfere with each other because of their linear relationship. In contrast, the latitude-longitude quadratic specification in Eq. 5 explicitly identifies

[3] Since hedonic regressions are sometimes estimated with multiple procedure types in addition to OLS, the R^2 statistic is often not a good choice for comparison. If, for instance, the model is estimated using a Maximum Likelihood method, most standard statistics packages report a pseudo-R^2, which often leaves σ^2 or the Root Mean Square Error (RMSE) as the better measure for comparative purposes.

the observations' relative position in space without the collinearity problem of distance variables. Also, while distance variables restrict the regression to defining the optimal location to be on a line between the two landmarks, including the quadratic terms of the latitude-longitude coordinates precludes this problem, as one can think of the regression as now determining the optimal x and y directions, respectively.

Using 400 observations and 1,000 trials once again, Table 3 presents the results from Eq. 5 for both the single and dual landmark hedonic DGP from Eqs. 1 and 2. In addition, Table 3 presents the estimates that would be generated if the researcher correctly identified the landmark(s) with D^A and D^B. According to the mean and standard deviations of the parameter estimates in Table 3, the model using the latitude-longitude specification produced almost identical estimates for the coefficient on *x* and model fit statistics. Therefore, researchers can learn as much as they want about optimal location in space and the effect of changes in *x* with perfect information about true landmarks as they could with no information about important landmarks, provided they use the latitude-longitude quadratic specification.

Comparing the latitude-longitude specifications in Table 3 to the distance variable approach in Table 2, it can be seen that the model fit statistics are considerably better in Table 3, with the adjusted R^2 almost at 1 for both HP^1 and HP^2. Furthermore, the mean estimate of β_1 is much closer to its true value of 2 and has a standard deviation that is considerably lower, suggesting the Eq. 5 specification provides the most

Table 3 Means and standard deviations of Monte Carlo II results

	HP^1	HP^1	HP^2	HP^2
Intercept	0.995	−66.644	1.010	−93.227
	(0.200)	(40.579)	(0.409)	(44.168)
x	1.994	1.994	1.994	1.994
	(0.101)	(0.102)	(0.101)	(0.102)
Distance to true landmark 1	−0.250		−0.250	
	(0.001)		(0.001)	
Distance to true landmark 2			−0.100	
			(0.002)	
Latitude		−0.155		−0.201
		(5.760)		(6.298)
$Latitude^2$		−0.250		−0.350
		(0.003)		(0.003)
Longitude		−0.146		−0.146
		(5.872)		(6.315)
$Longitude^2$		−0.250		−0.350
		(0.003)		(0.003)
σ^2	4.008	4.008	4.008	4.008
Adj-R^2	0.997	0.997	0.997	0.997

Based on 1,000 trials of 400 simulated observations. Reported above are means, with standard deviations in parentheses. HP^1 has only one true distance variable in data generating process, while HP^2 has two

reliable estimates for the x regressor. Like the cases using distance variables, the latitude and longitude coefficients are only helpful in gathering information about the optimal location in space. However, the absence of a collinearity problem that accompanies distance variables allows for much more favorable statistical features.

Conclusion

Hedonic studies frequently employ a variable capturing proximity to some potentially important amenity or location. For example, distances to the CBD, to an amenity, or to an environmental hazard are all common in the urban, real estate, and environmental economics literatures. Unfortunately, this practice of using "distance to" variables is difficult for applied researchers and can create misleading results. The likelihood of potentially biased coefficients and inflated standard errors from employing these variables requires careful investigation from researchers.

To demonstrate the problem, we use Monte Carlo simulations to show that choosing the incorrect location for the landmark can yield entirely wrong coefficient estimates for both the distance variable and the other regressors. Researchers are recommended 1) to refrain from drawing welfare inferences from their distance variables unless they have very strong theoretical reasons for doing so and 2) to control for individual latitude-longitudinal coordinates with a linear quadratic specification. This not only clears bias out of the other covariate parameter estimates, but it provides more reliable information regarding the optimal location in space.

The reason the overall (or bundled and indexed) location effect can be estimated but the individual (or element weightings) effects cannot is that these effects expose an identification problem, not a specification problem. Decomposing a bundled location effect into reliable individual component weights requires that the analyst identify the exact number of locations that affect household purchase decisions, and then identify the precise location of each of the effective landmarks. The overall location effect can be captured quite generally with distance variables as in these simulations, but the individual coefficients are typically highly sensitive to the inclusion or exclusion (and position) of other distance variables. Because of the inherent mathematical relationship between any two randomly chosen points on a Cartesian grid, the sensitivities caused by distance to (or from) variables is particularly acute. The resulting demands on the analyst to select the right number of landmarks and their precise locations to affect reliable welfare estimates is considerably, perhaps implausibly, high.

The identity of the optimal location, however, is not a trivial property. First, home location with distance to amenities or disamenities has a profound effect on home prices and on location choices; so tracking the impact, even as an aggregated, indexed good tracks those choices. Second, location cannot be expected to be separable from other household services satisfied by a home purchase; and some home attributes will complement services satisfied by location. So to index successfully the effect of location as an aggregated effect into one variable can be applied to extract consistent estimates of home attributes. Finally, and related, many modified generalized least squares (GLS) corrections for distance effects already

blend multiple, yet very local amenity and near neighbor effects in, say, a spatial weights matrix. While regional models often are designed to track region-wide landmark draws, requiring the analyst to control for a bundled mixture of very local effects by a spatial weights matrix, analysts often are interested in estimating the influence of those very local amenities on home price. Either defined by distance or derived from a distance measure, such as zonal effects (e.g. $x<200$ ft; $200<x<1000$ ft; $x>1000$ ft), to robustly estimate local effects often requires controlling for the larger location effects as a packaged effect. Palmquist (2004) argued for an estimation strategy relatively similar and, while not formalized, arguably for similar reasons.

References

Brasington, D. M., & Hite, D. (2005). Demand for environmental quality: a spatial hedonic analysis. *Journal of Urban Economics, 35*, 57–82.

Cameron, T. A. (2006). Directional heterogeneity in distance profiles in hedonic property value models. *Journal of Environmental Economics and Management, 51*(1), 26–45.

Deaton, B. J., & Hoehn, J. P. (2004). Hedonic analysis of hazardous waste sites in the presence of other urban disamenities. *Environmental Science and Policy, 7*(6), 499–508.

Fik, T. J., Ling, D. C., & Mulligan, G. F. (2003). Modeling spatial variation in housing prices: a variable interaction approach. *Real Estate Economics, 31*(4), 623–646.

Harrison, D., & Rubinfeld, D. (1978). Hedonic housing prices and demand for clean air. *Journal of Environmental Economics and Management, 5*, 81–102.

Ihlanfeldt, K. R., & Taylor, L. O. (2004). Externality effects of small-scale hazardous waste sites: evidence from urban commercial property markets. *Journal of Environmental Economics and Management, 47*(1), 117–139.

Kasriel, R. A. (1971). *Undergraduate topology*. Malabar: Krieger.

Noonan, D. S., Krupka, D. J., & Baden, B. M. (2007). Neighborhood dynamics and price effects of superfund site clean-up. *Journal of Regional Science, 47*(4), 665–692.

Pace, R. K., & LeSage, J. P. (2008). A spatial Hausman test. *Economic Letters, 101*(3), 282–284.

Palmquist, R. B. (1984). Estimating the demand for the characteristics of housing. *The Review of Economics and Statistics, 66*(3), 394–404.

Palmquist, R. B. (2004). Property value models. In K. G. Mäler & J. Vincent (Eds.), *Handbook of environmental economics edition 1* (Vol. 2, pp. 763–819). North-Holland: Elsevier.

Schmeidler, D. (1969). The nucleolus of a characteristic function game. *SIAM Journal on Applied Mathematics, 17*(6), 1163–1170.

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. Cambridge: The MIT.

Zietz, J., Zietz, E. N., & Sirmans, G. S. (2008). Determinants of house prices: a quantile regression approach. *Journal of Real Estate Finance and Economics, 37*(4), 317–333.