Proximity to a Water Supply Reservoir and Dams: Is There Spatial Heterogeneity in the Effects on Housing Prices?

Jeffrey P. Cohen[*], *Joseph Danko*[†], *Ke Yang*[‡]

Abstract

An understanding of the spatial variation in the impacts of living near reservoirs, dams, and undevelopable land is important in explaining residential property values. While there is a body of literature on the effects of proximity to dams and reservoirs on housing prices, little known research attempts to determine if various individual houses are impacted differently depending on their locations and years of sale. We examine properties in Barkhamstead, Connecticut that sold between 2001 and 2015. We utilize non-parametric regression techniques to allow for the possibility that the major reservoirs, dams and undevelopable land areas, affect various house prices differently, depending on their locations and when they are sold. We find that for the most part, proximity to dams leads to lower housing sale prices, with the magnitudes of these effects varying across geographic space and over time. A difference-in-differences approach indicates that the willingness to pay for distance from the dams decreased after the most recent housing crisis. In general, undevelopable land area is valued as an amenity in this rural town. The signs of the effects of proximity to the reservoir vary – some properties benefit from proximity while others experience lower sale prices when they are closer to the reservoir. We also control for other key housing characteristics and environmental variables, such as elevation, numbers of bedrooms and baths, age of properties, year of sale, square footage and acreage, and others. We generate maps of the signs and magnitudes of the coefficients for several of the key variables to illustrate the spatial heterogeneity.

Keywords: House prices; real estate; spatial dependence; non-parametric regression; kernel regression.

JEL Codes: R3

[*]Corresponding author, Center for Real Estate and the School of Business, University of Connecticut, 2100 Hillside Road, Unit 1041-RE, Storrs, CT 06269. Jeffrey.Cohen@business.uconn.edu

[†]Department of Geography, University of Connecticut.

[‡]Barney School of Business, University of Hartford, 200 Bloomfield Ave, West Hartford, CT 06117; Kyang@hartford.edu

1 Introduction

Proximity to reservoirs, wetlands, and dams can have both positive and negative impacts on house prices. For instance, wetlands and reservoirs can be considered "amenities" due to open space, wildlife, and aesthetics/views. On the other hand, wetlands locations can be restrictive to future development, which could be a detriment to a potential buyer who may want to expand a property on a wetland or near a reservoir. Also, there can be a higher risk of flooding near dams, reservoirs, and wetlands, which can be capitalized into house prices and lead to lower property values. An understanding of the potential positive and negative impacts of living near reservoirs, dams, wetlands and undevelopable land due to relatively steep slope, is important in justifying the operation of water (and possibly other) utilities near residential properties.

While there is a body of literature on the effects of proximity to dams and reservoirs on housing prices, little known research attempts to determine if various individual houses are impacted differently depending on their locations and years of sale. Also, relatively little is known about how proximity to these amenities affects house prices differently during a "boom" period versus a "bust" period. We examine properties in Barkhamstead, Connecticut that sold between 2001 and 2015. This covers a period of a significant housing "boom" (2002-2009) and also a "bust" (the housing crisis which began in 2009).

The reservoir in Barkhamstead supplies much of central Connecticut with its drinking water. We utilize non-parametric regression techniques (Geographically Weighted Regressions) to allow for the possibility that the major reservoir and dams in Barkhamstead affect various house prices differently, depending on their locations and when they are sold. We follow the approach of Saiz (2010) and generate estimates of land with sufficiently steep slopes that inhibit development. We also estimate a set of partial linear (i.e., semi-parametric) models. We find that for the most part, proximity to dams leads to lower housing sale prices, with the magnitudes of these effects varying across geographic space and over time. Properties with higher census block steep slope land area tend to sell for higher prices, implying this type of land is a amenity in this rural town. Our difference-in-differences approach in the context of a partial linear model leads us to

conclude that the housing crisis caused proximity to dams to be valued less after 2009. In other words, these lower effects due to proximity to dams are magnified in the post-2009 (housing crisis) period.

The signs of the effects of proximity to the reservoir vary – some properties benefit from proximity while others experience lower sale prices when they are closer to the reservoir. We also control for other key housing characteristics and environmental variables, such as elevation, numbers of bedrooms and baths, age of properties, year of sale, square footage and acreage, and others. We generate maps of the signs and magnitudes of the coefficients for several of the key variables to illustrate the heterogeneity (e.g., see Figures 2 and 3).

The remainder of this paper proceeds as follows. First, we review the literature on proximity to wetlands, dams, and water bodies. Then we describe our empirical approaches, followed by a discussion of the data. After presenting our results, we conclude with a summary of the key findings of the paper.

1.1 Literature Review

There are several studies on the relationships between housing prices and proximity to wetlands, water, and/or dams. However, no known research considers all of these effects together with the impacts of before and after a housing crisis, in a semi-parametric estimation framework.

Cohen et al (2015) consider wetlands and water impacts, but they ignore the important aspects of dams and elevation, and they examine a shorter sample period that stops before the beginning of the housing crisis. They find that various properties are affected differently by proximity to wetlands and water.

Other recent studies include Atreya et al (2016), who find a different effect of distance to the coastline in Texas, depending on flood risk. Ironically, they also find that the willingness to pay for avoiding flood risk is higher for properties that are further away from the shore. However, they do not use a semi-parametric estimation framework to arrive at these conclusions.

Rouwendal (2016) examine the effects of proximity to water, using a sample of identical Dutch houses. This simplifies the hedonic housing problem because it is not necessary to

"control" for differences in characteristics, other than proximity to water. They find that in this context, the potential benefits from water proximity are smaller, possibly due to "specification bias" that can occur in the typical hedonic model.

Lewis et al (2008) examine willingness to pay for removal of a dam in Maine. Their approach is rich in the sense that their examination of house prices pre- versus post- dam removal, for various distances from the dam, enables the identification of the benefits of living far from the dam.

Bohlen and Lewis (2009) study another river and dam in Maine, and in this instance, they find a 16% premium for living closer to the river. They also find a premium for living closer to the dam, although the level of statistical significance is lower for this variable. These conflicting findings of the impacts of a dam on housing prices, for two studies of different dams in the state of Maine, imply that a semi-parametric approach could be fruitful in our case of dams in Barkhamstead, Connecticut.

Saiz (2010) is a more broad study, at the U.S. Metropolitan Statistical Area (MSA) level, of the impacts of water bodies and elevation on the amount of developable land in each MSA. He finds that development is detrimentally affected in MSA's with greater amounts of "steep-slope terrain". He also finds that prices are endogenous, which is not surprising given the relative scarcity of land in many densely developed urban areas.

To further explore these issues and the importance of considering spatial heterogeneity, we control for elevation in a non-parametric manner in our analysis. Our difference-in-differences approach of impacts of distance to dams, before vs. after the housing crisis, is an identification strategy that leads to our key finding of the dis-amenities from proximity to dams being pronounced after the most recent housing crisis. We also find that properties in census block groups with greater steep-slope terrain sell for higher prices, which implies the seep-slope terrain is an amenity in this rural setting.

2 Approach

Our analysis of the impacts of water body, wetlands and dams on housing prices is based on a hedonic housing price model. A typical hedonic model with linear regression function takes the following form:

$$Y_i = X_i\beta + u_i, \quad i = 1, \cdots, N \tag{1}$$

where Y_i is the logarithm of sale price and X_i is a vector of house characteristic variables, including number of baths, bedrooms, square footage, acres, as well as neighborhood variables such as physical locations (longitude and latitude), logarithm of elevations, logarithm of distance to dam, water body and reservoir.

2.1 Locally weighted regression (LWR)

In addition to the ordinary least squares (OLS) estimation of the model, we use a non-parametric approach - locally weighted regressions (LWR), also commonly referred to as Geographically Weighted Regression (GWR) - to approximate the regression function, considering the fact that the data are prices of houses at fixed points with spatial coordinates and years of sale. In a LWR model, the spatial coordinates of the data are used to calculate distances that are used in a kernel function to determine weights of spatial dependence between observations. Time of sales are used similarly to determine weights of time dependence between observations. The hedonic house price function is assumed to take the following form:

$$Y = X_i\beta(s_i, t_i) + u_i, \quad i = 1, \cdots, N$$

where s_i is the geographic location variables of observation i and t_i is the time of sale on observation i, $\beta(s_i, t_i)$ is a column vector of regression coefficients, each of which is an unknown function of s_i and t_i. The coefficient vector at location $s_i = s$ and at time $t_i = t$, denoted by $\beta(s_i = s, t_i = t)$ is calculated by minimizing the following objective function with respect to a

and b,

$$\sum_{i=1}^{N}(y_i - a - b'x_i))^2 K\left(\frac{d_i}{h}\right) K\left(\frac{\tau_i}{h}\right) \tag{2}$$

where $K(\cdot)$ is a kernel function that determines the weight that observation i receives in the regression; d_i and τ_i are the distance between observation i and location (s,t) in geographic space and in time space, respectively;[1] and h is the bandwidth. The Gaussian kernel function is used to calculate the weight assigned to each observation, based on its distance from the target point, both in geographic location and time/year.[2]Many researchers have shown that choice of kernel function has little effect on the results. (See, e.g. McMillen and Redfearn, 2010). The performance of kernel estimator is much more sensitive to the choice of bandwidth, h . Given in the data sets that the houses are located densely in some areas and sparse in other areas, a fixed bandwidth would lead to over-smoothing in areas where many observations are present and under-smoothing in areas with sparse data. Following McMillen and Redfearn (2010) we use a "K^{th} nearest neighbor" (K-nn) approach in calculating the bandwidth. For a target point we chose a bandwidth to include a fixed percentage of the sample into the local averaging.[3]

Following the method suggested in McMillen and RedFearn (2010), we apply a version of an F-test on the significance of each explanatory variable in $f(z_i)$. Let L be the N by N matrix so that $\tilde{Y} = L\tilde{Y} + \epsilon$, where $\tilde{Y} = Y - X \cdot \hat{\beta}$ is the vector of the dependent variable and ϵ is the regression residuals in the GWR regression. Define $d_1 = tr(L)$, $d_2 = tr(L'L)$ and $\kappa = 2d_1 - d_2$. Then the F-test is simply:

$$\frac{(\tilde{Y}'R_r\tilde{Y} - \tilde{Y}'R_a\tilde{Y})/(\kappa_a - \kappa_r)}{(\tilde{Y}'R_aY)/(n - \kappa_a)} \sim F(\kappa_a - \kappa_r, n - \kappa_a)$$

where the subscript a and r are used to indicate whether the quantity is calculated from the

[1] The distances d_i and τ_i are normalized with the standard deviation of $\{d_i\}_{i=1}^N$ and $\{\tau_i\}_{i=1,}^N$.

[2] The kernel function on time assigns positive weight only for $\tau_i \leq 0$ and assigns 0 weight for $\tau_i > 0$, i.e. only those observations that precede the observation at (s,t) in time get positive weights.

[3] We use two window sizes: 50 percent and 100 percent. With a Gaussian kernel function (Standard normal density function) the bandwidth are chosen to include a specified percentage (25 percent or 100 percent) of the sample in the window - two standard deviations of the target point. Sample points outside of this window get near-zero weights and are essentially ignored in the averaging. One could potentially use different thresholds in bandwidth selection, e.g. three or four standard deviations, but this will not change the results significantly.

restricted model (null) or the alternative model. To test the significance of each variable, the above F-statistics can be calculated as with that variable dropped from the model. The P-values from these tests are probabilities of the null hypothesis that the coefficients equal zero. In this sense this F-test indicates whether an explanatory variable in the non-parametric component of the regression adds any explanatory power to the model.

2.2 Partial Linear Regression

While the OLS model imposes too many restrictions on how X affects Y , the locally weighted regression might gives too much degrees of freedom in each point of estimation (i.e. leads to too few observation being used in each point estimation), especially with a relatively small data set. As a compromise in modeling the hedonic price function we also take a semi-parametric approach - a partially linear model - in estimating the average effect of a single variable, say X, of our interest. The partial linear model takes the following form:

$$Y_i = X_i\beta + f(Z_i) + u_i, \quad i = 1, \cdots, N; \tag{3}$$

where X_i is of dimension one, β is a unknown parameter that is of our main interest, Z_i is of dimension $d \times 1$, $f(\cdot)$ is a smooth but otherwise unknown function. The advantage of using a semi-parametric model over a fully non-parametric one is for convenience in interpretation and the faster converging rate, the later of which is particularly important given our sample size. The estimate of β provides an estimate of the conditional expectation of Y_i given X_i after controlling in a general, non-parametric way for the effects of all other variables.

Following Robinson (1988), by taking the expectation of (3) conditional on variables in the non-parametric component, z_{it}, then subtracting it from (3) we have

$$Y_i - E(Y_i|Z_i) = [X_i - E(X_i|Z_i)]'\beta + u_i \tag{4}$$

If we use the following notations: $\nabla_i = Y_i - E(Y_i|Z_i)$, $V_i = X_i - E(X_i|Z_i)$, then we can write

the above equation as

$$\nabla_i = V_i\beta + u_i \tag{5}$$

Then a simple OLS regression of ∇ on V will give a consistent estimate of β, assuming $E(Y_i|Z_i)$ and $E(X_i|Z_i)$ are known. In practice, these conditional expectations can be approximated using locally weighted regression (LWR) following McMillen and Redfearn (2010). We follow Cohen, Osleeb and Yang (2014) by rotating each independent variables in the parametric part of the model, X, and leaving the rest of the independent variables in the non-parametric component of the model, f(z). With this approach we can obtain an estimate of the marginal impact of each individual factor on the housing price after controlling for the effects of all other variables in a non-parametric way.

One interesting question to ask regarding the real estate market is that how the economic downturn, like the great recession in 2008-2009, affects the housing prices. More specifically, we would like to understand if there is a significant change in how home buyers value amenities differently, before and after the economic downturn in 2008/2009. In Figure 1, the coefficients - estimated from the non-parametric LWR model - on variables measuring the distances to water body, wet land or dam are plotted over the year of sales. These plots suggest that the marginal impact of closeness to water/dams changed during 2008-2009. To further test whether these changes are statistically significant, we apply a F-test in the context of the partial linear model in section 4. In model (3) , we assume that β , which captures the marginal effects of variable X, applies to all the observations in our sample. It is straight forward to test whether the marginal effects from variables like distance to water bodies on housing prices change significantly before and after 2008/2009. Our data spans two very different periods. Up to 2009, denoted by period 1, the housing market was booming and prices had been rising steadily. The recession of 2008/2009 marked a transition in the housing market (denoted by period 2), with low demand and over supply, and dropping prices. Denote a dummy variable that takes value of 1 only for observations prior to year 2009 as D^1 and a dummy variable that takes value of 1only for observations in 2009 and later as D^2 , an unrestricted model that allows the coefficients to be

different in the two periods is :

$$Y_i = D_i^1 X_i \beta_1 + D_i^2 X_i \beta_2 + f(Z_i) + u_i, \quad i = 1, \cdots, N. \tag{6}$$

The model (6) can be estimated following the same steps described above for the partial linear model. Given that the estimators for $\beta_1$and β_2 , denoted as b_1, b_2, are well behaved we can test the hypothesis: $H_0 : \beta_1 = \beta_2$using the following F-statistics:

$$F(1, n-2) = \left[(b_1 - b_2)^2 R' \left[s^2 (X_i' D_1' D_2 X_1)^{-1} R \right]^{-1} \right],$$

where $R = [1, -1]$ and $s^2 = e'e/(n-2)$ with e being regression residual vector from regression (3). The tests are done on each individual independent variables with two different bandwidth: 50% and 25%.

2.3 Difference-in-differences Model

One model that is particularly helpful in identifying how the impact of dams have changed before and after the economic recession is known as the "Difference-in-difference" (DD) model. Following Kiel and McClain (1995), the "difference-in-difference" model is as the following:

$$Y_i = \beta_0 + D09_i \beta_1 + nearDM_i \beta_2 + D09_i * nearDM_i \beta_3 + f(Z_i) + u_i, \quad i = 1, \cdots, N. \tag{7}$$

where $D09_i$is the dummy variable that takes value 1 only for observations in 2009 and later (post recession observations), $nearDM_i$ is the dummy variable that takes value 1 for houses located within a certain distance (10% of the the maximum distance in the area) from the target (the nearest water body, wet land or dam), $Z_i$include other control variables. As described in Woodridge (2013), the coefficient β_3 has become known as the "difference-in-difference estimator". With our Barkhamsted housing data set, β_3 is an estimate of the change in the effect of a

dam on home values, after the recession, i.e.

$$\beta_3 = (\bar{Y}_{pos,nr} - \bar{Y}_{pos,fr}) - (\bar{Y}_{pre09,nr} - \bar{Y}_{pre,fr})$$

where $\bar{Y}_{.,.}$ is average home prices, "nr" stands for "near the dam" - homes located within one tens of the maximum distance among all homes, "fr" stands for "farther away from the dam", "pre09" stands for "before 2009" (observations in 2000-2008), "pos09" stands for "observations from 2009 or later". In other words, β_3 is the difference after recession in the average difference of housing prices in two different locations, near the dam versus further away from the dam.

We first estimated model (7) by ignoring $f(\cdot)$, i.e. by OLS with only dummies variables but no control variables; As described Woodridge (2013), β_0 is the average price of a home not near the dam. β_1 is the change in all housing prices in Barkhamsted caused by the recession in 2008-2009. The coefficient β_2 captures the location effect (whether being close to the dam or not) on housing prices without considering the effect from the housing crisis. The parameter of our interest is on the interaction term $D09_i * nearDM_i$, β_3 , measures changes in how consumers value a house's proximity to the dam, caused by the housing crisis.

Then we estimate the above model (7) by including various housing characteristics in Z_i and $f(Z_i)$ is assumed to be a linear function of control variables. The reason for doing this is that the houses selling before the housing crisis might have been systematically different than those selling after the housing crisis. This make it important to control for characteristics that might caused the difference.

At last, to render some flexibility on how those other characteristics might affect house prices, we assume that $f(Z_i)$ in model (7) is a unknown smooth function and estimate it non-parametrically. This model is a variation of the "partial linear model" in (3). The partial linear model allows us to estimate the "difference-in-difference" effect directly while controlling for the effects from other house characteristics in a flexible fashion. The procedure to estimate this non-parametric partial linear model is described in section 3. We use a window size of 50% in estimating the non-parametric components.

3 Data

Barkhamsted is a town in Litchfield County, Connecticut and contains two villages, Pleasant Valley and Riverton. According to the United States Census Bureau, the town has a total area of 38.8 square miles (100 km2), of which, 36.2 square miles (94 km2) of it is land and 2.6 square miles (6.7 km2) of it (6.72%) is water. A high percentage of the land in the town is owned by the State of Connecticut as state forest and by the Metropolitan District Commission as watershed land. Major bodies of water include the Barkhamsted Reservoir, Lake McDonough, and the Farmington River. The Barkhamsted Assessor Department provided the information regarding non-locational characteristics of the single-family houses sold between 2000 and 2015, including: sales price (nominal), year built, year sold, acreage, square footage, number of bedrooms and number of bathrooms. The variables included home address, living area square footage, the age of the property in years, and the year of the sale. Also, data on the number of bedrooms, number of bathrooms, the actual sale price (USD), and the number of acres were compiled. Among all of the single family properties in Barkhamsted, there were 495 houses sold in the period 2000-2015. Following Cohen, Cromley and Banach (2014), we use dummy variables to mark if the property was located either in Riverton or Pleasant Valley, the two of the three villages in Barkhamsted. Properties in neither of these areas are indicated to be in an "Other" category.

The locations of the single-family houses sold between 2000 and 2015 were identified in a two-step process. First, the location of the houses were georeferenced using the addresses provided the Barkhamsted Assessor Department via the mapping function of the Google Fusion Table software (tables.googlelabs.com). Second, the accuracy of the georeferenced data was verified using the MapGeo Barkhamsted GIS System (barkhamstedct.mapgeo.io) in order to ensure that the points representing the locations of the single-family houses sold between 2000 and 2015 were positioned atop (or as close to) the center of the appropriate house. The boundaries of the Riverton and Pleasant Valley neighborhoods, water bodies, wetlands and Barkhamsted reservoir were obtained from Cohen et al. (2014). Maps of elevation, slope, and the location of dams in

Barkhamsted were obtained from the Connecticut Department of Energy and Environmental Protection, or CT DEEP (www.ct.gov/deep/gisdata). Data utilized to calculate the amount of undevelopable land per census block (following the approach for the MSA-level by Saiz, 2010) include the CT DEEP slope map and 2010 United States Census block geography.

Descriptive statistics and a description of the variables are presented in Table 1. The average home sold for about US$247,642; there was no outwardly discernible pattern to the spatial distribution of sales price for individual homes. The highest and lowest quantiles were distributed in all parts of the town in proximity to one another. The average home also has about 1,800 square feet of living area, on a 3.2 acre property, about 850 feet above sea level, about 780 feet from the nearest water body, about 890 feet from the nearest wetlands and 4133 feet from the nearest dam. (see Figure 2 for a map with the distances of each house to the nearest dam). Because water features are not uniformly distributed across the town, homes that are near water features are clustered in different areas and homes that are distant from water are clustered in other areas.

4 Results

Parameter estimates for the different model specifications described above are presented in Tables 2 - 7. The OLS results estimated from model (1) are given in Table 2. The impact from basic house characteristic variables, including property acreage, house age, square footage, number of bedrooms and bathrooms, are consistent with expectations and they are statistically significant. For example, the parameter estimate on the log of the number of acres is 0.0068, implying that every 1 percent increase in lot size drive up the house sale price by 0.0068 percent. The parameter estimate on the log of age is -0.001, implying that sale prices fell by about 0.001 percent for every 1 percent increase in a property's age. In addition, the parameter estimates on the year-dummies are all positive and significant,[4] implying that the sale price was going up over the time period under study, despite the real estate "bust" experienced in some regions of

[4] The coefficients on these time dummies are not included in the table to make it consise but they are available upon request.

the U.S. that started in 2007 - 2008. Meanwhile, houses in Riverton and Pleasantvalley sold for significantly more than houses in the "other" neighborhood. This is consistent with a previous study by Cohen, Cromley and Banch (2014). The parameter estimate on undevelopable land in the census block group is positive and significant, implying undevelopable land is an amenity in this rural town. In contrast, the parameter estimates on geographic measurement of a house, including elevation, distance to water body, wetland or dam, are all insignificant. This makes it difficult to attribute changes in house prices to any of these geographic variables. For example, while the parameter estimate on distance to nearest water bodies is negative, one cannot infer that on average houses closer to water body sold for more than houses that were further from their nearest water body because that estimate is highly insignificant based on this linear model. However, the linearity assumption in OLS might be a over simplification and miss some important aspects of the data set. First, many of the characteristic variables and geographical variables might impact the sale price in a nonlinear fashion and this would be masked by a OLS model. Using distance to the nearest water body as an example, while a 10% increase in the distance from to water body might have a substantial impact on price of houses within immediate vicinity of a lake, the same increase might not affect price of houses at all that are located further away from the lake. Second, as common in real estate studies, spatial dependence, as well as dependence across time period, might play an important role in determining a house's market value. McMillen and Redfearn (2010) discuss how with LWR the "combination of functional form flexibility and spatially varying coefficients helps to reduce spatial auto-correlation without imposing arbitrary contiguity matrices or distributional assumptions on the data". While LWR accounts for spatial dependence, we, in this paper, extend it to allow coefficients varying across both space and time periods. See, for instance, an similar application of LWR in Cohen, Osleeb and Yang, (2014).

Parameter estimates from LWR, with two different bandwidths of 50% and 100%, are summarized in Table 3. Note that as a non-parametric model, actual parameter estimate values change across observations. Table 3 presents only the means of these estimates. Meanwhile, unlike in a parametric model, it is well known that a non-parametric estimate is biased in finite

samples, and the inferences are not possible in a usual manner. As an alternative, following McMillen and Redfearn (2010), we apply a set of F-tests for the significance of each of the explanatory variables. Based on these results, the means of coefficients for most characteristic variables, like a house's age, acreage, square footage, number of bathrooms, are consistent with the OLS model and significant, with the exception of the coefficient on the number of bedrooms being insignificant. The indevelopable land coefficient is once again positive and significant. On the other hand, parameter estimates for geographic measurements, including distances to wetlands, water body or dams, as well as variables on elevation, on the Pleasant Valley or Riverton neighborhood dummies, are mostly insignificant, which again makes it difficult to tell what the model implies in regard to the impact of these variables on the house prices. Reducing the bandwidth from 100% to 50% generally reduces P-value of the F-tests, but not enough to make these variables statistically significant. This issue might be attributed to the fact that non-parametric models typically require a large sample size in order to show consistency. The required sample size increases exponentially with the number of explanatory variables, which is known as the "curse of dimensionality". Given that our data set has only 495 observations but 13 explanatory variables, these results should not be unexpected. It actually provides another motivation for a semi-parametric model specification, as in the partial linear model.

Table 4 shows the partial linear model parameter estimates, with two different bandwidth of 50% and 100%. One advantage of this model is that the parameter estimate from the linear part of the model is well behaved statistically, i.e. converges at rate of square root of N, the same as that of a parametric model. Therefore tests of significance can be done based on the standard normal distribution. An immediate observation from the results in Table 4, as a contrast to the OLS or LWR results, is that all coefficients are highly significant. We argue that the smaller bandwidth is preferred in our partial linear model, because with the bigger window size (100%) we effectively used more observations in estimating a local effect, making it more similar to a parametric model. A smaller window size enables us to better capture the local effects presented in the data. For this reason, our interpretation will be focused on results obtained with smaller window size (50%).

Parameter estimates on house characteristic variables, including a house's age, acreage, square footage, number of bedrooms and bathrooms, are consistent with previous results. Parameter estimates on both Riverton (0.13) and Pleasant Vally (0.1) neighborhood dummies are positive, implying the houses in these two neighbor sold more than houses that are not in either one, with Riverton commanding a more significant premium compared to Pleasantvally. Higher elevation in general decreases a house's sale price, by a 0.046% for every 1% increase in elevation. Moving away from a water body generally drives down a house's sale price, by a magnitude of 0.005% for every 1% increase in distance. This again is consistent with the OLS results, although the magnitude of the estimated impact is slightly smaller. On the other hand, the parameter estimate on wetlands is positive (0.013), implying that consumers prefer to live away from wetlands. They are willing to pay on average 0.013% more for every 1% increase in the distance from the nearest wetland. The parameter estimate on distance to a dam (0.01) reveals that consumer's preference on dams are similar to that on wetlands. They prefer to live further away from a dam, and are willing to pay on average 0.01% more for every 1% increase in distance from a dam. The model suggests that in this particular area, dams as well as wetlands, are viewed as menace rather than amenities. But once again, the coefficient on undevelopable land area in the census block group is positive and significant. This implies that homeowners prefer open spaces, which can be viewed as an amenity.

Finally, Table 5 - 7 presents coefficients from three variations of the "Difference-in-difference" (DD) model. Tables 5a and 5b present results from basic DD models without any control variables. The parameter estimates of interest in Table 5a are: $\hat{\beta}_2 = 0.16$ and $\hat{\beta}_3 = -0.19$, the later of which is significant . This implies that being closer to the dam decreases a house's market value. When we add other housing characteristics as control variables (Table 6a), the estimates on the overall value of the dam and the decline caused by the recession are similar. The results with the non-parametric DD model (Table 7a) also result in more significant estimates. This provides stronger evidence that housing crisis negatively affects consumer's valuation of a house's distance from the nearest dam. In contrast, the DD model that tests for the impact of steep sloped land before versus after 2009, generates an insignificant treatment effects in Table

5b. But when we include other control variables in Tables 6b, the treatment effect for low undevelopable land is negative. This implies more undevelopable increases house prices, and this effect is stronger after the housing crisis. Once again, the results with the non- parametric DD model (Table 7b) also result in more signicant estimates.

5 Conclusions

We estimate a variety of non-parametric and semi-parametric hedonic housing models, and obtain estimates of the effects of proximity to water, wetlands, and dams on housing prices in a small Connecticut town. We find spatial heterogeneity in the effects of dams proximity on housing prices, with properties that are on the east side of the reservoir having a smaller marginal impact on prices than the properties on the west side of the reservoir (See Figure 2). However, the properties in the east are generally further away from the dams than the properties in the west, which implies that the detrimental effects of living near the dams diminish with distance. We also find that wetlands are a disamenity, likely due to the associated development restrictions imposed upon properties that are located on wetlands. Also, the benefits from distance to the dams diminishes after the housing crisis that began in 2009. Clearly, our empirical approaches generate a much richer set of results than we would have obtained with an OLS model.

We also incorporate a measure of "undevelopable land" as in Saiz (2010). While the Saiz (2010) analysis is at the Metropolitan Statistical Area (MSA) level, our undevelopable land estimates are at the Census block group level due to the fact that we are using transaction-level observations as opposed to MSA level data. In all of our models, the undevelopable land coefficient is positive and significant which implies open space is an amenity. This results is not surprising given that we are examining a rural town that is much less densely developed than a metropolitan area. Clearly, our analysis demonstrates that non-parametric and semi-parametric analysis have the potential to generate many additional insights about spatial heterogeneity for hedonic models in the context of properties near wetlands, water, and dams.

References

[1] Atreya, A & Czajkowski, J. Graduated Flood Risks and Property Prices in Galveston County. Real Estate Economics, forthcoming. http://dx.doi.org/10.1111/1540-6229.12163

[2] Baltagi, B.H., Li, Q. (2002). On instrumental variable estimation of semiparametric dynamic panel data models. Economics Letters, 76, 1–9.

[3] Bohlen, C., & Lewis, L. Y. (2009). Examining the economic impacts of hydropower dams on property values using GIS. Journal of Environmental Management, 90, S258-S269.

[4] Cleveland, W.S., and S.J. Devlin. (1988). Locally weighted regression: An approach to regression analysis by local fitting. Journal of the American Statistical Association 83: 596–610.

[5] Cohen, J. P., Cromley, R. G., & Banach, K. T. (2014). Are homes near water bodies and wetlands worth more or less? An analysis of housing prices in one Connecticut town. Growth and Change, 46(1), 114-132.

[6] Cohen, J. P. & Osleeb, J. P. & Yang, K. (2014). Semi-parametric regression models and economies of scale in the presence of an endogenous variable. Regional Science and Urban Economics, Elsevier, vol. 49(C), pages 252-261.

[7] Kiel, K. A., and K. T. McClain (1995). House Prices during Siting Decision Stages: The Case of an Incinerator from Rumor through Operation. Journal of Environmental Economics and Management 28, 241–255.

[8] Lewis, L. Y., Bohlen, C., & Wilson, S. (2008). Dams, dam removal, and river restoration: A hedonic property value analysis. Contemporary Economic Policy, 26(2), 175-186.

[9] McMillen, D. P., & Redfearn, C. L. (2010). Estimation and hypothesis testing for nonparametric hedonic house price functions. Journal of Regional Science, 50(3), 712-733.

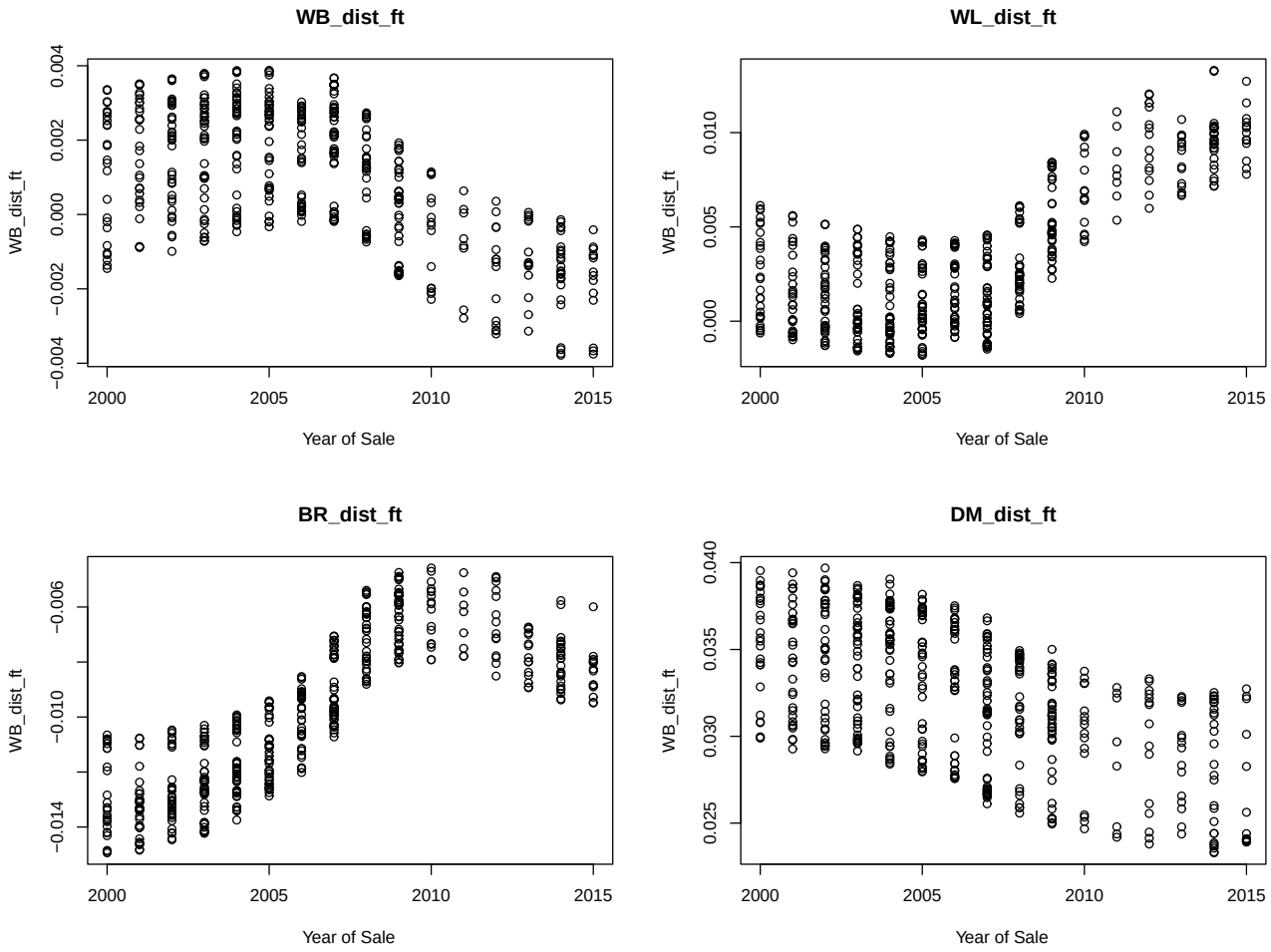
[10] Rouwendal, J., Levkovich, O., & van Marwijk, R. Estimating the Value of Proximity to Water, When Ceteris Really Is Paribus. Real Estate Economics, forthcoming. http://dx.doi.org/10.1111/1540-6229.12143

[11] Saiz, A. (2010). The geographic determinants of housing supply. Quarterly Journal of Economics, 125(3).

[12] Woodridge, J. M. (2012) Introductory Econometrics: A Modern Approach, 5^{th} edition, Cengage Learning.

A Figures

Figure 1: LWR Coefficients Values Over Time [5]



[5] Notation in the graph: "WB_dist_ft" – Log(distance to water body); "WL_dist_ft" – Log(distance to wet land); "BR_dist_ft" – Log(distance to reservoir); "DM_dist_ft" – Log(distance to dam).

Figure 2: Map with Distance to Dams (ft) Coefficients from LWR, Window Size = 50%

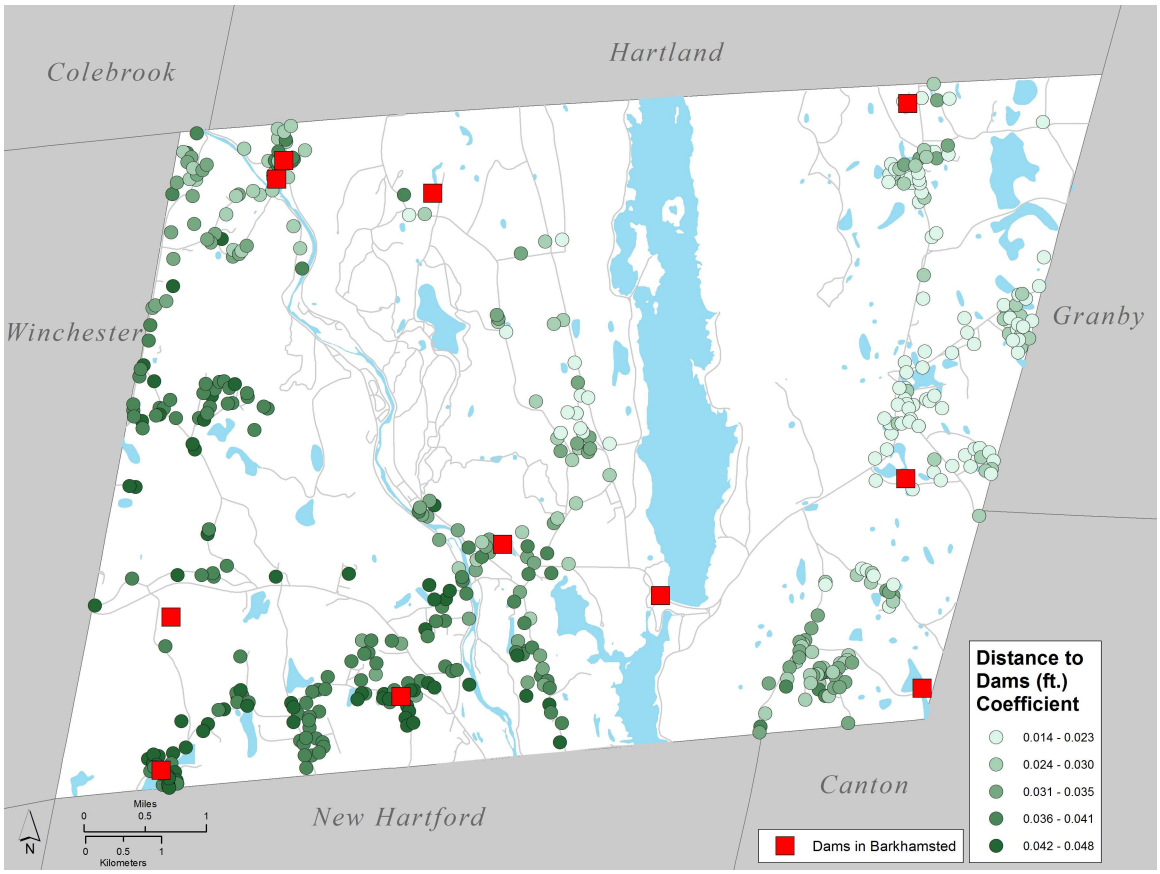
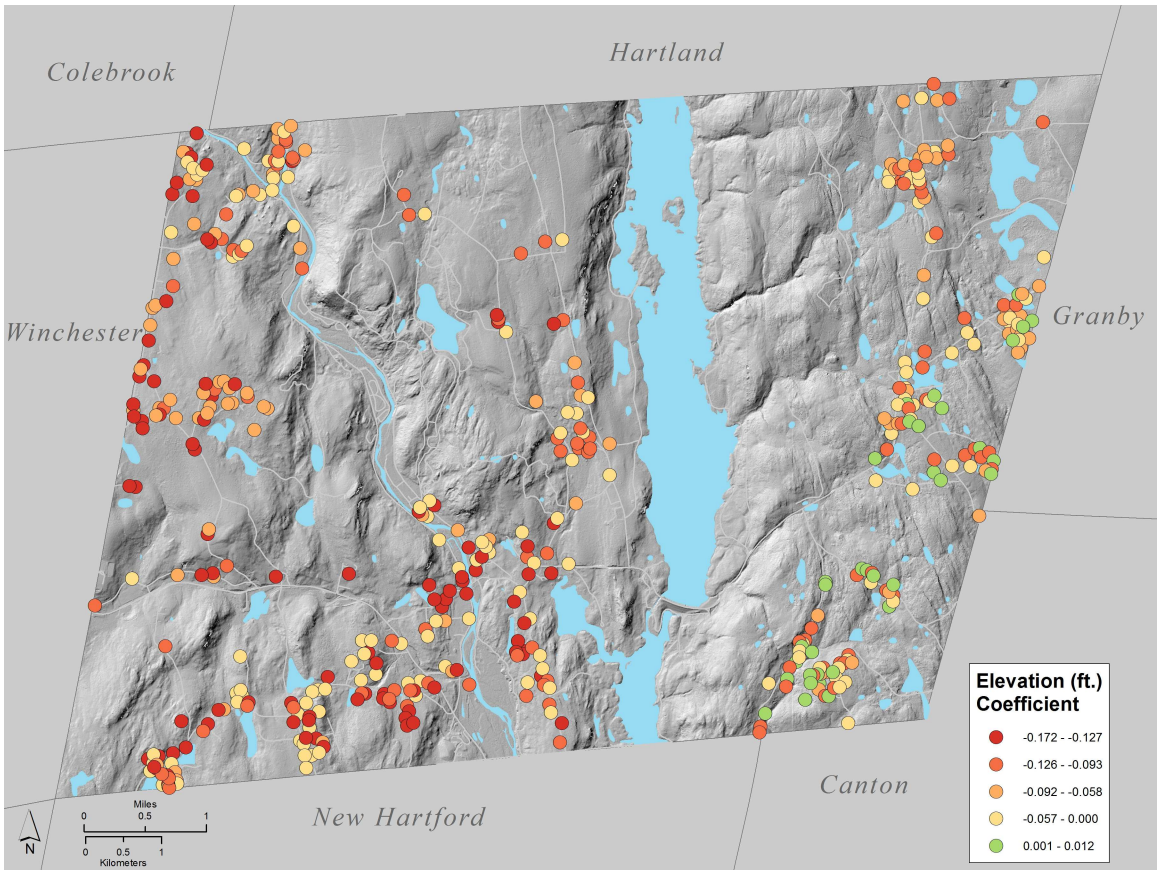


Figure 3: Map with House Elevation Coefficients from LWR, Window Size = 50%



B Tables

Table 1 - Descriptive Statistics/Explanation of Variables

Variables:	mean	std. dev.	min	max	median	Description
Log(house price)	247,642.64	103,216.06	48,500.00	962,500.00	232,000.00	Sales price expressed in historical nominal dollars. (NOMSalePr)
Sale year	7.27	4.09	1.00	16.00	7.00	A reclassification of the Year_Sale category, such that the earlier year sold (2000) is defined as 1, the next year (2001) is defined as 2,…and the final year (2015) is defined as 16. (RCLYr_Sale)
Age	42.18	46.02	0.00	269.00	31.00	The length of time between the actual year built and year of sale for the property in question. Calculated as the difference between Year_Built and Year_Sale. (Age_Yr)
Acreage	3.17	7.67	0.08	110.50	1.43	The amount of land on the property, measured in acres. (Acreage)
SF	1,818.61	679.88	512.00	5,704.00	1,744.00	The amount of living area in the property, measured in square footage.
Bedrooms	2.99	0.81	1.00	7.00	3.00	The number of bedrooms in the property.
Bathrooms	2.05	0.71	1.00	4.50	2.00	The number of bathrooms in the property.
Riverton	0.08	0.28	0.00	1.00	0.00	A dummy variable used to indicate whether the property is located in the Riverton neighborhood, where a value of 1 is used if this condition is true and 0 is used if it is not. (River_NGBH)
Pleasant Valley	0.07	0.26	0.00	1.00	0.00	A dummy variable used to indicate whether the property is located in the Pleasant neighborhood, where a value of 1 is used if this condition is true and 0 is used if it is not. (Pleas_NGBH)
Log(elevation)	855.46	224.79	410.00	1,230.00	935.00	The elevation of the property from sea level in feet, derived from a 5-foot contour map of Barkhamsted. (Elev_ft)
Log(water distance)	784.57	601.00	0.00	2,987.80	622.51	The straight line distance from the property to nearest water body, measured in feet. (WB_dist_ft)
Log(wetland distance)	890.34	637.30	0.00	2,745.98	738.00	The straight line distance from the property to nearest wetland, measured in feet. (WL_dist_ft)
Log(reservoir distance)	11,891.67	5,763.85	1,200.84	24,113.40	10,866.90	The straight line distance from the property to the Barkhamsted Reservoir, measured in feet. (BR_dist_ft)
Log(dam distance)	4,133.85	2,592.27	151.09	10,377.40	3,780.23	The straight line distance from the property to the dams in Barkhamsted, measured in feet. (DM_dist_ft)
Log(undevelopable land)	0.184	0.096	0	0.726	0.193	The share of undevelopable land at the census block level for each housing unit. (Undev_CB)

Table 2 – Ordinary Least Square Regression with year-dummy variables, number of observations: 495

Dependent Variable: Log(House Price)						
Independent Variable	**Estimate**	**standard Error**	**t-value**	**Prob>|t|**	**standardized Estimate**	**Cor with Dep Var**
Constant	11.849687	0.488472	24.258695	0	---	---
Age	-0.000995	0.000341	-2.921163	0.004	-0.11323	-0.096859
Acreage	0.006751	0.001761	3.834405	0	0.128024	0.24117
SF	0.000233	0.000029	8.157498	0	0.392005	0.587614
Bedrooms	0.010209	0.019372	0.526967	0.598	0.02054	0.332975
Bathrooms	0.102471	0.027044	3.789045	0	0.180267	0.50488
Riverton	0.12874	0.061419	2.096089	0.037	0.087848	0.079482
Pleasant Valley	0.067654	0.063818	1.060104	0.29	0.043496	0.12204
Log(elevation)	-0.074909	0.063284	-1.183689	0.237	-0.054972	-0.130079
Log(water distance)	-0.006151	0.017162	-0.358384	0.72	-0.015355	0.019864
Log(wetland distance)	0.010873	0.017703	0.614182	0.539	0.025704	0.025131
Log(reservoir distance)	-0.007059	0.025069	-0.281602	0.778	-0.00993	-0.050186
Log(dam distance)	0.012346	0.017883	0.690341	0.49	0.026628	0.069068
Log(undevelopable land)	0.034872	0.013451	2.592543	0.01	0.092051	0.264946

Notes: A set of dummy variables for each individual year from 2001-20015 are included in the regression. The coefficients on these dummy variables are not included to keep the results concise.

Table 3 – Locally Weighted Regression, number of observation: 495

Dependent Variable: Log(House Price)							
	Window size = 100%				**Window size = 50%**		
Independent Variables	**mean**	**f-test statistics**	**P-value**		**mean estimates**	**f-test statistics**	**P-value**
Constant,	11.835937	0.049908916	3.37E-06		11.930532		
Sale year	0.014637547	0.019530432	0.007536454		0.023307482	0.061702612	7.89E-07
Age	-0.000939874	0.031320475	0.000358089		-0.00093	0.032293676	0.002814858
Acreage	0.007536329	0.13240653	0		0.007341998	0.031649884	0.001277696
SF	0.000236873	0.001223101	0.71571815		0.000242488	0.14501238	3.87E-13
Bedrooms	0.006943928	0.023170991	0.003848697		0.008729055	0.004717384	0.63064273
Bathrooms	0.086006879	0.008490036	0.068033713		0.094762272	0.033790652	0.003020605
Riverton	0.15357422	0.006219834	0.14583896		0.13769192	0.007348239	0.15824658
Pleasant Valley	0.10082647	0.003994953	0.2938828		0.072668129	0.010739218	0.10345212
Log(elevation)	-0.057844936	0.000444276	0.86077814		-0.0754517	0.005485869	0.38067747
Log(water distance)	0.00193721	0.001956782	0.56641575		0.003427336	0.000967383	0.94634944
Log(wetland distance)	0.006825122	0.000806289	0.71252051		0.001346175	0.003922995	0.63372945
Log(reservoir distance)	0.001267426	0.002248557	0.50507399		0.001605199	0.002584599	0.62456476
Log(dam distance)	0.020304169	0.016270136	0.016369573		0.01945129	0.0020523	0.79240391
Log(undevelopable land)	0.036133167	0.049908916	3.37E-06		0.041177997	0.021307558	0.023907947

Table 4 – Partial Linear Model, number of observations: 495

Dependent Variable: Log(House Price)							
	Window size = 100%				**Window size = 50%**		
Independent Variables	**estimates**	**std dev**	**p-value**		**estimates**	**std dev**	**p-value**
Sale year	0.014703576	1.45E-05	4.45E-308		0.014593432	1.79E-05	4.45E-308
Age	-0.000985628	1.25E-07	4.45E-308		-0.00107	1.23E-07	4.45E-308
Acreage	0.007013448	3.36E-06	4.45E-308		0.006593148	3.23E-06	4.45E-308
SF	0.000234764	8.83E-10	4.45E-308		0.000230013	8.37E-10	4.45E-308
Bedrooms	0.005871107	0.000399396	6.45E-49		0.005020374	0.000375901	1.10E-40
Bathrooms	0.087854818	0.00078013	4.45E-308		0.089875722	0.000735428	4.45E-308
Riverton	0.14029175	0.00472622	1.25E-193		0.13130721	0.005499128	5.21E-126
Pleasant Valley	0.096429086	0.004533533	2.14E-100		0.10303589	0.004487951	1.22E-116
Log(elevation)	-0.058524186	0.004657135	3.22E-36		-0.04591	0.004938931	1.46E-20
Log(water distance)	-0.001152824	0.000325839	0.000403152		-0.00490	0.00031507	1.27E-54
Log(wetland distance)	0.009233136	0.000343236	2.18E-159		0.013283488	0.000342046	4.45E-308
Log(reservoir distance)	0.002444589	0.000782893	0.001793185		0.002126198	0.000924416	0.021445696
Log(dam distance)	0.015656994	0.000366879	4.45E-308		0.009542507	0.000374597	3.82E-143
Log(undevelopable land)	0.035460237	0.000201648	4.45E-308		0.035328732	0.000199565	4.45e-308;

Table 5a: Effects of Distance to DM on House Prices, before and after 2008-2009 Recession (A Difference-in-Difference Linear Model with No Control Variables)

Dependent Variable: Log(House Price)						
Independent Variable	**Coefficient Estimate**	**Standard Error**	**t-value**	**Prob >|t|**	**Standardized Estimate**	**Cor with Dep Var**
CONSTANT	12.533102	0.044542	281.378511	0	---	---
D2009	-0.205743	0.04914	-4.18684	0	-0.192991	-0.23477
nearDM	0.128306	0.130622	0.982267	0.326	0.110021	-0.140513
D2009*nearDM	-0.333636	0.141593	-2.356301	0.019	-0.267635	-0.199626
Other Control Variables	None					

Table 5B: Effects of Undevelopable Land on House Prices, before and after 2008-2009 Recession (A Difference-in-Difference Linear Model with No Control Variables)

Dependent Variable: Log(House Price)						
Independent Variable	**Coefficient Estimate**	**Standard Error**	**t-value**	**Prob>|t|**	**Standardized Estimate**	**Cor with Dep Var**
CONSTANT	12.547226	0.04274	293.572	0	---	---
D2009	-0.217103	0.047432	-4.5771	0	-0.203648	-0.23477
lowUnd	0.017101	0.198176	0.08629	0.931	0.013718	-0.200094
D2009*lowUnd	-0.257913	0.205962	-1.2522	0.211	-0.20067	-0.220501
Other Control variables	None					

Table 6a: Effects of Distance to DM on House Prices, before and after 2008-2009 Recession (A Difference-in-Difference Linear Model with Other Control Variables)

Dependent Variable: House Prices						
Independent Variable	**Coefficient Estimate**	**Standard Error**	**t-value**	**Prob>|t|**	**Standardized Estimate**	**Cor with Dep Var**
Constant	11.804565	0.513822	22.974029	0	---	---
D2009	0.063132	0.045979	1.373046	0.17	0.059219	-0.23477
nearDM	0.184014	0.105377	1.746244	0.081	0.15779	-0.140513
D2009*nearDM	-0.342867	0.117251	-2.924205	0.004	-0.27504	-0.199626
Sale year	0.014278	0.003409	4.188128	0	0.14441	0.168019
Age	-0.000879	0.000393	-2.238725	0.026	-0.100013	-0.096859
Acreage	0.00727	0.001877	3.87354	0	0.137849	0.24117
SF	0.00025	0.000032	7.798907	0	0.420687	0.587614
Bedrooms	-0.000819	0.020647	-0.039662	0.968	-0.001648	0.332975
Bathrooms	0.080567	0.028712	2.80605	0.005	0.141733	0.50488
Riverton	0.176275	0.065476	2.692219	0.007	0.120285	0.079482
Pleasant Valley	0.069666	0.068656	1.014713	0.311	0.04479	0.12204
Log(elevation)	-0.059892	0.066674	-0.898292	0.369	-0.043952	-0.130079
Log(water distance)	-0.005275	0.018298	-0.28827	0.773	-0.013168	0.019864
Log(wetland distance)	0.016206	0.018836	0.860346	0.39	0.038311	0.025131
Log(reservoir distance)	0.017284	0.027145	0.636749	0.525	0.024314	-0.050186
Log(undevelopable land)	0.028511	0.014478	1.969305	0.049	0.075259	0.264946

Note: The "lowUnd" is a dummy variable that takes value "1" for houses located in a census block with less (than 10% of the sample maximum) undevelopable land, and value "0" otherwise.

Table 6b: Effects of Undevelopable Land on House Prices, before and after 2008-2009 Recession (A Difference-in-Difference Linear Model with Other Control Variables)

Dependent Variable: House Prices						
Independent Variable	**Coefficient Estimate**	**Standard Error**	**t-value**	**Prob>|t|**	**Standardized Estimate**	**Cor with Dep Var**
Constant	11.688108	0.530025	22.051983	0	---	---
D2009	0.051347	0.045769	1.121867	0.262	0.048164	-0.23477
lowUnd	0.297856	0.162582	1.832032	0.068	0.238933	-0.200094
D2009*lowUnd	-0.385936	0.16739	-2.305604	0.022	-0.300278	-0.220501
Sale year	0.013916	0.003446	4.038475	0	0.140746	0.168019
Age	-0.001032	0.000393	-2.625753	0.009	-0.117464	-0.096859
Acreage	0.007371	0.0019	3.879905	0	0.139766	0.24117
SF	0.000255	0.000032	7.896986	0	0.42955	0.587614
Bedrooms	0.003775	0.020877	0.180813	0.857	0.007595	0.332975
Bathrooms	0.083576	0.028948	2.887148	0.004	0.147027	0.50488
Riverton	0.16536	0.066095	2.50186	0.013	0.112837	0.079482
Pleasant Valley	0.107009	0.068783	1.55574	0.12	0.068798	0.12204
Log(elevation)	-0.059091	0.068425	-0.863579	0.388	-0.043364	-0.130079
Log(water distance)	0.0029	0.018517	0.156634	0.876	0.007241	0.019864
Log(wetland distance)	0.005143	0.018913	0.271915	0.786	0.012157	0.025131
Log(reservoir distance)	-0.000578	0.026855	-0.021536	0.983	-0.000814	-0.050186
Log(dam distance)	0.02724	0.019433	1.401745	0.162	0.058753	0.069068;

Note: The "lowUnd" is a dummy variable that takes value "1" for houses located in a census block with less (than 10% of the sample maximum) undevelopable land, and value "0" otherwise.

Table 7a: Effects of Distance to DM on House Prices, before and after 2008-2009 Recession (A Semi-parametric Difference-in-Difference Partial Linear Model with Other Control Variables)

Dependent Variable: House Prices						
Independent Variable	**Coefficient Estimate**	**Standard Error**	**t-value**	**Prob>|t|**	**Standardized Estimate**	**Corr with Dep Var**
D2009	0.067434	0.043568	1.547806	0.122	0.072215	0.04058
nearDM	0.194833	0.099201	1.964018	0.05	0.183202	-0.045653
D2009*nearDM	-0.304508	0.110337	-2.759806	0.006	-0.25698	-0.093761
Control variables	Full set*					

Table 7b: Effects of Undev_CB on House Prices, before and after 2008-2009 Recession (A Semi-parametric Difference-in-Difference Partial Linear Model with Other Control Variables)

Dependent Variable: House Prices						
Independent Variable	**Coefficient Estimate**	**Standard Error**	**t-value**	**Prob>|t|**	**Standardized Estimate**	**Corr with Dep Var**
D2009	0.063711	0.043541	1.463234	0.144	0.067278	0.040029
lowUnd	0.307897	0.154673	1.990628	0.047	0.3077	-0.060371
D2009*lowUnd	-0.394937	0.158545	-2.491004	0.013	-0.386264	-0.086997
Control variables	Full set*					

Note: The "lowUnd" is a dummy variable that takes value "1" for houses located in a census block with less (than 10% of the sample maximum) undevelopable land, and value "0" otherwise. The control variables includes house characteristic variables as included in Table 6a & 6b. We did not include the estimates in the report to keep it concise.