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 bodies of water, 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. In general, undevelopable land area is valued as an amenity in this rural town. The signs of the effects of proximity to the nearest body of water vary – some properties benefit from proximity while others experience lower sale prices when they are closer to water. We also control for other key housing characteristics and environmental variables, such as elevation relative to the nearest dam, numbers of bedrooms and baths, age of properties, year of sale, square footage and acreage, and others. We plot the parameter estimates over time for some variables to demonstrate how the spatial heterogeneity changes after the recession that began in late 2008.

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

JEL Codes: R3

*Date: June 15, 2017. We thank participants in a 2017 ASSA session, the 2017 Florida State University and University of Florida Real Estate Critical Issues Symposium, the 2017 Midwest Economics Association, and the 2017 American Real Estate Society meetings, for helpful comments. Any remaining errors are our own responsibility. We also want to thank the Barkhamstead assessor's office for providing much of the data.

[†]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. The author gratefully acknowledges financial support from the Lincoln Fundation.

1 Introduction

Proximity to reservoirs and dams can have both positive and negative impacts on house prices. For instance, reservoirs can be considered "amenities" due to open space, wildlife, and aesthetics/views. On the other hand, there can be a higher risk of flooding near dams and reservoirs, 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, 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 a similar 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 with higher elevations than properties 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.

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 figures showing the range of the coefficients for several of the key variables to illustrate the heterogeneity (e.g., see Figure 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 describe some robustness checks and finally 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 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, undevelopable land, and elevation, and they examine a shorter sample period that stops before the beginning of the housing crisis. They find that while overall water is an amenity, various properties are affected differently by proximity to wetlands and water. Their results are different from the findings in our paper because their focus was on wetlands and water bodies, while here we have relatively few properties in the wetlands, and we focus more of our attention here on the impacts of elevation of nearest dam and undevelopable land.

Other relevant 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.

McKenzie and Levendis (2010) consider elevation of houses (although not the relative elevation with respect to dams), and they find that higher elevation houses tend to sell for higher prices.

Another important consideration is whether or not to examine flood zones. Speyer and Ragas (1991) note that there can be biases when using a dummy variable to represent flood zones, because the FEMA flood zones typically encompas broad areas. Therefore, a flood zone dummy likely also reflects the effect of other factors besides being in a flood zone.

In rural areas the issue of undevelopable land is worthy of attention. 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".

To further explore these issues and the importance of considering spatial heterogeneity, we control for elevation relative to the nearest dam, and undevelopable land area, in a non-parametric manner in our analysis. We find that properties in census block groups with greater steep-slope terrain sell for higher prices, which implies the steep-slope terrain is an amenity in this rural setting (in contrast to the Saiz, 2010 disamenity findings for MSAs, which mostly are

comprised of urban areas where land is relatively scarce).

2 Approach

Our analysis of the impacts of water bodies and dams on housing prices is based on a hedonic housing price model. Our 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 distance to the nearest water body, and dummy variables such as whether a property's elevation is below the nearest dam, and a shift dummy for whether a house is sold after the start of the 2008 recession.

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 the kernel estimator is much more sensitive to the choice of bandwidth, h . Given that the houses in our dataset 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 Cleveland and Devlin (1988), we apply a version of an F-test for 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 LWR regression. Define $d_1 = tr(L)$, $d_2 = tr(L'L)$ and $\kappa = 2d_1 - d_2$. Then the F-test is simply:

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

[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.

where the subscript a and r are used to indicate whether the quantity is calculated from the 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 may impose too many restrictions on how X affects Y , the locally weighted regression might give too many degrees of freedom in each point of estimation (i.e. it may lead to too few observations 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 Baltagi and Li (2002), and Cohen, Osleeb and Yang (2014), by rotating each independent variable 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.

3 Data

Barkhamsted is a town in Litchfield County, Connecticut and contains three villages, Pleasant Valley, Riverton, and the remainder of the town. 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. In total, there are 204 water features that we consider in our analysis. The Barkhamsted Assessor Department provided the information regarding non-locational characteristics of the single-family houses sold between 2001 and 2015, including: sales price (nominal),[4] 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

[4] Typically, in parametric hedonic models researchers include time fixed effects when using nominal sales prices. But in a nonparametric or semi-parametric model as we use here, we include time in the kernel function, so we are already controlling for year of sale in the estimations.

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 2001-2015. 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 2001 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. (2015). 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. We also determine whether each property is at the same or lower elevation than the nearest dam, and generate a dummy variable equal to 1 in this case, and 0 otherwise.

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, and 4133 feet from the nearest dam. See Figure 2 for a map of the relative elevation of each house with respect 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.[5]

4 Results

Parameter estimates for the different model specifications described above are presented in Tables 2 - 4. 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 0.0078, implying that every 1 percent increase in lot size drive up the house sale price by 0.0078 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 estimate on the post-2008 dummy is negative and significant, implying that sales prices were going down durng and after the real estate "bust" experienced in most parts of the U.S. that started in late 2008. Meanwhile, houses in Riverton and Pleasant Valley sold for significantly more than houses in the "other" neighborhood. This is consistent with the previous Barkhamsted study by Cohen et al (2015). The parameter estimate on undevelopable land in the census block group is positive and insignificant. Also, the parameter estimates on elevation relative to the nearest dam and distance to the nearest water body are insignificant. This makes it difficult to attribute changes in house prices to these geographic variables. For example, while the parameter estimate on distance to nearest water bodies is positive, one cannot infer that on average houses closer to water body sold for less 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

[5] A helpful reviewer suggested we consider developable land sales in addition to improved single family homes. While we have data on the undevelopable land area by Census block group, the Barkhamsted assessor indicated that she does not have data available to share with us on specific sales of developable land.

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 to the 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).

Finally, it is important to note that we avoiding including too many distance variables in the regressions, since changing the distance to one amenity also impacts the distance to another, despite the fact that the regression parameters assume all other regressors are held constant. Therefore, based on the guidance provided by Ross et al (2011), we include one distance variable, the distance to the nearest water feature.

Parameter estimates from LWR, with two different window sizes of 50% and 100%, are summarized in Table 3. Note that with a non-parametric model, actual parameter estimate values change across observations. Table 3 presents the means of these estimates. Meanwhile, unlike in a parametric model, it is well known that a non-parametric estimate is biased on finite samples, and the inferences are not possible in a usual manner. As an alternative, following Cleveland and Devlin (1988), 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 all of the characteristic variables are consistent with the OLS model and significant, with the exception of the coefficient on the number of bedrooms being insignificant. The undevelopable land coefficient is once again positive but insignificant. The parameter estimates on the post 2008 dummy, distance to nearest water body, and elevation relative to the nearest dam, are 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 window size 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 12 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 window sizes 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 window size 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 (0.150) and Pleasant Valley (0.145) 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 Pleasant Valley. Lower elevation relative to the nearest dam in general decreases a house's sale price. Moving away from a water body generally drives up a house's sale price, by a magnitude of 0.01% for every 1% increase in distance. This again is consistent with the OLS results, although the magnitude of the estimated impact here is slightly smaller. The model suggests that in this

particular area, locations below dams as greater distance from water, are viewed as disamenities. In contrast to the fully nonparametric model, with the semi-parametric approach the coefficient on undevelopable land area in the census block group is positive and significant. This implies that these rural homeowners, where land is less scarce than in urban areas, prefer open spaces. Therefore undevelopable land can be viewed as an amenity.

5 Robustness Checks

We consider several robustness checks in our analysis. First, a helpful reviewer pointed out that there is a 110 acre lot in our sample, and it might be worthwhile to think about dropping that observation. We considered dropping additional outlier observations, based on the fact that the mean lot size is 3.2 acres, and the standard deviation is 7.67 acres. Therefore, we expect approximately 99 percent of the observations to fall within 2 standard deviations of the mean, or less than 19 acres. This led us to drop the 13 observations that were above 19 acres in a robustness check. In this truncated sample, the signs and significance of all the variables were the same with our OLS specification. Therefore, we decided to proceed using the full sample.

A helpful reviewer also suggested we examine whether using the centroid of the property in the distance and kernel calculations would lead to different results. We found that using the centroids had no impact on the signs and significance of the results.

We also explored how the presence of water on a property impacted the results. We tried adding a dummy variable that equals 1 if a property had a water feature on it, and 0 otherwise. There were only 8 properties that had part or all of a water body on the property, and our parameter estimate for this coefficient was negative and insignificant, while all the signs and significance of the other coefficientes in the model were unaffected. Therefore, we omitted this dummy variable from our preferred specification.

Another consideration that suggested by a reviewer was to examine sales of developable land. Unfortunately the town assessor does not have historical data available on developable land for the time period of our analysis.

One might wonder why Cohen et al (2015) found that water proximity was an amenity, while in the present study it is a disamenity. Cohen et al (2015) include several distance variables, but as noted above, Ross et al (2011) indicate that greater than 2 distance variables in a hedonic model can lead to inconsistent estimates. Also, in the present study, we have incorporated data on undevelopable land and relative height of the nearest dam, while Cohen et al (2015) did not include these variables in their analysis. These issues may explain the discrepancies between Cohen et al (2015) and our findings.

While the concentration of dams across space might be an issue to be considered, the relative height of a property with respect to the nearest dam is a greater concern. If a property is next to, and higher than, a cluster of dams, there is no risk of flooding if the dam were to break. Therefore, we chose to explore the issue of elevation relative to the nearest dam.

Speyer and Vagas (1991) find that using flood zone dummy varables can lead to biases because flood zone data typically encompases broad areas. Therefore the flood zone dummy can reflect other factors besides flood risk. For this reason, together with the fact that the only flood zone data available for Barkhamsted, CT is the Q3 Flood Zone Data from a time period much earlier than all of the property sales in our sample, we do not consider flood zone data. But, we extensively consider how elevation of the closest dam relative to a property's elevation affects sale prices. Our hypothesis is that properties that are lower than the nearest dam sell for less than those that are higher than the nearest dam.

There are several potential issues related to the impacts of a recession that are worth addressing. One interesting question to ask regarding the real estate market is how an economic downturn, like the recession beginning in late 2008, 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 nearest water body, undevelopable land, and the dummy for elevation relative to the nearest dam, are plotted against the year of sale. These plots suggest that the marginal impact of closeness to water/dams changed around the time of the recession. Another issue related to the

recession impacts is contagion. The potential of contagion across properties from the recession is factored into the kernel weights already, since the LWR approach allows for the possibility that nearby property sales affect the price of a particular property. Contagion across towns would be interesting to consider, however the data on property sales in neighboring towns are not ready available to us, and such an analysis is therefore beyond the scope of this particular study. Real estate taxes after the recession change uniformly across the town over time, so these are also controlled for in the kernel weights. Also, changes in insurance costs are controlled for in the post-2008 dummy, since in 2009 there was a sharp increase in insurance costs in general. Fluctuations in household income and employment are not available to us on a household level.

Finally, one might argue that if some properties were short sales or foreclosure sales, these properties could have different effects than arms-length transactions. We compared the list of property sales in our dataset with a list of short sales and foreclosures obtained from the town assessor, and there was no overlap between these two datasets.

6 Conclusions

We estimate a variety of non-parametric and semi-parametric hedonic housing models, and obtain estimates of the effects of proximity to water, elevations relative to the nearest dam, and undevelopable land, on housing prices in a small Connecticut town. We find spatial heterogeneity in the effects of dams proximity on housing prices. Also, property values fell after the housing crisis that began in 2009 (which occurred simultaneously as increased flood insurance rates). Clearly, our semi-parametric and nonparametric empirical approaches generate a much richer set of results, with more significant parameter estimates, than we 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 result 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 analyses have the potential to generate many additional insights about spatial heterogeneity for hedonic models in the context of properties near water, undevelopable land, 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. (2015). 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] McKenzie, R., & Levendis, J. (2010). Flood hazards and urban housing markets: The effects of Katrina on New Orleans. The Journal of Real Estate Finance and Economics, 40(1), 62-76.

[10] 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.

[11] Ross, J. M., Farmer, M. C., & Lipscomb, C. A. (2011). Inconsistency in welfare inferences from distance variables in hedonic regressions. The Journal of Real Estate Finance and Economics, 43(3), 385-400.

[12] 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

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

[14] Speyer, J. & Ragas, W.R. (1991). Housing Prices and Flood Risk: An Examination Using Spline Regression. Journal of Real Estate and Finance Economics.

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

A Figures

Figure 1 – Locations of Water Bodies in Barkhamsted, CT

Figure 2 – Locations of Barkhamsted Single Family Property Sales and Their Elevations, and Elevations of Dams (N=495, t=2001,2002,. . . ,2015)

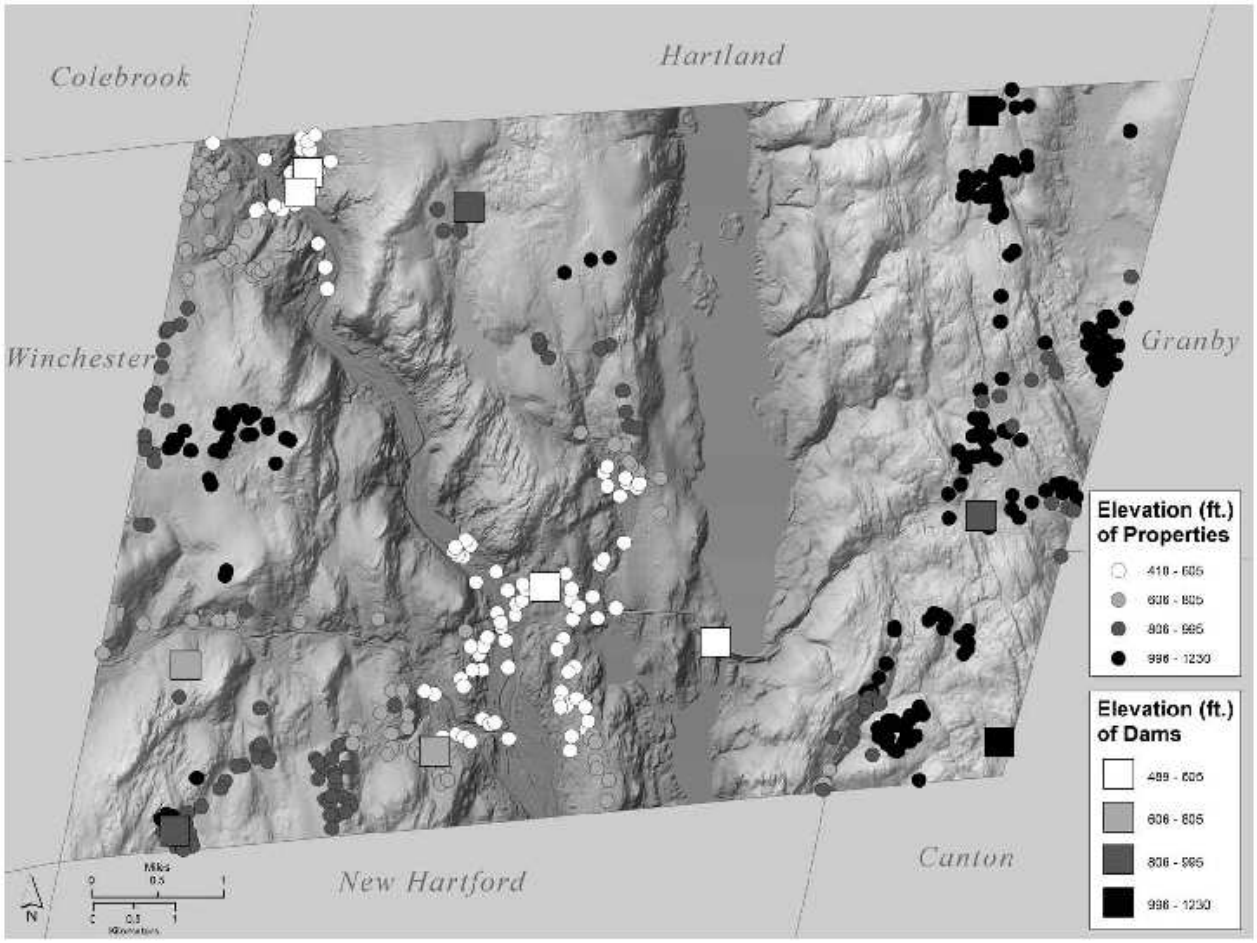
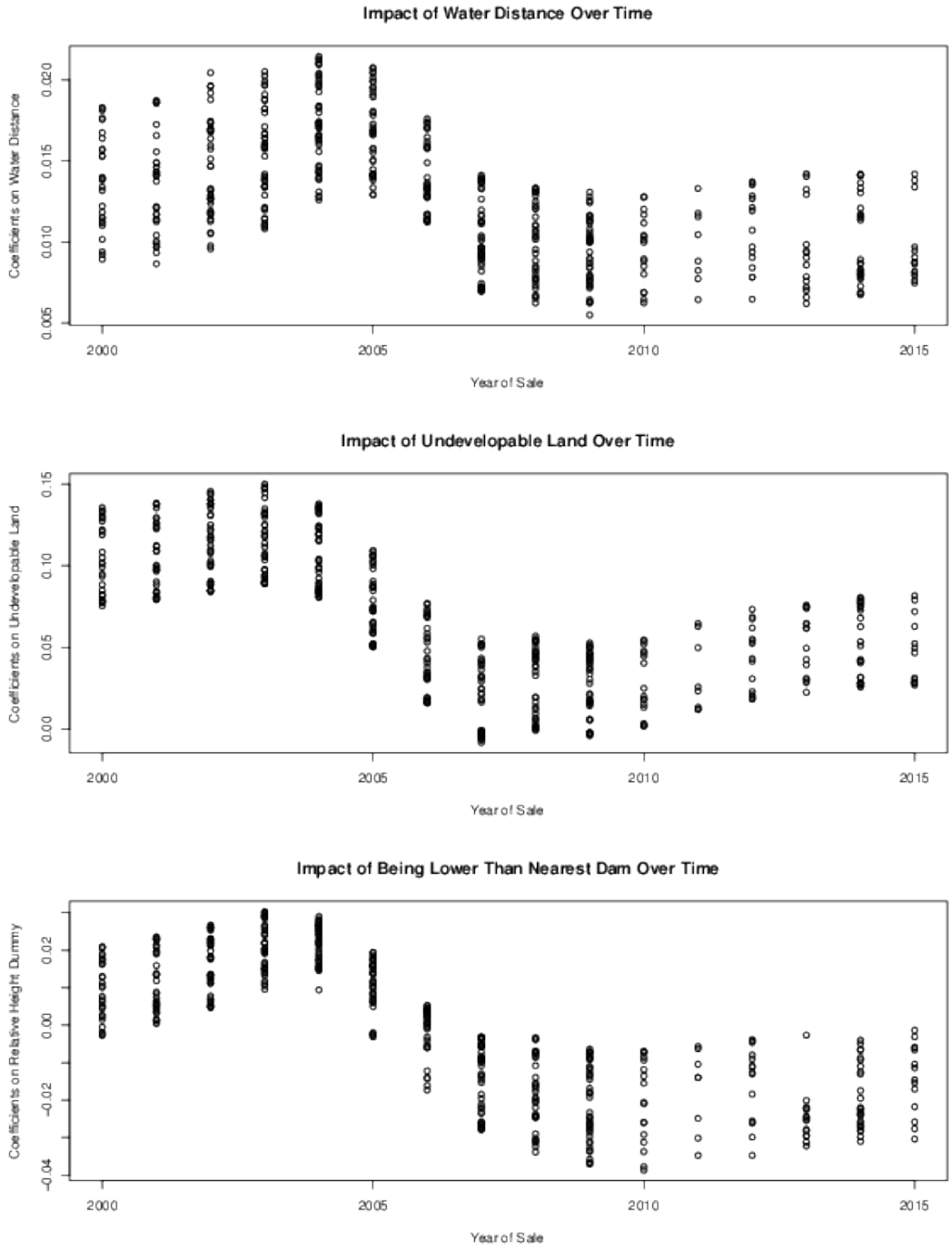


Figure 3: LWR Coefficients Values Over Time



B Tables

Table 1 - Descriptive Statistics/Explanation of Variables (Years 2001 to 2015, n=495)

Variables:	mean	std. dev.	min	max	median	Description
house price	247,642	103,216	48,500	962,500	232,000	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	679	512	5,704	1,744	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)
Water distance	784	601	0	2,987	622	The straight line distance from the property to nearest water body, measured in feet. (WB_dist_ft)
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)
Relative Height	0.309	0.463	0	1	0	A dummy variable used to indicate whether the property is higher or lower elevation than the nearest dam, where a value of 1 if the property is lower than the nearest dam and a value of 0 if it is higher.
Dummy 2008	0.283	0.451	0	1	0	A dummy equal to 1 if a property sale occurred after 2008, and 0 otherwise.

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|**
constant,	11.69719	0.12556	93.16284	0.00000
Dummy 2008	-0.00677	0.03200	-0.21153	0.83300
Age	-0.00101	0.00036	-2.79429	0.00500
Acreage	0.00783	0.00191	4.09444	0.00000
Square Footage	0.00028	0.00003	10.42544	0.00000
Bedrooms	0.00638	0.02114	0.30176	0.76300
Bathrooms (>1)	0.12867	0.04539	2.83470	0.00500
Riverton	0.15802	0.05934	2.66302	0.00800
Pleasant Valley	0.13609	0.05877	2.31583	0.02100
Log(Water distance	0.01116	0.01461	0.76378	0.44500
Log(Undevelopable land)	0.05950	0.05672	1.04902	0.29500
Relative height dummy	-0.00614	0.03314	-0.18516	0.85300

Table 3 – Locally Weighted Regression, number of observations: 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	11.69077	----------	----------		11.66488	----------	----------
Dummy 2008	-0.01013	0.00530	0.16139		-0.01168	0.00855	0.10301
Age	-0.00100	0.01964	0.00727		-0.00098	0.02934	0.00495
Acreage	0.00792	0.03479	0.00015		0.00789	0.03628	0.00043
Square Footage	0.00028	0.23697	0.00000		0.00029	0.25723	0.00000
Bedrooms	0.00734	0.00199	0.58419		0.00914	0.00560	0.54780
Bathrooms (>1)	0.13124	0.01936	0.00845		0.14083	0.02442	0.01626
Riverton	0.15782	0.01425	0.01631		0.15964	0.01329	0.04770
Pleasant Valley	0.13083	0.01325	0.02202		0.10997	0.01706	0.02177
Log(Water distance)	0.01139	0.00137	0.66331		0.01275	0.00211	0.84103
Log(Undevelopable land)	0.05843	0.00451	0.30447		0.06629	0.00974	0.26338
Relative height dummy	-0.00533	0.00178	0.60022		-0.00051	0.00536	0.53400

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

Dependent Variable: Log(House Price)							
	Window size = 100%				**Window size = 50%**		
Independent Variables	**estimates**	**std error**	**p-value**		**estimates**	**std error**	**p-value**
Dummy 2008	-0.02283	0.00121	0.00000		-0.02360	0.00147	0.00000
Age	-0.00102	0.00000	0.00000		-0.00107	0.00000	0.00000
Acreage	0.00751	0.00000	0.00000		0.00713	0.00000	0.00000
Square Footage	0.00028	0.00000	0.00000		0.00028	0.00000	0.00000
Bedrooms	0.00506	0.00043	0.00000		0.00167	0.00041	0.00004
Bathrooms (>1)	0.13105	0.00196	0.00000		0.12925	0.00188	0.00000
Riverton	0.15621	0.00361	0.00000		0.15007	0.00379	0.00000
Pleasant Valley	0.14019	0.00347	0.00000		0.14596	0.00357	0.00000
Log(Water distance)	0.01064	0.00021	0.00000		0.00950	0.00021	0.00000
Log(Undevelopable land)	0.05669	0.00314	0.00000		0.05673	0.00312	0.00000
Relative height dummy	-0.00574	0.00105	0.00000		-0.00232	0.00101	0.02237