Online Appendix to "Natural Amenities, Neighborhood Dynamics, and Persistence in the Spatial Distribution of Income"

Sanghoon Lee   Jeffrey Lin

March 8, 2017

A Theory Appendix

A.1 Proof of Proposition 5

We write the Markov transition matrix as

$$M \equiv \left\{ \begin{array}{cc} \Pr(S_1|S_1) = 1 - \Pr(S_2|S_1) & \Pr(S_1|S_2) \\ \Pr(S_2|S_1) & \Pr(S_2|S_2) = 1 - \Pr(S_1|S_2) \end{array} \right\} \tag{13}$$

and define the steady-state vector π as

$$\pi = M\pi \tag{14}$$

where the elements of π are positive and sum to 1. The steady-state vector π is a time-invariant probability distribution over the two states, which we can also interpret as the long-run probability distribution over the states in the city.

Any Markov chain with a regular transition matrix (defined as a matrix whose elements are all positive for some power of the matrix) is known to converge to a steady state. Since M is a regular Markov matrix, the probability distribution over states converges to the steady-state vector π. By solving equation (14), we obtain π:

$$\pi \equiv \left\{ \begin{array}{c} p^*_{S_1} \\ p^*_{S_2} \end{array} \right\} = \frac{1}{\Pr(S_2|S_1) + \Pr(S_1|S_2)} \left\{ \begin{array}{c} \Pr(S_1|S_2) \\ \Pr(S_2|S_1) \end{array} \right\}. \tag{15}$$

Denote by $Var(r_{j,t})$ the over-time variance in percentile income rank of a neighborhood j. In steady state, the beach's over-time variance in income rank can be written as

$$\begin{aligned} Var(r_{j,t}|j = b) &= \{p^*_{S_1}(r_H)^2 + (1 - p^*_{S_1})(r_L)^2\} - \{p^*_{S_1} r_H + (1 - p^*_{S_1}) r_L\}^2 \\ &= (1 - p^*_{S_1}) \cdot p^*_{S_1} \cdot (r_H - r_L)^2. \end{aligned}$$

Since the average income of the beach takes exactly the opposite value to that of the desert, the over-time variances are equal. Thus, the average rank variance for the city can be written as

$$E(Var(r_{j,t}|j)) = (1 - p^*_{S_1})p^*_{S_1} \cdot (r_H - r_L)^2.$$

This city average variance is maximized when $p^*_{S_1} = 0.5$ and decreases monotonically as $p^*_{S_1}$ moves away from 0.5. Equations (4), (5), and (15) imply that, conditional on $r_H - r_L$, $p^*_{S_1}$ increases from 0.5 with $|\alpha_b - \alpha_d|$. Therefore, $E(Var(r_{j,t}|j))$ decreases with $|\alpha_b - \alpha_d|$. Intuitively, the city's spatial distribution of income experiences the least persistence over time when there is no natural heterogeneity within the city (i.e., the city is in a flat, featureless plain). As the beach's natural advantage increases, the likelihood of churning between states declines, leading to stability and persistence in the spatial distribution of income.

A.2 Proof of Proposition 6

Just as we derive equations (6) and (7), we can calculate expected income percentile rank change for each neighborhoood and initial income:

[Table A1 about here.]

Equations (4) and (5) imply that, as $|\alpha_b - \alpha_d|$ increases, $Pr(S_2|S_1)$ decreases and $Pr(S_1|S_2)$ increases. It follows from Table A1 that, as $|\alpha_b - \alpha_d|$ increases, the gap between beach and desert in expected income change increases regardless of initial income level. In other words, the anchoring effect of natural amenities is stronger in naturally heterogeneous cities.

A.3 Full model

This section presents the full model that allows the city to have more than two neighborhoods and the amenity shocks $\epsilon_{j,t}$ to be correlated over time. The full model differs from the simple model presented in Section 2.1 in the main text in the following ways. First, the city has $J \in \mathbb{N}$ neighborhoods and J unit measure of workers. Second, the aggregate amenity shock $\epsilon_{j,t}$ follows an AR(1) process: $\epsilon_{j,t+1} = \rho\epsilon_{j,t} + \nu_t$ where ν_t is independent and identically distributed. Third, we extend the equilibrium selection rule in Section 2.3 as follows: When there are multiple equilibria, we choose the one that is closest to the selected equilibrium in the previous period, in terms of the Euclidean distance in the vector of average incomes across neighborhoods.

A.3.1 Equilibrium within a period

Lemma 1, which states that higher income workers sort into superior aggregate amenity neighborhoods, holds with the full model, because it is driven solely by workers' preferences.

To precisely describe the sorting with J neighborhoods, we introduce new notation. First, we partition the set of worker incomes $[\underline{\theta}, \overline{\theta}]$ into J intervals $\{\Theta_1, \Theta_2, ..., \Theta_J\}$ so that each group has a unit measure of workers; Θ_1 is the top income group, and Θ_J is the bottom group. $\hat{\theta}_{i,i+1}$ denotes workers who divide group i from group $i+1$; i.e., $\Theta_J \equiv [\underline{\theta}, \hat{\theta}_{J-1,J}]$, $\Theta_{J-1} \equiv [\hat{\theta}_{J-1,J}, \hat{\theta}_{J-2,J-1}]$, ..., $\Theta_1 \equiv [\hat{\theta}_{1,2}, \overline{\theta}]$. Second, we define a neighborhood rank function $\tilde{r}_t : J \to J$ such that $\tilde{r}_t(j)$ is the rank of neighborhood j in terms of aggregate amenities in period t. For example, suppose that neighborhood 1 is the third best neighborhood in terms of aggregate amenities in period 2. Then we have $\tilde{r}_2(1) = 3$. Note that its inverse function $\tilde{r}^{-1}$ maps back to neighborhood index numbers, given aggregate amenity rankings. For example, suppose that the second best neighborhood in period 3 is neighborhood 4. Then we have $\tilde{r}_3^{-1}(2) = 4$.[38]

Since each Θ_j group of workers consumes one unit measure of land and each neighborhood has one unit measure of land, each Θ_j group of workers occupies one and only one neighborhood. Further, since higher-skill workers select into better aggregate amenity neighborhoods, each group Θ_j occupies neighborhood $\tilde{r}_t^{-1}(j)$ in each period t.

We characterized workers' location choices as a function of the distribution of aggregate amenities across neighborhoods. In turn, workers' location choices must generate a distribution of aggregate amenities across neighborhoods that is consistent with their location choices. In other words,

$$a_{\tilde{r}_t^{-1}(j)} + E(\theta|\theta \in \Theta_j) + \varepsilon_{\tilde{r}_t^{-1}(j)} \geq a_{\tilde{r}_t^{-1}(j+1)} + E(\theta|\theta \in \Theta_{j+1}) + \varepsilon_{\tilde{r}_t^{-1}(j+1)}. \tag{16}$$

Proposition 2, which states that an equilibrium exists in each period and there can be multiple equilibria, holds with the full model. First, an equilibrium always exists because condition (16) is satisfied if higher-income workers choose to live in neighborhoods with greater exogenous amenities: $a_j + \varepsilon_{j,t}$. Second, there

[38] Note also that this rank function $\tilde{r}$ differs from the percentile rank income r in three ways. First, $\tilde{r}$ ranks neighborhoods based on their aggregate amenity levels A_j while r on average income (i.e., endogenous amenity level) $E(\theta|j)$. Second, $\tilde{r}$ assigns a lower number for a better aggregate amenity neighborhood, while r assigns a higher number for a better average income neighborhood. Third, $\tilde{r}$ gives an integer rank, while r gives a percentile rank.

can be multiple equilibria. For example, condition (16) is satisfied for any matching pattern between income groups and neighborhoods if exogenous amenities (i.e., $\alpha_j + \epsilon_{j,t}$) are identical across all neighborhoods.

Now we characterize how rents are determined in each period. We normalize rent for the least favored neighborhood to be 0, that is,

$$R_{\tilde{r}_t^{-1}(J)} = 0.$$

For the other neighborhoods, equilibrium rent $R_{\tilde{r}_t^{-1}(j)}$ is recursively determined so that $\hat{\theta}_{j,j+1}$ workers (i.e., workers who divide Θ_j and Θ_{j+1}) are indifferent between neighborhood $\tilde{r}_t^{-1}(j)$ and $\tilde{r}_t^{-1}(j+1)$:

$$A_{\tilde{r}_t^{-1}(j)} \cdot (\hat{\theta}_{j,j+1} - R_{\tilde{r}^{-1}(j),t}) = A_{\tilde{r}_t^{-1}(j+1)} \cdot (\hat{\theta}_{j,j+1} - R_{\tilde{r}_t^{-1}(j+1)}).$$

This equation recursively pins down rent for each neighborhood. Note that neighborhood rents follow the same order as average incomes, as with the simple model.

A.3.2 Equilibrium selection and history dependence

When there are multiple equilibria, we choose the one that is closest to the selected equilibrium in the previous period, in terms of the Euclidean distance in the vector of average incomes (i.e., endogenous amenities) across neighborhoods.

Partly because the number of possible location-choice patterns increases dramatically with that of neighborhoods (i.e., $J!$ with J neighborhoods), we cannot analytically prove Propositions 4, 5, and 6 with the full model. Instead, we use numerical methods to demonstrate that the results are robust with more than two neighborhoods and serially correlated amenity shocks.

For various combinations of parameters, we calculate the equilibrium path for 100,000 periods and test whether the propositions hold. The following list of parameters are used in the simulations. For the number of neighborhoods J, we use 3, 5, and 7 neighborhoods. For the natural amenity distribution across neighborhoods, we use $\xi \times (1, 2, ..., J)$ and vary ξ to be 1, 3, 5, and 10. Note that the variance in natural amenity levels increases as ξ increases. For the average income distributions across neighborhoods, we use $\psi \times (1, 2, ..., J)$ and vary ψ to be 1, 3, 5, and 10. For amenity shocks, we assume that $\epsilon_{j,t}$ follows an AR(1) process $\epsilon_{j,t} = \rho\epsilon_{j,t-1} + \nu_{j,t}$, where $\nu_{j,t}$ follows a Normal distribution $(0, \sigma^2)$. We vary ρ to be 0, 0.2, 0.6, 0.9, 0.95, 0.98, 0.99, and 1 and vary σ to be 1, 3, 5, and 10. ρ determines how much the amenity shocks are correlated over time. Note that the amenity shocks are stationary if ρ is less than 1. σ determines how volatile the shocks are. This grid of parameters generates 1,536 unique combinations of parameters.

We begin with Proposition 4. For each combination of all parameters with $\rho < 1$ (1,344 combinations in total), we obtain a number $J \times 100,000$ of neighborhood-period level simulated data. For each combination with $\rho < 1$, we regress change in percentile rank income of a neighborhood on its percentile rank natural amenity level and its current period percentile rank income. This is the base specification we use in our empirical analysis.

Proposition 4 implies that the coefficient on natural amenity should be positive. Our simulation results confirm this prediction with stationary amenity shocks. With $\rho \leq 0.98$, the coefficients were weakly positive for all 1,152 combinations. With $\rho = 0.99$, only four parameter combinations out of 192 show small negative values. The small number of negative outcomes seem to be driven by numerical errors, as neighborhood shocks become close to a nonstationary unit-root process. With a unit-root process (i.e., $\rho = 1$), our predictions do not hold: 128 of 192 cases show negative values.

Next, we test Proposition 5. We calculate $E(Var(r_{j,t}|j))$ for each parameter set. The Proposition implies that $E(Var(r_{j,t}|j))$ decreases with ξ, and our simulation results show that $E(Var(r_{j,t}|j))$ indeed decreases with a stationary amenity shock.

Finally, we test Proposition 6. The effect of superior natural amenities is captured by the coefficient on percentile rank natural amenity level in the previous regressions used to test Proposition 4. We test if the coefficients tend to increase with ξ. (Recall that natural amenity values are $\xi \times (1, 2, ..., J)$ across neighborhoods. As ξ increases, heterogeneity in natural amenity values increases.) We calculate the mean value of the coefficient estimates for each ξ=1, 3, 5, 10. Each ξ group has 384 parameter sets. The results show that the mean coefficient increases monotonically with ξ.

Table A1: Expected change in income conditioned on natural amenity and initial income

	Initial Income	
Neighborhood	r_H	r_L
Beach	$(r_L - r_H)Pr(S_2\|S_1)$	$(r_H - r_L)Pr(S_1\|S_2)$
Desert	$(r_L - r_H)Pr(S_1\|S_2)$	$(r_H - r_L)Pr(S_2\|S_1)$
Difference	$(r_H - r_L)(Pr(S_1\|S_2) - Pr(S_2\|S_1))$	$(r_H - r_L)(Pr(S_1\|S_2) - Pr(S_2\|S_1))$

B Data appendix

B.1 Figure and tables referenced in footnotes

[Figure B1 about here.]

[Table B1 about here.]

[Table B2 about here.]

B.2 Census data and boundary normalization

We use 2010 census tract data from the American Community Survey (ACS) five-year summary file, via the National Historical Geographic Information System (NHGIS) (Minnesota Population Center, 2011). These data cover the entire geographic extent of the U.S., although we focus on metropolitan (core-based statistical) areas only. The ACS is the annual replacement for the decennial long-form data, and it includes much of the detailed information on population and housing (e.g., income) that is no longer reported in the decennial census. However, the ACS has one important limitation. Because of small annual sample sizes and privacy concerns, these data represent five-year averages of residents and houses located in each tract. Thus, although we refer to these data as coming from the year 2010 throughout the paper, they really represent an average over 2006–2010. Finally, since these data already follow 2010 census tract boundaries, no normalization is required.

Census data for 1970–2000 are from the Geolytics Neighborhood Change Database (NCDB) (Tatian, 2003). These data are already normalized to 2000 (n.b., not 2010) census tract boundaries. The NCDB methodology compares maps of the 2000 census tract and block boundaries with earlier years. Then, 1990 census block information (each tract is composed of many blocks) is used to determine the proportion of people in each historic tract that should be assigned to each overlapping 2000 tract. These proportions are then used as weights to normalize the data to 2000 boundaries.[39]

To normalize the NCDB data to 2010 census tract boundaries, we use the Longitudinal Tract Database (LTDB) (Logan, Xu, and Stults, 2014). The LTDB uses the same block-weighting methodology as the NCDB. Thus, our analysis uses weights defined by 2000 census block populations to normalize all of the Geolytics NCDB data, from 1970 to 2000, to 2010 census tract boundaries. It is important to note that in 1980 and earlier, the entire geographic extent of the U.S. was not completely organized into tracts, and missing data problems are more severe for earlier years. However, since we focus on metropolitan areas, data quality is quite good as early as 1970. (We also drop tract observations in years when their respective metropolitan area is incompletely tracted. See more on sample selection below.)

For census years 1910–1960, we use decennial census information from the NHGIS. The 1940, 1950, and 1960 NHGIS extracts are collectively known as the Bogue files (2000a, 2000b, and 2000c), and they are also available from the Inter-university Consortium for Political and Social Research. These files contain tract information for selected cities and metropolitan areas. The 1910, 1920, and 1930 NHGIS extracts are known as the Beveridge files. Note that data availability is sparse, especially before 1950. Even for cities that are completely tracted, sometimes the data do not contain complete information on population, housing, or income. (For example, in 1910, tract information on household income is only available for New York City; in 1920, such information is only available for New York City and Chicago. Ten metropolitan areas have valid data in 1930, and 43 metropolitan areas have valid data in 1940.) We normalize these data to 2010 census tract boundaries ourselves using NHGIS map layers. For each decade, we compare historical tract boundaries with 2010 census tract boundaries. Since subtract or block information on population is unavailable for these historical years, we are unable to exactly follow the NCDB and LTDB methodologies of constructing weights using block populations. Instead, we normalize using a simple apportionment based on land area.

Finally, we draw 1880 census information from the Integrated Public Use Microdata Series (IPUMS) (Ruggles et al., 2010). We use both the 100% census and the 10% population sample; the 10% sample includes information on literacy, while the 100% census does not. The IPUMS includes data on each person's place of residence, via the enumeration district variable. Enumeration districts were areas assigned to census enumerators to gather data, and they are comparable in population size with modern-day census tracts. (In fact, on average, they are slightly smaller than modern-day census tracts.) We use enumeration districts

[39] We make a small adjustment to the 1980 Geolytics NCDB. The 1980 census prized identification of "places" (e.g., towns, villages, boroughs) over tracts when confidentiality restrictions were binding. The NCDB propagates this censoring in its normalization procedure, even if the proportion of households in the tract with suppressed income data is negligible. We restore this income information from the original 1980 census as long as the proportion of censored households in a census tract is less than 20%.

to normalize the historical 1880 data to 2010 census tract boundaries. First, we obtain maps on historical enumeration district boundaries from the Urban Transition Historical GIS Project (UTHGIS) (Logan et al., 2011). Maps are available for 32 present-day metropolitan areas (totaling 29 consolidated metropolitan areas). (Note that our ability to match households to neighborhoods is limited both by availability of household data—the 1890 census was destroyed by fire—and the availability of maps showing the spatial extent of historical census tracts or enumeration districts.) Second, using the same procedure as for 1910–1960, we compare historical enumeration district boundaries with 2010 census tract boundaries. We apportion to 2010 census tract boundaries using land area.

B.3 Sample selection

We exclude a number of tract observations according to the following criteria. We drop tracts in Alaska, Hawaii, and Puerto Rico. We exclude tracts with zero land area (these are typically "at sea" populations, i.e., personnel on ships) or zero population (e.g., airports or zones otherwise reserved for nonresidential uses).

We do not consider tracts outside of metropolitan areas defined in 2009. One problem with nonmetropolitan tracts is that many of them are not available before 1990, the first year that the U.S. was fully organized into census tracts. Another problem with rural tracts is the difficulty in grouping these tracts into units that share common labor, housing, product, and input markets. (Exceptions are the core-based statistical areas called micropolitan areas. However, many of these micropolitan areas feature a very small number of tracts, making them unsuitable for our analysis. The very small number of tracts means that the entry of even one new neighborhood can elicit a volatile response in within-micropolitan area rankings.)

We drop tracts in particular years that are clearly nonurban. This restriction is more salient in historical years, when tracts or enumeration districts on the urban fringe were not subject to urban land uses. We classify tracts as nonurban if (i) the entire tract population is classified by the census as "rural" or (ii) population density is less than 32 people per square mile, or one person per 20 acres. (Lowering this threshold to one person per 40–160 acres affects the number of excluded tracts minimally. Population densities of less than 32 people per square mile are already well short of standard definitions of urban population densities.) We reason that while these tracts are within counties that contain urban uses, at the time of observation, they are likely to be outside of metropolitan areas and urban household location decisions. In this way, we also address concerns about changing metropolitan area boundaries over time.

We exclude tracts where our normalization procedure is likely to be poor. In some cases, especially for early census years and tracts on the urban fringe, historical tracts cover only a portion of 2010 census tract areas. This is more likely to be the case when historical city boundaries are much smaller than present-day extents. When historical tracts cover less than 50% of the land area of the present-day tract, we exclude these data from our analysis.

We also eliminate tract observations that disappear from one year to the next. This problem is partly mechanical; we cannot compute income changes for a tract that does not appear in the next period. It also is mostly limited to the transition between the 1880 UTHGIS data and the subsequent NHGIS data. The reason this problem arises is because the UTHGIS maps, which we use for our normalization procedure, typically cover entire counties, whereas the NHGIS data and maps used in the early 20th century are confined mostly to city boundaries. Thus, many of the UTHGIS tracts outside city boundaries are dropped anyway because they are nonurban (as previously noted), but to avoid the problem of contracting metropolitan boundaries, we exclude the remaining earlier tracts that do not appear in subsequent years.

A consequence of the unbalanced nature of the data is that forward lags vary by metropolitan area and year. For example, after 1880, it is only 30 years until our next observation of New York neighborhoods (in 1910), but it is 70 years until our next observation of Omaha neighborhoods (in 1950). Out of 1,684 metropolitan area-year groups in our data, 1,342 follow the standard 10-year gap between census year observations. As a result, the actual number of neighborhoods used in regressions varies according to whether the specification requires balancing across two subsequent census years or balancing over a large number of years. In addition, some variables, such as flood hazard or average housing unit age, are unavailable in some years, further affecting sample selection.

B.4 Natural amenity data

We spatially match our consistent-boundary neighborhoods to a number of natural and persistent geographic features.

Water features—coastlines, lakes, and rivers. We use data on water features from the National Oceanic and Atmospheric Administration's (2012) Coastal Geospatial Data Project. These data consist of high-resolution maps covering (i) coastlines (including those of the Atlantic, Pacific, Gulf of Mexico, and Great Lakes), (ii) other lakes, and (iii) major rivers. For each 2010 census tract, we separately calculate the distance to each of the nearest water features (ocean, lake, river) from the centroid of the tract.

Elevation and slope. We use the elevation map included in the Esri 8 package. These data have a 90-meter resolution. In ArcGIS, we use the slope geoprocessing tool to generate a slope map. Then we use the zonal statistics tool to calculate average slope in each 2010 census tract.

Floodplains. The Federal Emergency Management Agency (FEMA, 2012) publishes National Flood Hazard Layer (NFHL) maps covering much of the U.S. The NFHL maps show areas subject to FEMA's flood zone designations. We assign to tracts either a high-risk or low-risk indicator. High risk means that an area has at least a 1% annual chance of flooding (a 26% chance of flooding over a 30-year period), as determined by FEMA. Note that flood maps are unavailable for some metropolitan areas. In our data, 261 metropolitan areas have valid flood zone information.

Temperature and precipitation. We match tracts to temperature and rainfall data available from the PRISM Climate Group (2004) at Oregon State University. These data are 1971–2000 averages, collected at thousands of weather monitoring stations and processed at a spatial resolution of 30 arcseconds for the entire spatial extent of the U.S., of annual precipitation, July maximum temperature, and January minimum temperature.

B.5 Other data

City centers. Data on principal city center locations for 293 metropolitan areas were generously provided to us by Dan Hartley. Fee and Hartley (2013) identify the latitude and longitude of city centers by taking the spatial centroid of the group of census tracts listed in the 1982 Census of Retail Trade for the central city of the metropolitan area. Metropolitan areas not in the 1982 Census of Retail Trade use the latitude and longitude for central cities using ArcGIS's 10.0 North American Geocoding Service.

Seaports. Data on seaport locations are from the *World Port Index*, 23rd edition, published by the National Geospatial-Intelligence Agency (2014).

Land use regulation. The Wharton Residential Land Use Regulatory Index is from Gyourko, Saiz, and Summers (2008).

Neighborhood names. Information on neighborhood names comes from the U.S. Geographic Names Information System (GNIS), maintained by the U.S. Geological Survey (USGS), and the U.S. Board on Geographic Names, which maintains uniform usage of geographic names in the federal government. We use named populated places, which represent "named communities with a permanent human population" (USGS, 2014). These communities range from rural clustered buildings to metropolitan areas and include housing subdivisions, trailer parks, and neighborhoods. These names are assigned point coordinates by the USGS. with no defined boundaries. In our database of consistent-boundary neighborhoods of U.S. cities, 8,983 neighborhoods (out of 60,758) have no named populated place. This could be because a neighborhood has no name or (more likely) the neighborhood's name is also associated with a nearby census tract that happens to have been assigned the point coordinate by the USGS. The remaining tracts have one or more associated place names.

Note that the populated place names database excludes names of natural features, and it includes both incorporated and unincorporated place names.

B.6 Neighborhood percentile ranks

Figure B2 shows the evolution of several New York neighborhoods over our sample period. Recall that each neighborhood corresponds to data normalized to 2010 census tract boundaries. The solid lines show the relative rankings of three neighborhoods—tracts corresponding to the Upper East Side, East Harlem, and Tribeca. (Levittown was unpopulated in 1880 so a corresponding solid line does not appear in the figure.) An interesting feature of this graph is variation in income dynamics across neighborhoods. For example, the Upper East Side has remained a high-income neighborhood throughout our sample period. East Harlem, which was a relatively high-income neighborhood in 1880, experienced decline and has been a low-income neighborhood since 1910. Tribeca saw a large increase in average household income in the 1980s.

[Figure B2 about here.]

The dotted lines show the relative rankings of these three neighborhoods after 1960 and the relative ranking of a fourth neighborhood, Levittown, which first appeared in that census year. In comparing the solid with the dotted lines, note that we have changed the universe used to compute neighborhood ranks from 1880 to 1960 neighborhoods, but the dynamic patterns for the extant three neighborhoods remain qualitatively similar.

In our sample, most neighborhoods experience changes in percentile rank that are close to zero—that is, neighborhood income ranks are largely persistent over time, especially over the 10-year changes that are predominant in our sample. Few neighborhoods experience dramatic increases or declines in rank. The distribution of percentile rank changes has a mean zero and standard deviation of 0.164.

B.7 Mean reversion

We begin by noting the overall relationship between neighborhood change and initial income. In short, neighborhood status tends to mean revert. Figure B3 shows local polynomial smoothing of sample 10-year changes in neighborhood income percentile rank, $\Delta r_{i,t}$, versus initial ranks $r_{i,t}$. (The right axis smooths log changes in average neighborhood income, net of the metropolitan area–year mean.) Neighborhoods that are initially highly ranked tend to decline, and low-ranked neighborhoods tend to improve.

[Figure B3 about here.]

Several features of Figure B3 are noteworthy. First, despite using a nonlinear technique, the pattern of mean reversion is close to linear in initial rank, especially in the middle of the income distribution. In our anchoring regressions, we condition linearly on initial rank. (In robustness checks, we allow differences by decile in initial rank.)

Second, by construction, the sum of changes in income percentile ranks is zero, because one neighborhood's improvements in rank are offset by other neighborhoods' declines. An important implication is that changes for a bottom-ranked neighborhood are restricted to the interval $(0, 1)$, and changes for a top-ranked neighborhood are restricted to $(-1, 0)$. Despite this, mean reversion appears to be driven not by neighborhoods at the top and bottom of the income distribution (as would be expected if mean reversion were purely mechanical) but by neighborhoods near the middle of the income distribution (whose changes in rank are less restricted). The fact that the negative correlation between changes in income and initial income is weaker for neighborhoods with initial rank below the 20th percentile and above the 80th percentile suggests that mechanical effects contribute little to the overall pattern.[40]

[40] Weak mean reversion in the tails of the initial income distribution, despite mechanical censoring effects, may lend further credence to the role of natural amenities: The existence of very strong amenities and disamenities may lead to neighborhoods that are persistently very poor or very rich. See Section 5 in the main text.

Finally, there is mean reversion even in nominal incomes, showing that this pattern is not exclusively driven by our use of percentile ranks (Figure B3, right axis). The expected 10-year sample change in average household income for a neighborhood at the bottom of the income distribution, relative to the average neighborhood in the metro–year, is an increase of about 2%. In contrast, the expected relative change for a top neighborhood is about a 6% decline.

B.8 Robustness to proximity thresholds

In Figure B4, we show that the effect of proximity to an ocean or a Great Lake is consistent for varying definitions of proximity. The gray line connects estimates from 100 separate regressions of neighborhood change on proximity, varying the proximity indicator. (As we move to the right along the horizontal axis, our indicator variable classifies more neighborhoods as being "close" to the natural amenity and fewer neighborhoods as being "far" from the natural amenity.) The black line displays regression estimates when we use only natural features near top-income neighborhoods, as in Table 2, column 6 in the main text. (The intersection of these lines with the vertical dotted line are the estimates from our baseline estimates using a 500-meter definition, shown in Table 2.) Recall that we expect these features to more likely be positive, versus negative, amenities. As expected, the results using this variable are always stronger than the results using oceans and Great Lakes unconditioned on initial income.

[Figure B4 about here.]

Figure B4 also shows the same results for lakes, rivers, and hills. These results show consistent patterns. The important feature of this figure is that it shows that conditioning rivers on their proximity to high-income neighborhoods improves the estimated effect of (positive-amenity) rivers on neighborhood change. This is consistent with the view that, on average, rivers in our sample are a disamenity for households.

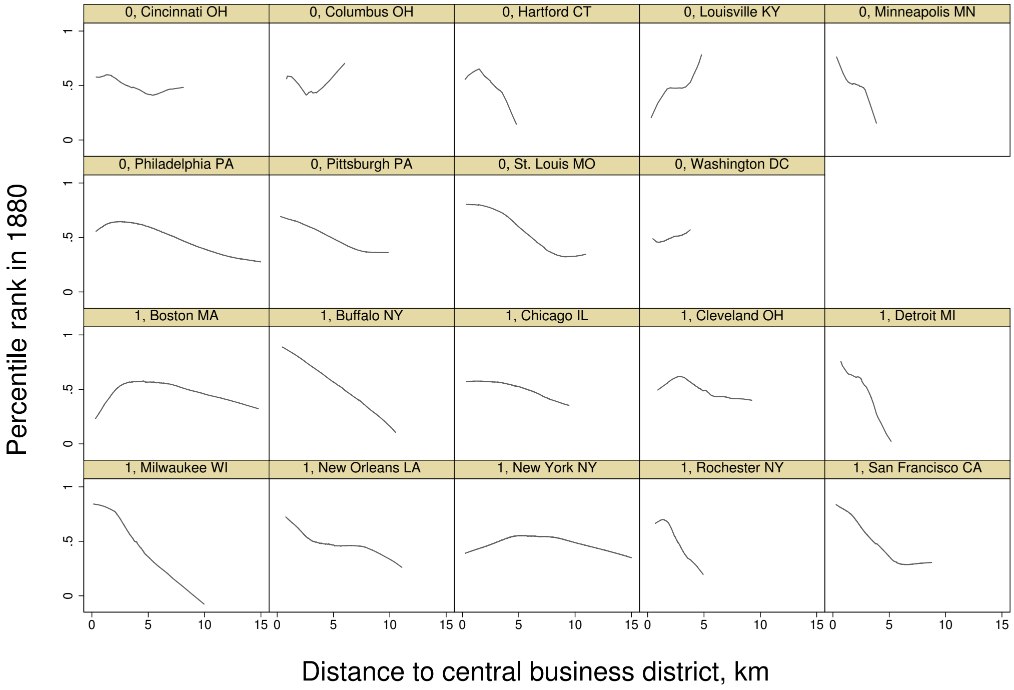

Figure B1: Income and residential location for large cities, 1880

These plots show the pattern of neighborhood percentile rank on the vertical axis versus neighborhood distance to the city center (up to 15km) on the horizontal axis for the 19 largest cities in our 1880 sample. Ten metropolitan areas with 20 or fewer neighborhoods in 1880 are not shown. Cities are organized by coastal status (0=interior, 1=coastal) and then alphabetically. The plotted lines are results from lowess smoothing with bandwidth 0.9.

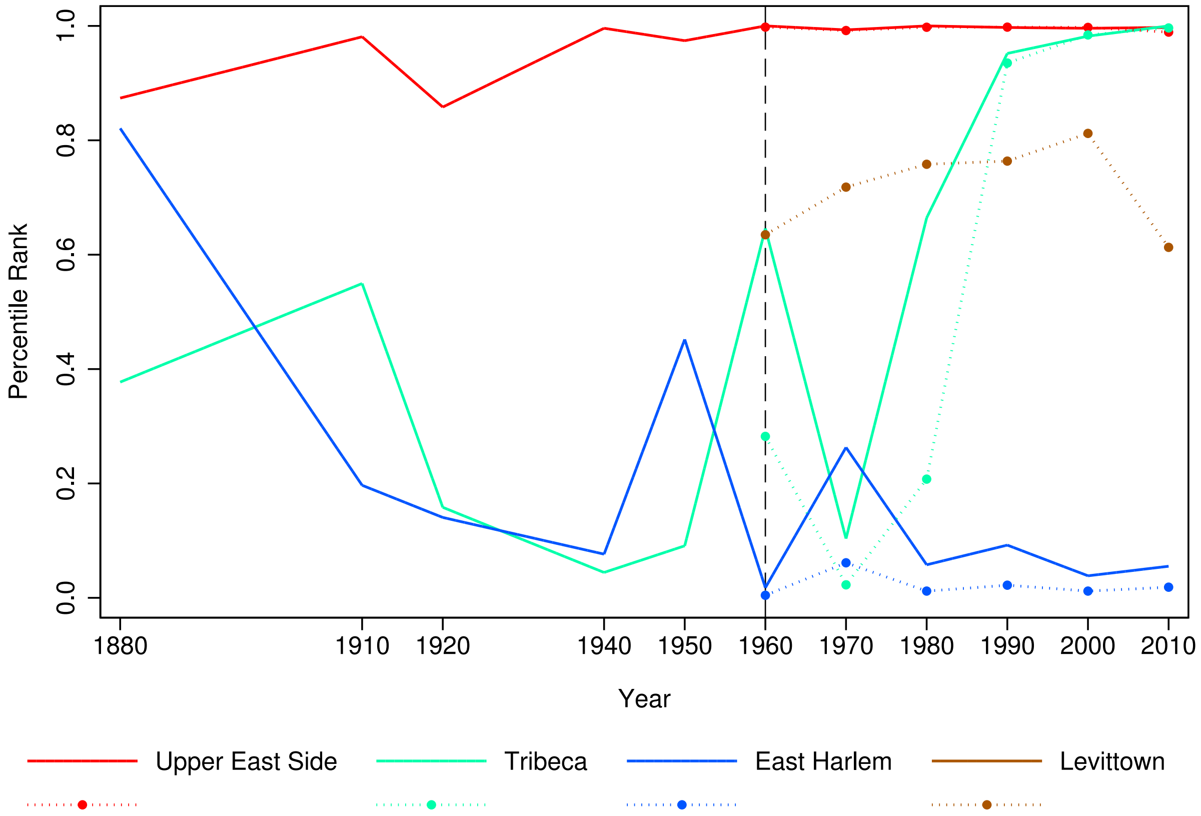

Figure B2: Selected New York neighborhood rankings over time

This plot shows percentile ranks of 3 neighborhoods among 1880 neighborhoods and 4 neighborhoods among 1960 neighborhoods. Solid lines connect neighborhood percentile ranks among 1880 neighborhoods. Dotted lines connect neighborhood percentile ranks among 1960 neighborhoods.

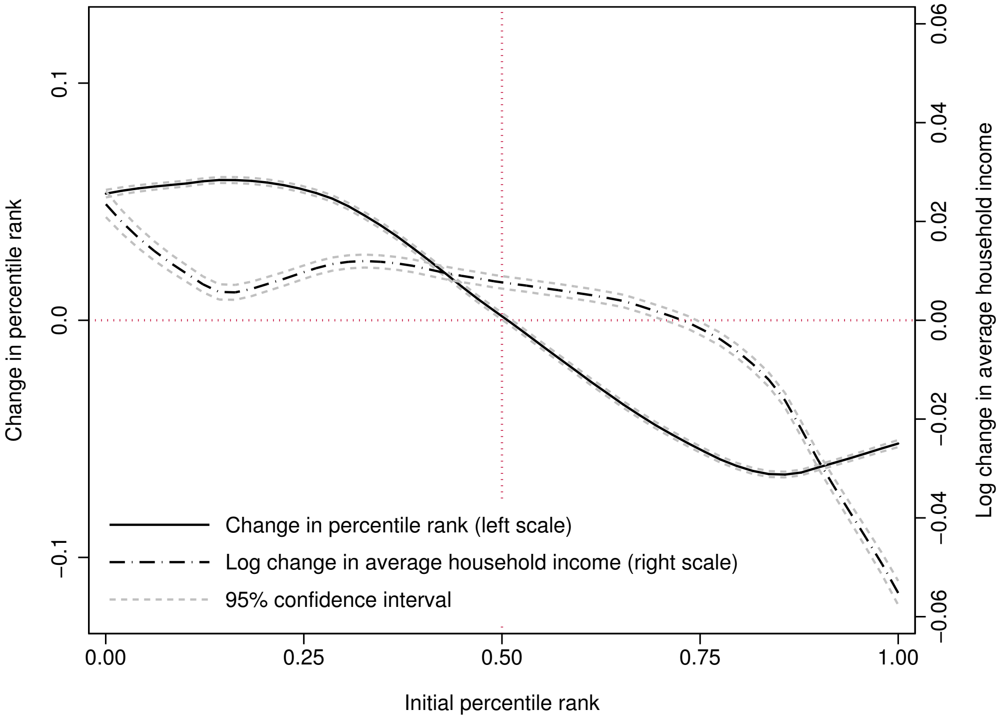

Figure B3: Mean reversion in neighborhood percentile rank by income

This plot shows kernel-weighted local polynomial smooths using the Epanechnikov kernel and rule-of-thumb bandwidth. Neighborhood log change in average household income (right scale) is normalized by metropolitan area–year mean.

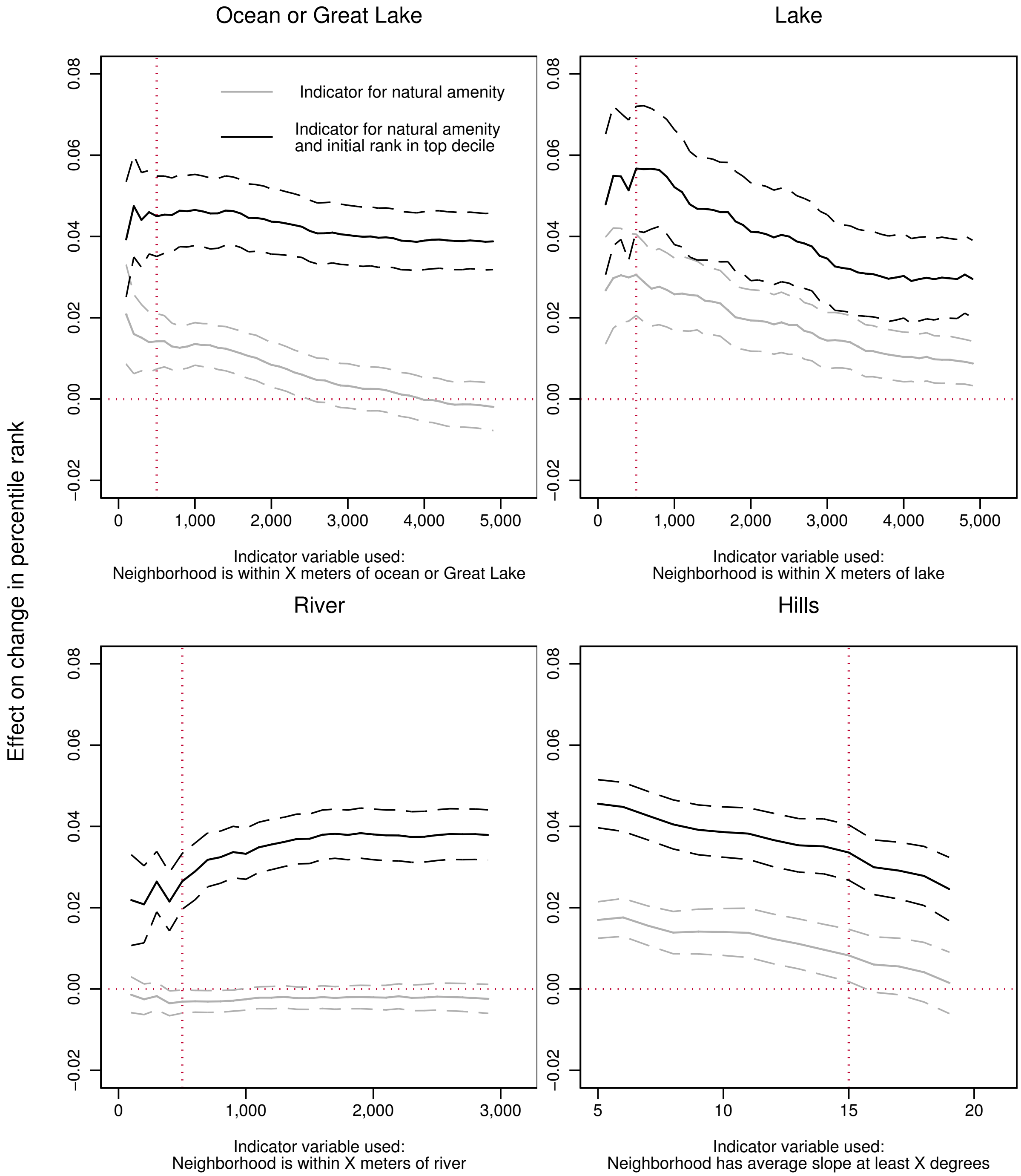

Figure B4: Robustness to indicator variable thresholds

These plots show the conditional effect of natural features on neighborhood change for varying indicator definitions of proximity to natural features. Each connected point is from a separate regression. The gray solid line connects estimates of the effect of natural amenities, varying the definition of the natural amenity on the horizontal axis, as in Table 3, Panel B. The black solid line connects estimates of the effect of natural amenities conditioned on initial income $r_{i,t} > 0.9$, varying the definition of natural amenity on the horizontal axis, as in Table 3, Panel C. Dashed lines show 95% confidence intervals. Vertical dotted line shows baseline definition used in Table 3.

Table B1: Summary statistics

	μ	(σ)
A. Consistent-boundary neighborhoods		
Δr, 10-year forward change in percentile rank, 1910–2000	0.00	(0.16)
$\sigma(r_{i[m]}) \times 100$, Std. dev. in 1960–2010 percentile rank	12.9	(7.9)
Population, 2010	4,283	(1,912)
Land area (km^2)	27.5	(73.5)
Persons per square km, 1880	5,940	(12,406)
Persons per square km, 1960	2,901	(6,159)
Persons per square km, 2010	2,335	(4,807)
Distance from centroid to nearest seaport (km), 2010	163	(240)
Distance from centroid to city center (km), 2010	29.9	(27.2)
Mean age of housing units (years), 2010	37.3	(14.1)
B. Share of 2010 neighborhoods		
... with centroid within 500m of ocean or Great Lake	0.047	
... with centroid within 500m of lake (ex. Great Lakes)	0.007	
... with centroid within 500m of major river	0.098	
... with average slope greater than 15 degrees	0.069	
... with moderate temperatures*	0.091	
... ... and less than 800mm average annual precipitation	0.063	
... with less than 1% average annual flood risk†	0.586	
C. Metropolitan areas		
$\overline{\sigma(r_m)} \times 100$, Mean std. dev. in 1960–2010 percentile rank	13.1	(2.8)

*–Average January minimum temperatures between 0 and 18 degrees Celsius and average July maximum temperatures between 10 and 30 degrees Celsius. †–Flood information available for 27,133 neighborhoods.

Table B2: Coastal and interior cities

Coastal	Interior
Boston, MA	Albany, NY
Buffalo, NY	Atlanta, GA
Charleston, SC	Cincinnati, OH
Chicago, IL	Columbus, OH
Cleveland, OH	Hartford, CT
Detroit, MI	Indianapolis, IN
Milwaukee, WI	Kansas City, MO
Mobile, AL	Louisville, KY
New Orleans, LA	Memphis, TN
New York, NY	Minneapolis, MN
Rochester, NY	Nashville, TN
San Francisco, CA	Omaha, NE
	Philadelphia, PA
	Richmond, VA
	Pittsburgh, PA
	St. Louis, MO
	Washington, DC

These are the principal cities for consolidated metropolitan areas shown in Figure 7 in the main text.

C Monte Carlo simulations

We analyze the biases in estimates reported in Section 4 of the anchoring effect of natural amenities, using Monte Carlo simulations. There are two identification issues. First, we observe only whether a neighborhood is near natural features (e.g., rivers, oceans), but we do not know whether these natural features are truly amenable. For example, a polluted river can be disamenity. This is a type of measurement error. Second, our anchoring regression includes a lagged dependent variable and may include unobserved neighborhood factors, as in equation (9). Unfortunately, we cannot employ the time-differenced approaches proposed by Arellano and Bond (1991) or Caselli, Esquivel, and Lefort (1996) because differencing would eliminate our time-invariant variable of interest. Therefore, in this Appendix, we sign these biases with Monte Carlo simulations. Our simulations show that the regression estimates reported in the paper are lower bounds for the true anchoring effect of natural amenities.

Section C.1 describes a data-generating process (DGP) for our simulated data. Section C.2 examines the role of measurement error. Section C.3 examines the role of unobserved neighborhood factors. Section C.4 examines both issues together.

C.1 Data-generating process

For exposition, we use a set of parameters chosen to roughly match data. (Our results are robust to a wide range of other parameter values.[41]) We assume 50,000 neighborhoods, of which 10% are "beaches" ($\mathbf{1}^*(a_i) = 1$). Since not all coastal neighborhoods are amenable, we randomly assign only a half of these beach neighborhoods to have positive natural amenity value ($\mathbf{1}(a_i) = 1$).

We generate $r_{i,t}$, neighborhood incomes over time, using equation (9) (repeated here for convenience):

$$\Delta r_{i,t} = \beta_0 + \beta_1 \mathbf{1}(a_i) + \beta_2 r_{i,t} + u_i + \epsilon_{i,t} \tag{9}$$

where we use $\beta_0 = 0$, $\beta_1 = .1$, $\beta_2 = -.1$ and $\epsilon_{i,t} \sim$ i.i.d. $\mathcal{N}(0, .05^2)$. Recall that β_1 is true natural amenity effect. Proposition 4 predicts $\beta_1 > 0$.

The u_i are unobservable neighborhood characteristics, where $u_i \sim$ i.i.d. $\mathcal{N}(0, \sigma_u^2)$. The standard deviation of u_i, σ_u, captures the importance of unobservable characteristics in determining neighborhood income growth $\Delta r_{i,t}$. It turns out that σ_u plays an important role in the bias coming from the lagged endogenous regressors. In Section C.3, we vary σ_u to see the effect of u_i on β_1 estimate.

We generate the data for 20 periods and keep only the last five periods. This is to allow time to arrive at the steady state.[42] Note in equation (9) that we use true natural *amenity* dummy $\mathbf{1}(a_i)$, not natural *feature* dummy $\mathbf{1}^*(a_i)$, in the data-generating process. This is because households in the real world can observe whether a natural feature is an amenity or not.

C.2 Measurement error

This section examines the issue of the measurement error when we estimate equation (9) using natural feature indicator $\mathbf{1}^*(a_i)$ instead of natural amenity $\mathbf{1}(a_i)$. To focus on the measurement error issue, we use the DGP with $\sigma_u = 0$ which, as we show in Section C.3, removes the bias caused by the lagged endogenous regresssor.

Our identification strategy in Table 2, column (5) is to condition observed natural features on historical income. For each set of simulated data, we estimate the following model:

$$\Delta r_{i,t} = \beta_0 + \beta_1 \mathbf{1}^*(a_i) \cdot \mathbf{1}(\text{PRank}(r_{i,t}) \geq \tilde{\theta_H})) + \beta_2 r_{i,t} + \epsilon_{i,t}$$

[41] Replication files are available at the authors' websites. Readers may try other parameter values with this program.

[42] Since we have 50,000 neighborhoods, the simulated data set has 250,000 observations. This is roughly similar to the number of observations used in Table 2 (291,321 to 298,776 across specifications.)

where $\mathbf{1}(\text{PRank}(r_{i,t}) \geq \tilde{\theta_H})$ is an indicator variable for a historically high-income neighborhood with a percentile rank of income greater than or equal to $\tilde{\theta_H}$. Note that when we use $\tilde{\theta_H} = 0$ and thus $\mathbf{1}(\text{PRank}(r_{i,t}) \geq \tilde{\theta_H})) = 1$ for all $r \in [0, 1]$, the regression model becomes our base regression in equation (8) of the paper that estimates the natural amenity effect without conditioning on historical income.

We vary the cutoff $\tilde{\theta_H}$ and report the mean β_1 estimates. We also report the mean of $\rho_{\text{AF}}|\text{H}$, the correlation between observed natural amenity $\mathbf{1}(a_i)$ and measured natural features $\mathbf{1}^*(a_i)$ among the top income neighborhoods H whose percentile rank incomes are greater than $\tilde{\theta_H}$. This correlation illustrates how well the indicator for a natural feature measures the true natural amenity. The following table shows the means of $\hat{\beta}_1$ and $\rho_{\text{AF}}|\text{H}$ with their standard errors in parentheses, from 1,000 trials:

$\tilde{\theta_H}$	$\hat{\beta}_1$	$\rho_{\text{AF}}\|\text{H}$
0	0.032 (0.001)	0.698 (0.007)
0.5	0.049 (0.001)	0.809 (0.006)
0.7	0.064 (0.001)	0.872 (0.006)
0.9	0.087 (0.001)	0.951 (0.004)

The results are consistent with our predictions. As the cutoff for historical income ($\hat{\theta_H}$) increases, the estimate $\hat{\beta}_1$ increases toward the true value of 0.1. The correlation $\rho_{\text{AF}}|\text{H}$ also increases with the cut-off, and this suggests that the natural feature is a better indicator for natural amenities after conditioning on historical income.

C.3 Lagged endogenous variables as regressors

We pick σ_u and generate the data as described in Section C.1. With each resulting data set, we estimate the following model by OLS:

$$\Delta r_{i,t} = \beta_0 + \beta_1 \mathbf{1}(a_i) + \beta_2 r_{i,t} + \epsilon_{i,t}.$$

This regression model is the same as the base regression in equation (8), except that it uses true amenity $\mathbf{1}(a_i)$ rather than natural feature $\mathbf{1}^*(a_i)$. This removes the measurement error issue regarding natural amenity and allows us to focus on the roles of the lagged dependent variable and unobserved neighborhood factors.

We vary the standard deviation of u_i, σ_u, from 0 to 0.02, and, for each value, we repeat the entire Monte Carlo exercise 1,000 times and report the mean of the β_1 estimates and its standard error:

σ_u	$\hat{\beta}_1$
0	0.1 (0.001)
0.01	0.069 (0.001)
0.02	0.036 (0.001)

With $\sigma_u = 0$, the β_1 estimate is virtually equal to its true value 0.1. Intuitively, when there are no unobserved time-invariant neighborhood factors, the OLS estimator is consistent even when including the lagged dependent variable. However, as σ_u increases from 0, the estimated $\hat{\beta}_1$ decreases away from the true value.

C.4 Two issues combined

Now we combine the two effects together. The following table shows mean estimates $\hat{\beta}_1$ and standard errors from 1,000 trials when both σ_{u_i} and $\tilde{\theta_H}$ vary.

		$\hat{\beta}_1$	
	$\sigma_{u_i} = 0.00$	0.01	0.02
$\tilde{\theta_H} = 0.5$	0.049 (0.001)	0.037 (0.001)	0.021 (0.001)
0.7	0.064 (0.001)	0.047 (0.001)	0.026 (0.001)
0.9	0.088 (0.001)	0.062 (0.001)	0.033 (0.001)

The two patterns we observed in the previous sections are robust. Either increasing σ_{u_i} with $\hat{\theta_H}$ fixed or decreasing $\hat{\theta_H}$ with σ_{u_i} fixed decreases the estimate $\hat{\beta}_1$ away from its true value of 0.1. Moreover, the estimates are always lower than the true effect, and this suggests that the true anchoring effect of natural amenities is bounded from below by estimates from the regressions we report in the paper.