Amenities in Quantitative Spatial Models*

Amanda Ang†
Daniel M. Angel‡
Andrii Parkhomenko§

July 31, 2024

Abstract

An indispensable feature of quantitative spatial models (QSM) are structural residuals that fit the distribution of residents across locations to the data. While they are often interpreted as amenities, do they actually represent observed amenities? We collect data on 41 amenities in the Los Angeles County, and then build a QSM to study the relationship between the structural residuals and observed amenities. We find that 45% of the variation in the residuals is explained by observed amenities. This suggests that one should be judicious when interpreting these residuals as amenities. We also find that 14 percentage points of the explained variation is accounted for by natural amenities and 31 by man-made amenities. This supports modeling amenities as endogenous.

Key Words: quantitative spatial model, QSM, amenities, structural residuals.
JEL Codes: R10, R15, R23.

*First draft: July 31, 2024.
†University of Southern California, Department of Economics: qang@usc.edu.
‡University of Southern California, Department of Economics: aqmateo@usc.edu.
§University of Southern California, Marshall School of Business: parkhome@marshall.usc.edu.

1 Introduction

Just as physicists rely on dark matter to account for unexplained mass in the universe, urban economists rely on structural residuals, often interpreted as amenities, to account for unexplained variation in residents across locations. And similar to dark matter, the exact nature of these residuals is not well understood: do they represent observed amenities or something else entirely? The goal of this paper is to examine the nature of these residuals in quantitative spatial models (QSM) and provide suggestions on how to model them.

We collect data on 41 amenities in the Los Angeles County from publicly-available sources. Our amenity variables characterize many important local features that households take into account when making location choices: natural amenities, pollution, traffic, crime, built environment, school quality, and consumption amenities. We aggregate our data at the Census tract level and show that there is large within-county variation in amenities. Then we build a canonical QSM of the Los Angeles Commuting Zone (CZ). In the model, households commute between residence and work, and choose residences depending on their access to jobs, housing cost, as well as amenities. As is standard in QSMs, we infer these amenities as structural residuals that fit the distribution of residents in the model to the observed distribution across Census tracts. Then, we compare these model-implied amenities to the amenities in the data.

Our analysis suggests that observed amenities account for 45% of the variation in model-implied amenities. In other words, over one-half of the variation in the structural residuals cannot be accounted for by an extensive array of amenities that we observe. What we often label as "amenities" in QSMs is largely unexplained by actual amenities.

We then conduct an R^2 decomposition to understand which of the observed amenities matter more for explaining model amenities. We find that out of the 45% of the variation that the data can account for, 14 percentage points come from natural amenities, such as weather and ruggedness, and 31 percentage points come from man-made amenities. Among the latter, consumption amenities and school quality are particularly important.

One possible explanation for the seemingly low R^2 is that in the model individual valuations of amenities are identical. In practice, individuals differ in their valuation of amenities: for example, families with children may prefer to live in a good school district. To address this issue, we extend our QSM to have multiple types and type-specific local residential amenities. Our types differ by education, income, age, gender, and race. We find that for some types, such as college graduates and high-income individuals, observed amenities account for over 50% of model-implied amenities specific to the type. For some other types, such as Hispanic or middle-income individuals, the R^2 is less than 30%. There is also substantial heterogeneity in the relative importance of different types of amenities. For instance, the variation in school quality accounts for a larger share of model-implied amenities for college graduates, and white and Asian individuals.

Our results inform how amenities should be modeled in QSMs and offer two main conclusions. First, because even an extensive battery of 41 amenity variables explains less than one-half of the variation in model-implied amenities, researchers should be cautious when interpreting the structural residuals of QSMs as actual amenities. Second, because man-made amenities account for more variation in model-based amenities than natural amenities, researchers should give preference to models that specify amenities as an equilibrium variable that depends on local population density and composition, and not an exogenous parameter.

This paper contributes to the vast literature that uses QSMs for evaluating spatial effects of shocks and policies. Examples include Ahlfeldt et al. (2015), Heblich et al. (2020), Severen (2021), Allen and Arkolakis (2022), Tsivanidis (2023), Zárate (2024), Chen et al. (2024), and many others. The QSMs and their applications are reviewed in Redding and Rossi-Hansberg (2017) and Redding (2023). We extend this literature by shedding light on the relationship between residential structural residuals in QSMs and the data on amenities.

This paper also contributes to the literature on the role of amenities for the location choices within cities (Almagro and Domínguez-Iino, 2024; Couture et al., 2024). While the empirical analysis in Diamond (2016) showed which types of amenities are important for sorting of college and non-college graduates *across* cities, our paper examines which amenities are important for sorting *within* cities.

The remainder of the paper is organized as follows. Section 2 describes the data on amenities, as well as the data used to build the QSM of Los Angeles. Section 3 describes the model. Section 4 analyzes the relationship between model-based and observed amenities, while Section 5 concludes.

2 Data

2.1 Data on Amenities

We collected data on a wide array of amenities in the Los Angeles County at the Census tract level. Our analysis focuses on six groups of amenities: (1) natural, (2) pollution, heat, and traffic, (3) crime, (4) built environment, (5) school quality, and (6) consumption amenities. The first group includes natural phenomena, such as precipitation or radiation, that are unlikely to be affected by local variation in human activity within Los Angeles. We label them as "natural" amenities. Other amenities are those that arise from human activity and we label them as "man-made" amenities. Table 1 summarizes the amenities that we use on our analysis. We note that all the data we use to construct amenity variables is publicly-available.

Natural amenities. To capture differences in microclimate in Los Angeles, we collect the data on precipitation, solar radiation and average temperature, separately for January, April, July, and October, from the Oregon State University PRISM dataset.[1] We also use a measure of ruggedness from Nunn and Puga (2012). The ruggedness index captures differences in elevation within a given unit of surface area.[2] For example, locations with greater ruggedness may offer better views and, thus, be more desirable.

Pollution, heat, and traffic. Pollution data was sourced from The California Environmental Protection Agency's Office of Environmental Health Hazard Assessment (OEHHA), which compiles a wide range of environmental indices for every census tract in the state, as documented in its CalEnviroScreen annual report. We use the pollution score for 2018 as our primary indicator of environmental amenities. This score measures the exposure to different climate factors such as air quality, drinking water contamination, pesticide use, and toxins from facilities.

Heat exposure is measured using the urban heat island index produced by the California Environmental Protection Agency.[3] The index measures the difference in temperatures between an urban Census tract and a nearby upwind rural tract. Daytime temperatures can be 1–6°C higher in urban areas and up to 22°C higher at night because of heat radiating from buildings and pavement.

We also take data on traffic from CalEnviroScreen. In particular, we use the log of the number of vehicle-kilometers per hour divided by the total road length (kilometers) within a 150-meter radius of the census tract boundary.

Crime. Crime data was obtained from the Los Angeles County Sheriff's Department. The dataset contains incident-level data from 1933. We measure crime in each location by calculating the annual average number of crimes recorded per Census tract from 2017 to 2019. There are 30 different types of offense types in the data, which we classify as either

[1] PRISM Climate Group, Oregon State University: https://prism.oregonstate.edu

[2] Data on Terrain Ruggedness and Other Geographic Characteristics of Countries: https://diegopuga.org/data/rugged/

[3] Understanding the Urban Heat Island Index, CalEPA: https://calepa.ca.gov/climate/urban-heat-island-index-for-california/understanding-the-urban-heat-island-index/

Table 1: Observed Amenities

Variable	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
Natural amenities						
Precipitation (mm)						
January	87	15	26	79	94	157
April	20	3.9	7	18	22	43
July	1.2	0.52	1	1	1	8.5
October	15	3.2	6.2	12	16	33
Solar Radiation						
January	10	0.39	9.8	10	10	11
April	22	0.6	20	22	23	25
July	26	0.8	23	25	26	29
October	15	0.61	14	15	16	18
Temperature (deg. C)						
January	14	1.3	6.8	14	14	15
April	17	0.85	10	16	17	18
July	23	1.6	20	22	24	28
October	20	0.81	15	20	21	22
Ruggedness Index	10	1.1	0	9.8	11	13
Pollution, heat and traffic						
Traffic	6.9	0.95	0	6.6	7.4	8.5
Pollution Score (0-10)	6.2	1.4	0	5.4	7.1	9.9
Urban Heat Island Index	7.6	2.6	0	7.2	9.2	11
Crime (No. of incidents)						
Property Crime	24	78	1	1.5	31	2169
Personal Crime	50	83	1	1	77	932
Built environment (shares)						
Multifamily Residences	0.29	0.3	0	0.014	0.52	1
Mixed Use	0.04	0.11	0	0	0.026	1
Non-determined	0.018	0.1	0	0	0	1
Non-residential	0.34	0.3	0	0.084	0.54	1
Single-family Residences	0.28	0.3	0	0	0.52	1
School Quality (0-100)						
Elementary School	49	26	3.2	28	73	99
Middle School	48	26	0	25	69	100
High School	51	26	8	26	71	99
Consumption Amenities (log)						
Eating Places	2	1	0	1.4	2.7	5.2
Full Restaurants	1.6	1	0	0.69	2.3	4.9
Fast Food Places	0.57	0.71	0	0	1.1	3.2
Coffee Shops	0.19	0.47	0	0	0	2.7
Bars	0.16	0.4	0	0	0	3.1
Furniture Stores	0.47	0.68	0	0	0.69	4.1
Electronics Stores	0.5	0.69	0	0	1.1	3.5
Supplies Stores	0.3	0.53	0	0	0.69	2.8
Clothing Stores	1.1	1	0	0	1.8	6.9
Sports / Music Stores	0.44	0.66	0	0	0.69	4.7
Department Stores	0.058	0.25	0	0	0	2.2
Pet Stores	0.034	0.17	0	0	0	1.4
Used Merchandise Stores	0.082	0.29	0	0	0	2.9
Cosmestics Stores	0.094	0.33	0	0	0	3
Food / Health Stores	0.076	0.26	0	0	0	1.9

a personal or property crime using definitions from the FBI.[4]

Built environment. We use residential zoning data from the Othering & Belonging Institute, published in the Single-Family Zoning in Greater Los Angeles Report. This dataset includes the share of the Census tract area dedicated to each of the following zoning types as of 2021: single-family, multi-family, mixed-use residential, non-residential, and non-developable.

School quality. We collected data on test scores for all public elementary, middle, and high school from SchoolDigger.com, a platform that offers comprehensive information on schools to assist parents in making informed school choice decisions. To measure school quality, we calculate the average test scores between 2016 and 2021 of the elementary, middle, and high school that is the nearest to the centroid of each Census tract.

Consumption amenities. We extracted the data on consumption amenities from openICPSR, a repository of research data that includes The National Neighborhood Data Archive (NaNDA) administered by the Social Environment and Health program at the University of Michigan Institute for Social Research. This dataset contains information on places of interest at the block level across the United States. We consider the number of eating places, full-service restaurants, fast food restaurants, coffee shops, bars, furniture stores, electronics stores, supply stores, clothing stores, sports and music stores, department stores, pet stores, cosmetic stores, used merchandise stores, and health stores in each Census tract in the Los Angeles County in 2017.

2.2 Data for the Model

While we use the data on amenities for Los Angeles County only, we build a quantitative model of the entire Los Angeles Commuting Zone (CZ) in order to represent a self-contained labor market. The Los Angeles CZ comprises five counties (Los Angeles, Orange, Riverside, San Bernardino, and Ventura) and has a total population of 18.7 million as of 2018. Below we describe the data we use to quantify the model and Appendix Table A.1 reports summary statistics.

Residents and employment. To construct Census tract-level data on the residential and workplace employment, we use the LEHD Origin-Destination Employment Statistics (LODES) data for years 2012 to 2016. To exclude nearly empty desert and mountain

[4] FBI classification of crimes against persons, property, and society: https://ucr.fbi.gov/nibrs/2011/resources/crimes-against-persons-property-and-society

tracts with large land areas, we exclude tracts that are below the 2.5th percentile of both residential and employment density. This excludes less than 1% of workers and leaves us with 3,847 tracts. The LODES dataset includes the total number of workers in each tract, separately by the place of residence and the place of work. It also includes the number of workers by education, race, age, income and gender.

Commuting times and flows. Time taken to commute between pairs of census tracts is taken from Delventhal et al. (2022). Data is obtained using the Census Transportation Planning Products (CTPP) dataset for years 2012–2016. Commuting flows between tracts are obtained from the LODES dataset.

Wages and housing rents. We construct a wage index for each tract using blockgroup-level wage and employment composition data from the IPUMS NHGIS that is based on the 2012–2016 American Community Survey (ACS). We also construct a rent index for each tract using tract-level data on self-reported rents and housing characteristics from the same source. Appendix Section A.1 provides more details.

3 Quantitative Spatial Model

Next, we build a QSM of the Los Angeles CZ. Our model follows the structure of the canonical model of Ahlfeldt et al. (2015) and is based on the model of Los Angeles in Ang (2024). Each model location corresponds to a Census tract.

3.1 Model

Workers. In the model, workers choose where to live and where to work by maximizing utility. Their utility depends on local amenities, housing prices, wages, and commuting costs. They consume a traded good and housing. We assume a Cobb-Douglas utility function with housing expenditure share γ, which implies that the indirect utility function is

$$V_{ij} = \frac{X_i \, E_j \, w_j}{q_i^{\gamma} \, d_{ij}},$$

where V_{ij} represents the utility associated with living in location i and working in location j, X_i is the residential amenity, E_j is the workplace amenity, q_i is the cost of housing in i, w_j is the wage available in j, and d_{ij} is the cost of commuting from i to j. We parameterize the commuting cost function as

$$d_{ij} = \exp(\kappa t_{ij}),$$

where t_{ij} is time in minutes required to travel from location i to location j.[5]

Individuals draw idiosyncratic preference shocks for residence-workplace location pairs from a Fréchet distribution with shape parameter θ. Thus, the fraction of households living in location i and commuting to j is

$$\pi_{ij} = \frac{V_{ij}^{\theta}}{\sum_{i'} \sum_{j'} V_{i'j'}^{\theta}}. \tag{3.1}$$

We can then obtain the fraction of workers living in i by summing the choice probabilities across all workplace locations:

$$N_{Ri} = \sum_{j} \pi_{ij} = X_i q_i^{-\gamma} \sum_{j} \frac{E_j \, w_j}{d_{ij}} = X_i q_i^{-\gamma} CMA_i. \tag{3.2}$$

The previous expression highlights that fact that, besides housing costs and commuter market access CMA_i, the equilibrium number of residents depends on residential amenities X_i.[6] Similarly, the fraction of jobs in each location is given by

$$N_{Wj} = \sum_{i} \pi_{ij}. \tag{3.3}$$

Traded-good firms and floorspace developers. In each location, there are perfectly competitive firms that produce a traded good using labor and commercial floorspace. Their production function is

$$Y_i = A_i \, N_{Wi}^{\alpha} \, H_{Wi}^{1-\alpha},$$

where A_i is total factor productivity and α is the labor share. The good is costlessly traded across locations.

Each location is also populated by perfectly competitive real estate developers that produce floorspace. Their production function is

$$H_i = K_i^{1-\eta_i} \, (\phi_i L_i)^{\eta_i},$$

where K_i and L_i are final good and land inputs, ϕ_i is land-augmenting construction productivity, and η_i is the share of land in production. The supply of floorspace is split

[5]The exponential commuting cost function is standard in QSMs.

[6]We abstain from modeling amenities as endogenous, as is common in many QSMs, for two reasons. First, there is no consensus on the correct functional form of amenities. Second, in most models amenities are a function of population density, income, or neighborhood composition, and these features are correlated with amenities in the data.

into residential and commercial use, $H_i = H_{Ri} + H_{Wi}$, and is rented to households and firms at price q_i.

Traded-good firms choose labor and floorspace to maximize profits. Together with the zero-profit condition, this gives us the following expression for equilibrium wages:

$$w_i = \alpha\, A_i^{\frac{1}{\alpha}} \left(\frac{1-\alpha}{q_i}\right)^{\frac{1-\alpha}{\alpha}}. \tag{3.4}$$

Developers choose land and traded good inputs to maximize profits. Each location has a maximum amount of available land Λ_i and, since there is no alternative use of land, developers optimally choose $L_i = \Lambda_i$. The equilibrium floorspace supply is given by

$$H_i = \phi_i (1-\eta_i)^{\frac{1-\eta_i}{\eta_i}}\; q_i^{\frac{1-\eta_i}{\eta_i}}\, \Lambda_i,$$

where $\frac{1-\eta_i}{\eta_i}$ corresponds to the housing supply elasticity. Household utility maximization implies that local demand for residential floorspace is $\gamma W_i/q_i$, where $W_i \equiv \sum_j \pi_{ij} w_j$ is the total income or residents of location i. Profit-maximization of traded good firms yields the commercial floorspace demand $((1-\alpha)A_i/q_i)^{\frac{1}{\alpha}} N_{Wi}$. Thus, the equilibrium floorspace price is implicitly defined by the following market clearing condition:

$$\phi_i (1-\eta_i)^{\frac{1-\eta_i}{\eta_i}}\; q_i^{\frac{1}{\eta_i}} \Lambda_i = \gamma W_i + ((1-\alpha)A_i)^{\frac{1}{\alpha}}\, q^{-\frac{1-\alpha}{\alpha}} N_{Wi}. \tag{3.5}$$

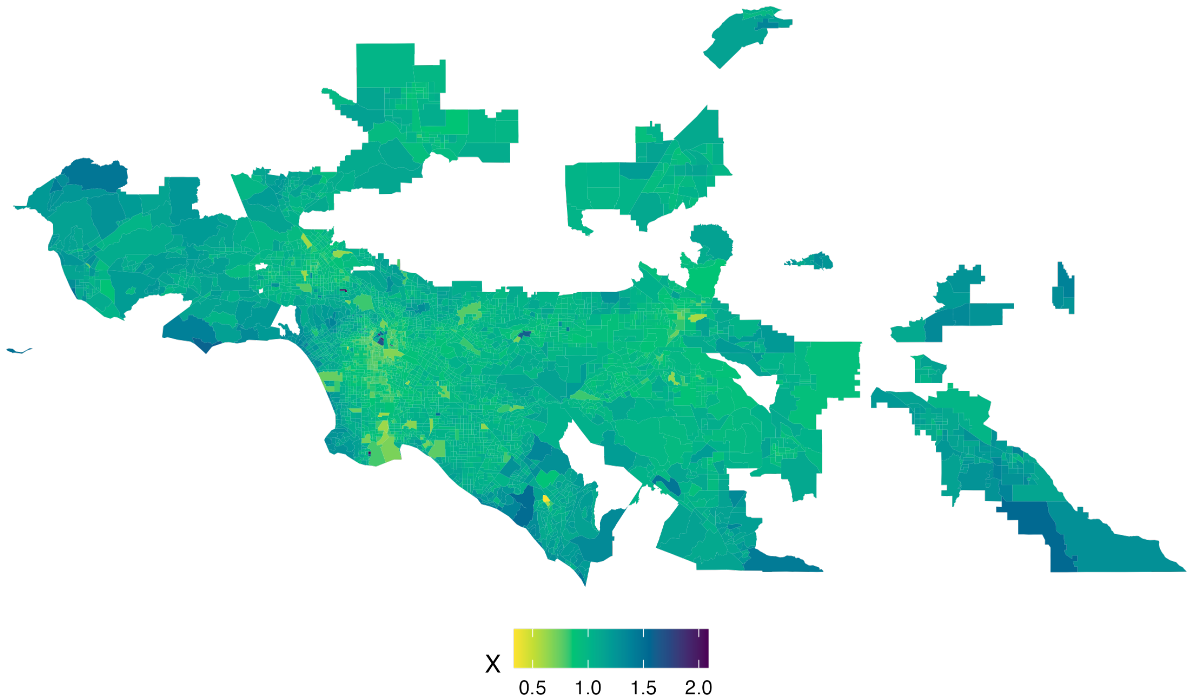
Equilibrium. Given economy-wide parameters $\{\alpha, \gamma, \kappa, \theta\}$, exogenous location fundamentals $\{X_i, E_i, A_i, \phi_i, \eta_i\}$, and the commuting time matrix $\{t_{ij}\}$, an equilibrium is given by $\{N_{Ri}, N_{Wi}, q_i, w_j\}$ such that equations (3.2), (3.3), (3.4), and (3.5) hold. The Walras law implies that the market for the traded good clears when all other markets clear.

3.2 Parameter Values

We take several parameters from the literature. The labor share is set to $\alpha = 0.82$ using the land and structures share estimate of 0.18 from Valentinyi and Herrendorf (2008). The share of housing in utility is $\gamma = 0.24$, following Davis and Ortalo-Magné (2011). Finally, we take the commuting elasticity $\kappa = 0.0126$ from Severen (2021), who estimates it for Los Angeles.

To obtain the values of local parameters, we invert the model and obtain values of X_i and E_i that ensure that population and employment shares in the model are equal to the shares in the data. Similarly, we obtain values of A_i and ϕ_i such that wages and rents in

Figure 1: Inverted Values of X_i



Notes: This map shows the values of inverted X_i.

the model are identical to those in the data. We also let housing supply elasticities, $\frac{1-\eta_i}{\eta_i}$, to differ by tract and use the 2011 FMM-IV estimates of total floorspace elasticity from Baum-Snow and Han (2024).

The values of structural residuals X_i represent the model-implied amenities that we will compare to the amenities we observe in the data. Figure 1 shows the spatial distribution of inverted residential amenities X_i. While X_i is positively correlated with local population, note that the R^2 in the regression of X_i on population is only 0.22, i.e., local population accounts for just about one-fifth of the variation in model-implied amenities.

To estimate the Fréchet elasticity θ, we construct the maximum likelihood function

$$\ln \mathcal{L} \equiv \sum_i \sum_j \pi_{ij}^{\text{data}} \ln \left[\frac{\varphi_{Ri} \varphi_{Wj} e^{-\kappa\theta t_{ij}}}{\sum_{i'} \sum_{j'} \varphi_{Ri'} \varphi_{Wj'} e^{-\kappa\theta t_{i'j'}}} \right]$$

using observed commuting flows π_{ij}^{data} and the "gravity equation" (3.1) for equilibrium commuting flows. All local variables are subsumed into residence and workplace fixed effects φ_{Ri} and φ_{Wj}. We estimate $\kappa\theta$ using Poisson pseudo maximum likelihood (PPML) and, as Appendix Table A.2 reports, the estimated value of $\kappa\theta$ is 0.0876. Note that we cannot separately identify κ and θ. Thus, in order to infer θ we use the calibrated value of κ described above, and obtain $\theta = 0.0876/0.0126 = 6.95$, which is within the range of other estimates in the literature.

4 Analysis

In the theoretical model, the level of amenities X_i accounts for the variation in residential population that is not explained by housing costs and access to jobs. In the quantitative model, it is a structural residual that ensures that the distributions of residents in the model and in the data are identical. It is the main object of interest in this section, and our analysis will aim to determine whether the values of X_i inferred from the quantitative model described in Section 3 can be explained by observed amenities that we described in Section 2. We estimate

$$X_i = \beta_0 + \beta_1 Z_i + \varepsilon_i, \tag{4.1}$$

where X_i is the value of model-implied amenities in location i and Z_i is the vector of 41 observed amenities in i.

Do observed amenities explain X_i? The regression (4.1) yields an R^2 of 0.452 (see Table 2). That is, the observed amenities in our dataset can explain less than half of the variation in model-implied amenities. The standardized estimated coefficients of each component β_1 and their 95% confidence intervals are reported in Appendix Figure A.1.[7] For example, amenities such as ruggedness, school quality, and single-family housing share are positively correlated with X_i, whereas solar radiation in July, traffic, and the number of fast food restaurants are negatively correlated with X_i.

Which amenities explain X_i? In order to determine the relative importance of different observed amenities in explaining the variation in model-based amenities, we run the Shapley R^2 decomposition (Shapley, 1953; Shorrocks, 1999). In particular, we examine the contribution of each of the six groups of amenities.[8]

Table 2 shows that school quality accounts for over 29% of the R^2, while consumption amenities account for over 25%. Other man-made amenities, such as pollution, heat, and traffic, crime, and built environment jointly account for 15%. Natural amenities are responsible for the remaining 30%. These findings provide support for modeling

[7]Because amenity variables differ in magnitudes, they are not directly comparable. For comparability of regression coefficients, we multiply each component of β_1 and its 95% confidence interval bounds by the ratio of the standard deviation of the regressor to the standard deviation of the dependent variable.

[8]Full R^2 decomposition with k regressors requires running 2^k regressions. Given that we have 41 amenity variables, to reduce the computational burden we run the decomposition for the six amenity groups.

Table 2: R^2 Decomposition

Category	Shapley Value	Share of R^2
Natural Amenities	0.137	30.43 %
Pollution, Heat & Traffic	0.035	7.69 %
Crime	0.011	2.51 %
Built Environment	0.022	4.85 %
School Quality	0.132	29.20 %
Consumption Amenities	0.114	25.32 %
Total	0.452	100.00 %

Notes: This table lists the contributions of the six amenity groups to the total R^2 value of the regression (4.1), estimated using the Shapley R^2 decomposition.

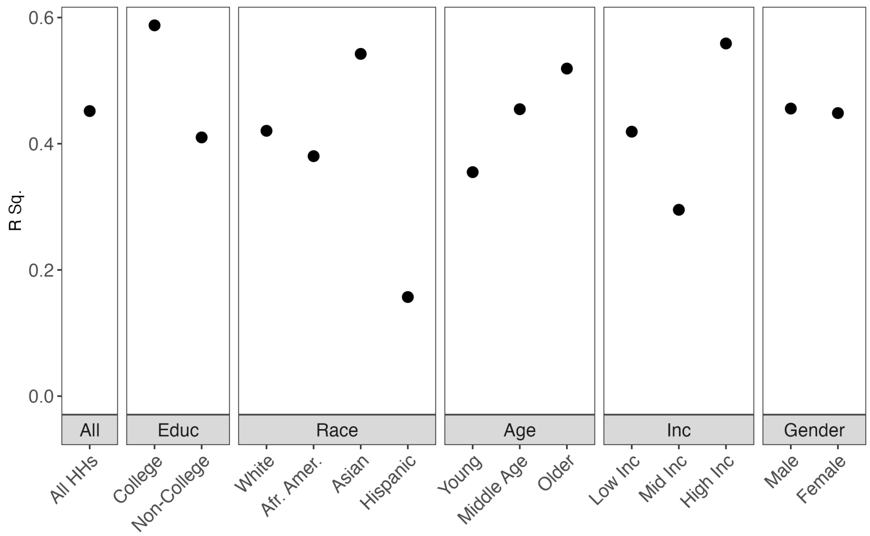
amenities as an endogenous variable and not an exogenous feature of natural geography. Moreover, given that school quality and the concentration of consumption amenities are positively correlated with income, these findings support models where the quality of local amenities depends on income or education of local residents.

Heterogeneity. It is possible that observed amenities account for less than half of model-based amenities because in our model workers have identical valuations of amenities, whereas in reality these valuations differ across individuals. To address this issue, we extend the quantitative model by introducing different types of individuals using categories in the LODES dataset. We consider versions of the model where individuals differ by education (college graduates and those without a college degree), race (white, African American, Asian, and Hispanic), age (29 and younger, 30–54, and 55 and older), income ($\leq \$1,250$, \$1,251–\$3,333, and $> \$3,333$), and gender (male and female). We assume that those types have the same preferences and are perfectly substitutable in production, but have different valuations of residential amenities X_i. We invert X_i separately for each individual type using the data on the distribution of local population by education, race, age, income, and gender. Then, as before, we regress it on the vector of observed amenities.

Figure 2 shows the regression R^2 associated with each individual type. We find that observed amenities are more successful at explaining the residential structural residuals for college-educated, Asian, older and high-income individuals. At the same time, they are less successful in explaining the residuals for Hispanic, young, and middle-income individuals.

Figure 3 describes the R^2 decomposition for each type. Panel A shows that school quality is the most important determinant of model-implied amenities of college-educated individuals, while natural amenities are the most important determinant of X_i of those

Figure 2: R^2 Values Across Individual Types

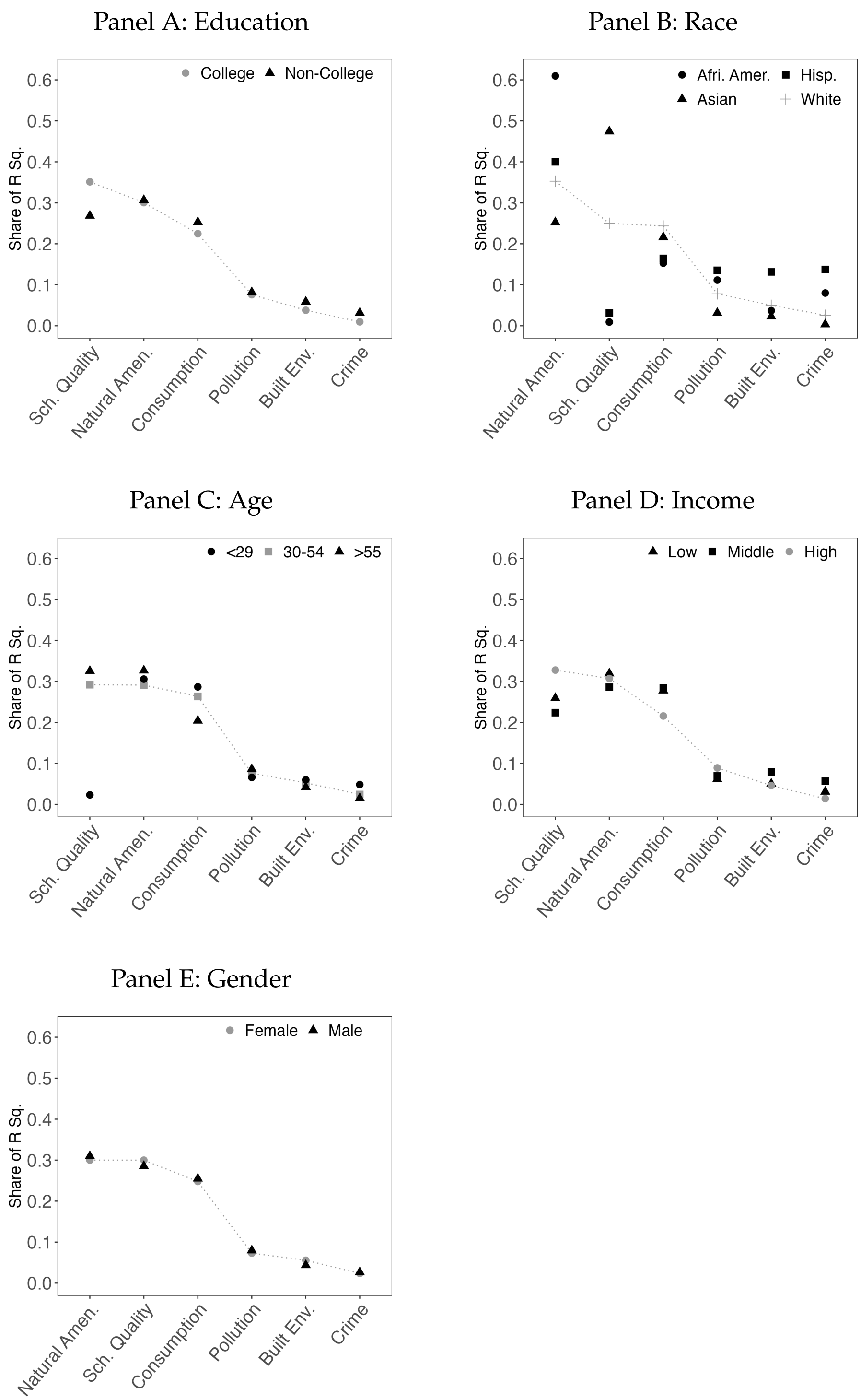


Notes: This figure shows the R^2 from a regression of type-specific residuals X_i on observed amenities, separately for each type.

without a college degree. Panel B shows that, while consumption and natural amenities are important determinants of X_i for all races, school quality has little effect on the estimated X_i for Hispanics and African Americans, possibly reflecting that those races tend to be concentrated in areas with lower school rankings. Looking at panel C, we can see that the high importance of consumption and natural amenities holds for all age groups. At the same time, school quality is unimportant for the young individuals' X_i, presumably because they are unlikely to have school-age children in the household. The decomposition by income in panel D resonates with the findings in panel A and shows that school quality is particularly important in account for the X_i of the high-income, while consumption amenities are more important for the X_i of middle- and low-income individuals. Panel E shows that the relative importance of each group of amenities is similar for men and women.

What accounts for the unexplained variation in X_i? While we collected the data on as many as 41 different amenity variables that include most types of amenities emphasized in the literature, there could be other amenities that are important for location choices and, therefore, the values of X_i. It is also worth noting that X_i is positively correlated with local population (see Appendix Table A.3) and inferred from a static model. However, in practice the population of a given location does not only respond to current amenities but also, due to moving costs, depends on past amenities that we do not observe.

Figure 3: R^2 Decomposition by Household Type



Notes: This table lists the contributions of the six amenity groups to the total R^2 value of the regression of type-specific X_i on observed amenities, estimated using the Shapley R^2 decomposition.

5 Conclusions

In this paper we studied the nature of the residential structural residuals that QSMs rely on and that are often interpreted as amenities. We found that about 45% of the variation in model-implied amenities can be explained by an extensive array of 41 different amenities in the Los Angeles County. We also found that man-made amenities are more important than natural amenities in accounting for the variation in the structural residuals. These results suggest that one should be cautious when treating the structural residuals in QSMs as amenities and that amenities should be modeled as endogenous objects.

We also note that the availability of spatially granular data on many kinds of amenities, such as the ones we use in this paper, should allow researchers to model amenities not just as a single variable that is a simple function of population density, income, or skill composition, as is standard in QSMs. It is also possible to incorporate different types of amenities in a QSM, each as a separate equilibrium variable. A more detailed modeling of amenities may allow for a more precise measurement of welfare consequences of neighborhood change in policy counterfactuals.

Bibliography

Ahlfeldt, G. M., Redding, S. J., Sturm, D. M., and Wolf, N. (2015). The Economics of Density: Evidence from the Berlin Wall. *Econometrica*, 83(6):2127–2189.

Allen, T. and Arkolakis, C. (2022). The welfare effects of transportation infrastructure improvements. *The Review of Economic Studies*, 89(6):2911–2957.

Almagro, M. and Domínguez-Iino, T. (2024). Location sorting and endogenous amenities: Evidence from Amsterdam. *Working paper*.

Ang, A. (2024). Paradise lost: Population growth and wildfire mitigation. *Working paper*.

Baum-Snow, N. and Han, L. (2024). The microgeography of housing supply. *Journal of Political Economy*, 132(6):1897–1946.

Chen, L., Hasan, R., Jiang, Y., and Parkhomenko, A. (2024). Faster, taller, better: Transit improvements and land use policies. *Journal of Development Economics*, page 103322.

Couture, V., Gaubert, C., Handbury, J., and Hurst, E. (2024). Income growth and the distributional effects of urban spatial sorting. *Review of Economic Studies*, 91(2):858–898.

Davis, M. A. and Ortalo-Magné, F. (2011). Household expenditures, wages, rents. *Review of Economic Dynamics*, 14(2):248–261.

Delventhal, M. and Parkhomenko, A. (2023). Spatial implications of telecommuting. *Available at SSRN 3746555*.

Delventhal, M. J., Kwon, E., and Parkhomenko, A. (2022). JUE insight: How do cities change when we work from home? *Journal of Urban Economics*, 127:103331.

Diamond, R. (2016). The determinants and welfare implications of US workers' diverging location choices by skill: 1980–2000. *American Economic Review*, 106(3):479–524.

Heblich, S., Redding, S. J., and Sturm, D. M. (2020). The making of the modern metropolis: evidence from London. *The Quarterly Journal of Economics*, 135(4):2059–2133.

Nunn, N. and Puga, D. (2012). Ruggedness: The blessing of bad geography in africa. *Review of Economics and Statistics*, 94(1):20–36.

Redding, S. J. (2023). Quantitative urban models: From theory to data. *Journal of Economic Perspectives*, 37(2):75–98.

Redding, S. J. and Rossi-Hansberg, E. (2017). Quantitative Spatial Economics. *Annual Review of Economics*, 9:21–58.

Severen, C. (2021). Commuting, labor, and housing market effects of mass transportation: Welfare and identification. *Review of Economics and Statistics*, pages 1–99.

Shapley, L. S. (1953). A value for n-person games.

Shorrocks, A. F. (1999). Decomposition procedures for distributional analysis: a unified framework based on the shapley value. Technical report, University of Essex.

Tsivanidis, N. (2023). Evaluating the impact of urban transit infrastructure: Evidence from Bogotá's TransMilenio. *Working paper*.

Valentinyi, A. and Herrendorf, B. (2008). Measuring factor income shares at the sectoral level. *Review of Economic Dynamics*, 11(4):820–835.

Zárate, R. D. (2024). Spatial misallocation, informality, and transit improvements: Evidence from Mexico City. *Working paper*.

A Appendix

A.1 Wage and Rent Indices

Wages. We use blockgroup-level data from IPUMS NHGIS (ACS 2012-2016) and construct a wage index following Delventhal and Parkhomenko (2023). First we regress the log of median self-reported income per blockgroup b in Census tract i on a set of observable characteristics:

$$\log \text{Income}_b = \beta_0 + \sum_a \beta_{2,a} \text{ Age Bins}_{a,b} + \beta_3 \text{ Female Shares}_b + \sum_r \beta_{4,r} \text{ Race Shares}_{r,b} \\ + \sum_k \beta_{5,k} \text{ Industry Shares}_{k,b} + \sum_o \beta_{6,o} \text{ Occupation Shares}_{o,b} + \mu_i + \epsilon_b.$$

We then generate the tract-level wage index by taking the constant $\hat{\beta}_0$ and the tract-level fixed effect μ_i:

$$w_i = \hat{\beta}_0 + \mu_i.$$

When self-reported wage data is not available for a tract, we use the average wage index value of its neighboring tracts.

Rents. We use tract-level data from IPUMS NHGIS (ACS 2012-2016) and construct an index of housing rents following Delventhal and Parkhomenko (2023). We regress the log of median self-reported rent per Census tract on a set of observable characteristics:

$$\log \text{Rent}_i = \beta_0 + \beta_1 \text{ Rooms}_i + \beta_2 \text{ Year Built}_i + \beta_3 \text{ Number of Units}_i + \iota_i,$$

where Rooms_i is the median number of rooms per dwelling, Year Built_i is the median year built across all dwellings, and Number of Units_i is the modal number of units per dwelling in each tract i. We then generate the tract-level rent index by taking the constant and the tract-level residual:

$$q_i = \hat{\beta}_0 + \hat{\iota}_i.$$

When self-reported rent data is not available for a tract, we use the average rent index of its neighboring tracts.

A.2 Additional Figures and Tables

Table A.1: Data for the Model

Variable	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
Wage Index	61,454	13,061	17,286	53,508	68,446	170,987
Floorspace Price Index	-0.12	0.52	-4.5	-0.38	0.14	3.3
Residents	1,921	788	55	1,384	2,351	7,710
Education						
College	442	268	6.8	235	589	2,923
Non-College	1,039	426	18	754	1,273	4,454
Race						
White	1,417	641	22	967	1,788	5435
African American	149	172	3.6	59	167	2,177
Asian	264	274	3.8	90	329	3,284
Hispanic	711	453	12	369	954	3,965
Age						
29 and Younger	421	176	11	303	515	1,690
30-54	1,077	478	20	755	1,330	5066
55 and Older	403	177	6.2	276	506	1,415
Income						
$1,250 a month or less	477	189	11	352	577	1,779
$1,251 a month - $ 3,333 a month	656	280	14	468	806	2,545
More than $ 3,333 a month	769	452	12	436	1,006	4,909
Gender						
Male	977	413	20	701	1,202	4,036
Female	924	387	16	662	1,143	3,789
Employment	1,938	4,745	1.4	372	1,781	136,332
Education						
College	447	1431	0.6	67	346	45,825
Non-College	1,042	2,528	0.33	211	959	78,604
Race						
White	1,427	3,360	1.2	267	1,317	81,748
African American	148	537	0	22	124	25,651
Asian	266	772	0	33	225	23,391
Hispanic	708	1,637	0	138	664	47,875
Age						
29 and Younger	424	895	0.2	76	408	16,621
30-54	1,082	2,857	0.4	198	941	89,611
55 and Older	406	1,042	0	83	369	34,818
Income						
$1,250 a month or less	480	1,089	0	129	486	31,899
$1,251 a month - $ 3,333 a month	659	1,370	0.8	128	650	25,322
More than $ 3,333 a month	773	2,826	0	85	545	114,383
Gender						
Male	982	2,633	1	158	845	73,054
Female	930	2,135	0.4	196	870	63,278

Table A.2: Estimation of the Gravity Equation

	π_{ij}
t_{ij}	-0.0876***
	(0.000110)
N	14,799,409
pseudo R^2	0.627

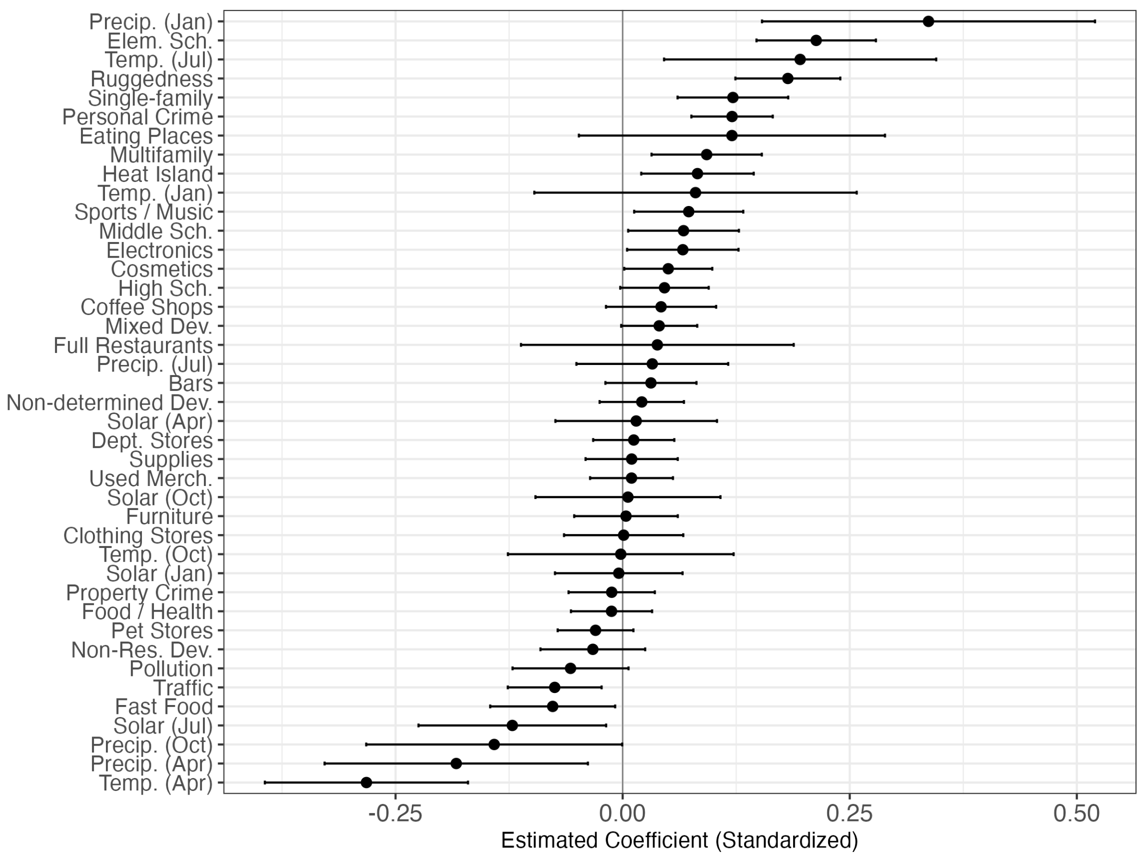
Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A.3: Relationship between X and Population

	(1)
	X
Pop. (log)	0.138***
	(0.00420)
N	3847
adj. R^2	0.219

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure A.1: Regression Output



Notes: This figure shows standardized estimated coefficients of each component β_1 and their 95% confidence intervals. We standardize each coefficient by multiplying each component of β_1 and its 95% confidence interval bounds by the ratio of the standard deviation of the regressor to the standard deviation of the dependent variable.