If You Zone It, Who Will Come, and How Will They Travel?

The Effects of Relaxed Zoning Regulations on Travel Behavior

by

Matthew Wigginton Conway

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved Month Year by the
Graduate Supervisory Committee:

Deborah Salon, Chair
Deirdre Pfeiffer
Stewart Fotheringham
Michael van Eggermond

ARIZONA STATE UNIVERSITY

April 2021

ABSTRACT

Urban areas across the Unites States are facing a housing affordability crisis. One approach some cities and states have taken is to reduce or eliminate single-family zoning. Single-family zoning prevents the construction of more-affordable apartments in vast swaths of the American urban landscape. This policy shift has already occurred in Minneapolis, Sacramento, and Oregon, and is under discussion in California, Massachusetts, and North Carolina, among others.

Independent of any effects on housing affordability, changes to land use will have effects on transport. I evaluate these effects using a microsimulation framework. In order for land use policies to have an effect on transport, they need to first have an effect on land use, so I first build an economic model to simulate where development will occur given a loosening of single-family zoning. Transport outcomes will vary depending on which households live in which parts of the region, so I use an equilibrium sorting model to forecast how residents will re-sort across the region in response to the land use changes induced by new land-use policies. This model also jointly forecasts how many vehicles each household will choose to own. Finally, I apply an activity-based travel demand microsimulation model to forecast the changes in transport associated with the forecast changes from the previous models.

I find that while there is opportunity for economically-feasible redevelopment of single-family homes into multifamily structures, the amount of redevelopment that will occur varies greatly depending on the exact expectations of developers about future market conditions. Redevelopment is focused in higher-income neighborhoods.

The transport effects of the redevelopment are minimal. Average car ownership across the region does not change hardly at all, although residents of new housing units do have somewhat lower car ownership. Vehicles kilometers traveled, mode choice, and congestion change very little as well. This does not mean that upzoning does not affect transport in

general, but that more nuanced proposals may be necessary to promote desirable transport outcomes. Alternatively, the results suggest that upzoning will not worsen transport outcomes, promising for those who support upzoning on affordability grounds.

ACKNOWLEDGMENTS

If it takes a village to raise a child, it takes at least a Census tract or two to write a dissertation. I am indebted to many who have supported me on this journey through graduate school, and in my life leading up to it. Academically, Deborah Salon has been an excellent mentor, collaborator, and friend, guiding me through this endeavor. I have truly appreciated the time I've spent with her, both as I worked through my dissertation, and as we collaborated on a myriad of other projects. My other committee members have likewise provided invaluable support. I come from a transportation background, and Deirdre Pfeiffer has excelled at teaching me about housing issues. Stewart Fotheringham helped spark my interest in advanced econometrics during my first semester at ASU, and has continued to ask challenging and useful questions throughout the dissertation process. I've known Michael van Eggermond longer than any of my other committee members, first meeting him before graduate school while working in consulting. He has always adeptly bridged the gaps between my prior experience with accessibility models, and this new world of microsimulation models.

I am indebted to numerous professors I have worked with throughout my undergraduate and graduate careers, including (at Foothill College) K. Allison Lenkeit Meezan, who encouraged me to pursue geography and to present at my first conference; Christine Hansell, who first suggested I should pursue a PhD; and Jennifer Carlin-Goldberg, who taught my first statistics class and helped me realize I actually liked math; (at UC Santa Barbara) Helen Couclelis, who treated me like a graduate student even though I wasn't; Kostas Goulias, who encouraged me to pursue transportation modeling; Stuart Sweeney, who taught me spatial statistics; (at Arizona State University, in addition to my committee) Kevin McHugh and Rashad Shabazz, who impressed upon me the value and importance of critical approaches to geography; David King, who has been an excellent collaborator on several projects; Peter Kedron, who helped me make sense of the academic job market, and Ram Pendyala, who directs the Center for Teaching Old Models NEw Tricks and who taught me survey methods and advanced transportation data analysis.

I want to thank my parents, Michele and Dan, and my brother, for raising me in a household where knowledge and inquiry were valued, and for providing a respite from graduate school during several family vacations. I thank my aunt Mary Lou Davis (in memoriam) and Jim Love for welcoming me to the Phoenix area and making me feel at home.

A number of close friends have supported me through the process of completing my dissertation. Anson Stewart, Bonnie Bounds, Connor Larsen, Kush Bhagat, and Sam Zhang have all been available to assist me, be sounding boards, and give advice from their own graduate school experiences. I owe convergence of my sorting models to Sam, without whom I would probably still be trying to get the Los Angeles housing market to clear. Anson Stewart and Conveyal also provided me with consulting opportunities and office space at various points in my degree program. Elaine Hoffman answered my questions about architecture and construction cost estimating.

I thank the team at California YIMBY for first getting me interested in this topic, bringing me to your workshop in January 2018. Salman Ahmad, Paavo Monkkonen, and Salim Furth provided valuable feedback and advice on the models in this dissertation.

And of course, I could not have done this without my partner, Dhruti Bhagat. I'm excited that soon I'll be able to call you my wife. Our guinea pigs, Adam, Bernard, and Bella, have also been supportive throughout the process—may Adam rest in peace.

While final computations were carried out on Amazon Web Services, significant portions of the computation for this dissertation was computed using the Arizona State University Agave high-performance computing cluster. I am indebted to ASU Reseach Computing for supporting these efforts.

Portions of this dissertation are based on the Zillow Transaction and Assessment

Dataset. Data provided by Zillow through the Zillow Transaction and Assessment Dataset (ZTRAX). More information on accessing the data can be found at http://www.zillow.com/ztrax. The results and opinions are those of the author(s) and do not reflect the position of Zillow Group.

During this research, the author was supported by by the School of Geographical Sciences and Urban Planning NEXUS/SNRF fellowship, the Dwight David Eisenhower Transportation Fellowship Program, and the Center for Teaching Old Models NEw Tricks.

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

Table | Page

LIST OF FIGURES

Figure Page

Figure Page

Chapter 1

INTRODUCTION AND MOTIVATION

Many of America's most economically productive urban areas are facing a housing affordability crisis (Shaw 2018). These areas demonstrate low supply and high demand for housing units. One mechanism that many argue has contributed to this shortage is restrictive zoning that constrains the ability of developers to produce housing—particularly multifamily housing, which is generally less expensive than single family homes, but which is not allowed to be built in many neighborhoods (*e.g.* Glaeser 2011; Molloy 2018). While exclusive single-family zoning is ubiquitous in the US, it is largely absent in many other nations (Hirt 2014).

Recently, a number of US cities and states have relaxed or are considering relaxing their zoning ordinances, to promote additional housing supply and by extension housing affordability (Infranca 2019; Dillon 2020; Mervosh 2018; Bliss 2019; Phillips 2020). There is significant evidence that this works (Shaw 2018; Been, Ellen, and O'Regan 2019; Logan 2021), although others question the efficacy of these policies (Rodríguez-Pose and Storper 2019).

Infranca (2019) details many of these policy changes. Massachusetts has used financial incentives to encourage communities to create dense, mixed-use zones. California has simplified approval processes for infill projects in cities that fail to meet their affordable housing obligations under the law discussed previously. California was until recently considering a bill that would preempt local single-family zoning in areas near job centers or public transportation across the state (Dillon 2020). In 2019, Oregon preempted local zoning laws prohibiting medium-density housing (Bliss 2019). North Carolina is currently considering allowing fourplexes in any residential zone statewide (*Increase Housing Opportunities* 2021).

City-level efforts have also occurred; Minneapolis recently stopped zoning any parts of the city for exclusive single family use, allowing small apartment buildings in residential areas throughout the city (Mervosh 2018). Sacramento recenly voted to eliminate single-family zoning citywide in their next general plan update (Clift 2021).

Separate from any effects on housing affordability, the built environment also has an effect on travel behavior (*inter alia,* Ewing and Cervero 2010; Aston et al. 2020). Therefore, changes to the zoning regulations that control the built environment will also induce changes in travel—either positive, by inducing walking, biking, and transit use, or negative, inducing more driving and congestion. Indeed, in the early 2010's, California passed SB 375 in an attempt to use housing and land use policy to reduce transportation emissions (Barbour and Deakin 2012). While the current land use policy discourse is largely focused on housing affordability outcomes, changes to land use policy to effect housing affordability will also have effects on transportation. This dissertation examines those effects.

Travel behavior is not directly affected by land use policy. Rather, it is affected by actual land use, which is in turn affected by land use policy. Thus, it is necessary to address the question of how land use policy affects travel behavior in two parts: first, how land use policy affects land use, and second, how land use affects travel. I examine two hypothetical changes to land use policy in Southern California. Since single-family zoning is dominant across the US (Hirt 2014), I focus on this type of zoning. The two scenarios I model are a blanket elimination of single-family zoning and an elimination of single-family zoning only in areas around transit stops, consistent with the aims of the now-defunct SB 50 in California (Wiener et al. 2019, § 65918.51–3) and a bill currently being considered in Boston (Logan 2021). These proposals both represent geographically-dispersed upzoning, which should help with housing affordability, as they do not concentrate development in a few neighborhoods (Phillips 2020). I simulate both proposals in the Southern California Association of Governments (SCAG) region, which includes all of Southern California except for San Diego County (Figure 1). This dissertation uses simulation models to forecast the effects of these proposed policies on land use change, the effects of that land use change on residential location choice, and the effects of the land use change and residential location choice on transport.

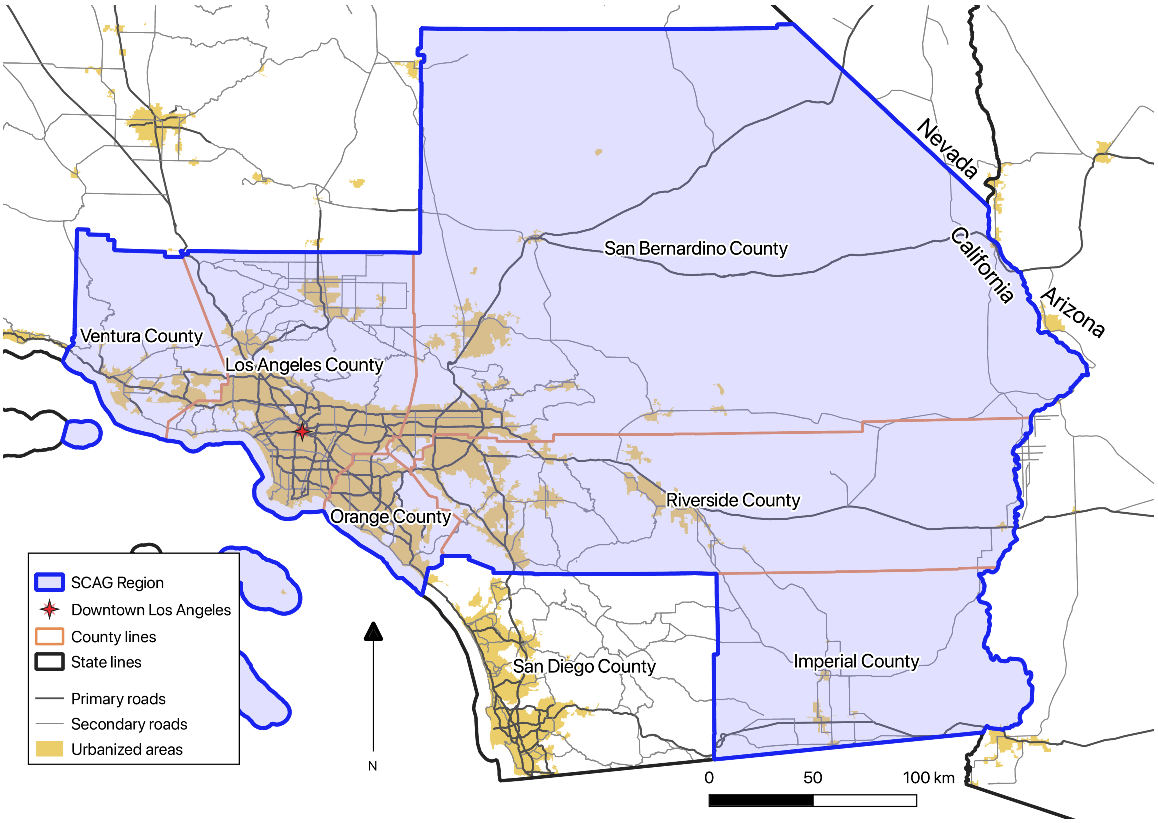


Figure 1. Southern California Association of Governments Region

Changes to land use policy will not necessarily induce changes to land use. Rather, it must be both permitted by policy *and* economically advantageous for changes in land use to occur. For instance, low-price California cities may have permissive zoning for new housing, but they still see little construction due to low demand and thus low profits for developers (Monkkonen, Lens, and Manville 2020). Thus, modeling the economic feasibility of new housing is the first step in determining how travel behavior is affected by these land use policies. I focus on the changes that would be newly permitted by a change from single-family to multifamily zoning: replacing a single-family home (or a vacant lot) previously zoned for single-family housing with a multifamily home. I use a net-present-value model using construction costs and discounted cash flows to model developer decisionmaking about the profitability of new construction. This model and its results are presented in Chapter 2.

The famed Alonso-Mills-Muth model of urban development provides theoretical justification for why upzoning leads to construction in some areas but not in others. This model posits that for every distance from the center of the city, there is an economically-optimal density. In the absence of zoning, the highest densities occur in the center of the city, where land costs are highest and developers substitute capital for land by building taller, denser buildings. Moving away from the center, land costs fall and density does as well (Alonso 1964). However, zoning may prevent these optimal densities from occurring. Consider the theoretical situation shown in Figure 2. As predicted by the Alonso-Mills-Muth model, the optimal density declines smoothly from the central business district to the exurbs. However, zoning causes the density to decline dramatically at a point just outside downtown where multifamily housing ceases to be allowed. In the absence of zoning, the market would provide higher density housing in these locations.

This simplified model demonstrates a key economic aspect of the effect of zoning on cities. Zoning changes will only cause housing to be constructed in places where there is sufficient market demand to support it. If the existing zoning regulations permitted a level of density *higher* than the optimal density, changing the zoning regulations to permit even higher density would have no effect on development, and thus no effect on the actual built environment.

Inspection of zoning maps from many metropolitan areas suggests that there are significant areas where multifamily housing would likely be profitable, but is prohibited by regulation. Figure 3 shows the residential zoning in Southern California. While the downtown

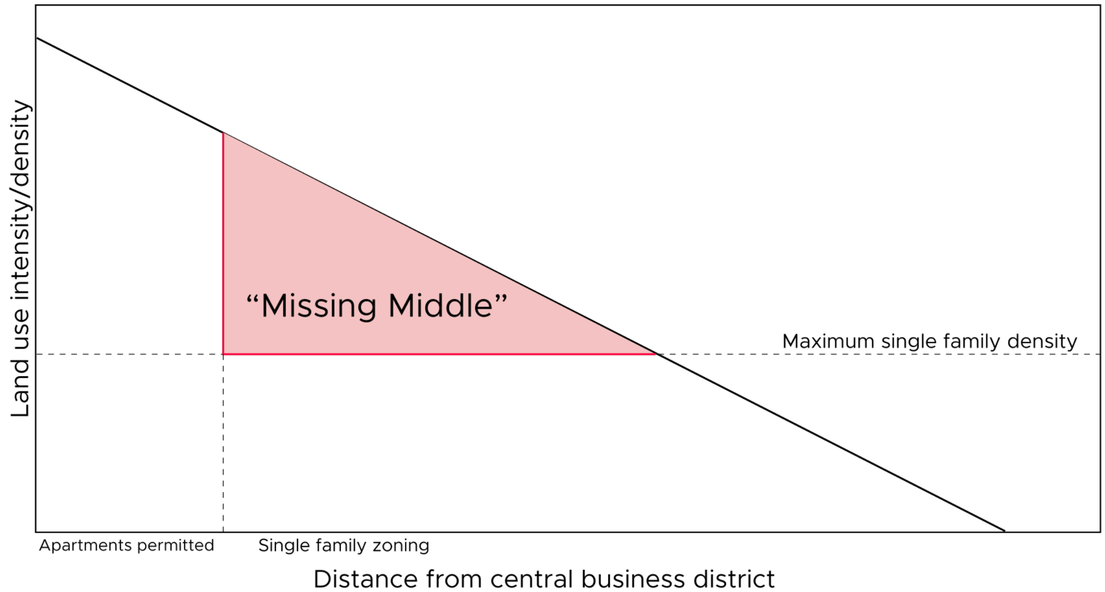


Figure 2. Optimal Density as Predicted by the Alonso-Mills-Muth Model, and Actual Density Constrained by Zoning

areas are zoned for multifamily development (blue), many areas not terribly distant from downtown (and often well-served by transit) are zoned single-family (orange). I expect to find that in many of these closer-in areas, it would be profitable to build multifamily housing if it were permitted, particularly in this example given the high cost of housing in California.

Minimum density regulations are rare, and likely to be ineffectual. Levine (2006) argues that zoning cannot increase density *above what the market would otherwise provide*, because any regulations that attempted to do so would simply result in developers choosing not to develop in that location. The true story is likely to be somewhat more nuanced—a regulated minimum density slightly above what the market considered optimal might not cut into developers' profits enough to make the development infeasible—but the general argument stands.

This dissertation focuses on a single type of development: the development of small

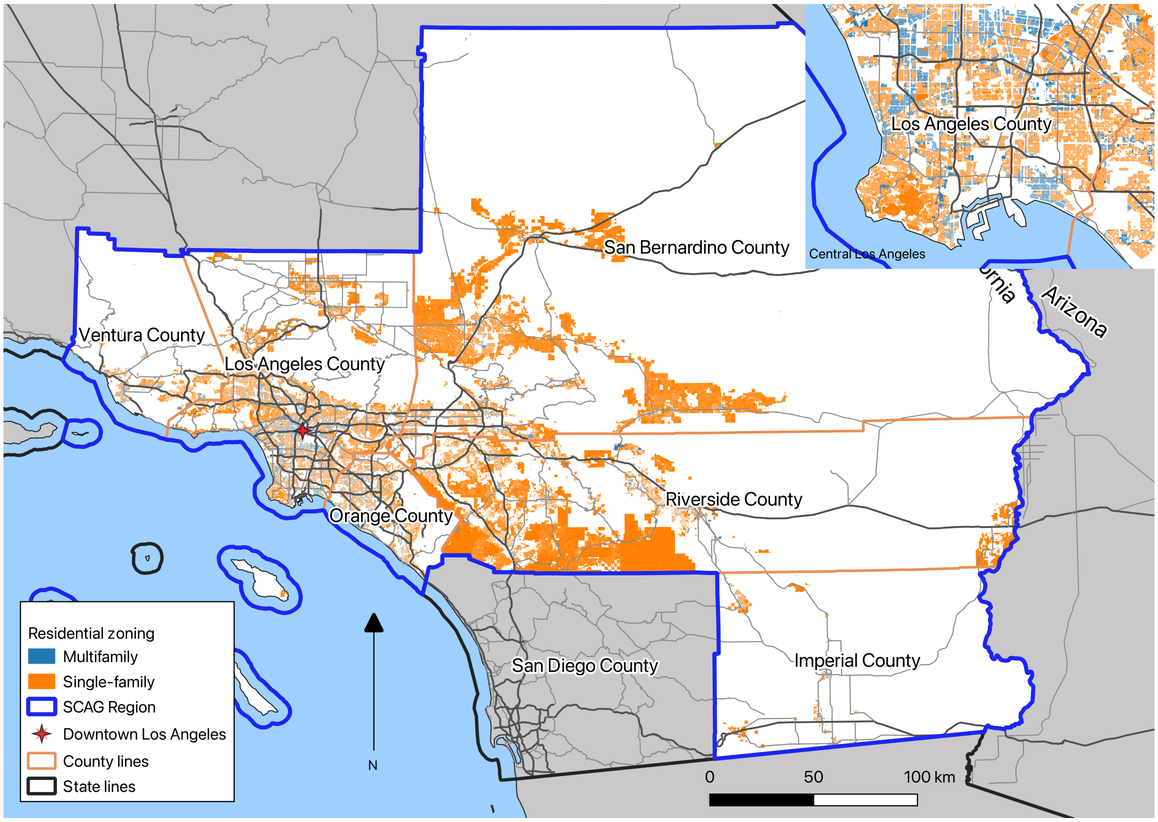


Figure 3. Residential Zoning in Southern California

multifamily homes on existing single-family home lots. I focus on small developments that could fit on a single family home lot because land assembly for larger developments in the vast single-family neighborhoods of American cities is very difficult, and carries a significant (and difficult-to-predict) price premium (Brooks and Lutz 2016; Miceli and Sirmans 2007; Gammage 2016, 163). Small, single-lot developments are the most likely developments to be implemented in the short to medium term, as they do not have to contend with land assembly delays and may be less likely to face strong neighborhood opposition (Pendall 1999). Policy changes to allow such developments have recently been proposed or implemented in Oregon, California, and Minneapolis.

I take a microsimulation approach, simulating the location and transportation choices

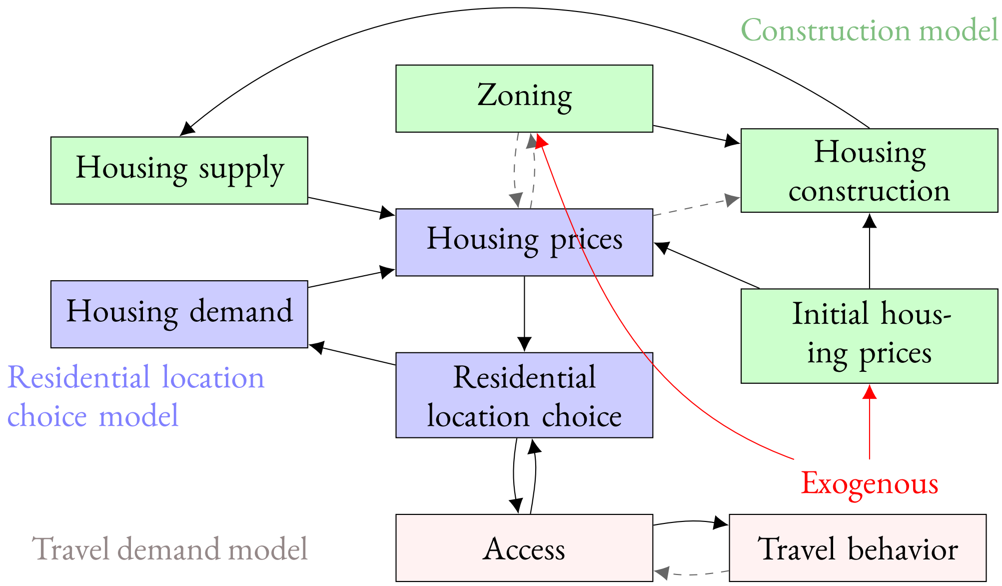


Figure 4. Conceptual Model of Relationships Between Zoning and Travel Behavior

of developers and households. Figure 4 presents a conceptual model of the relationships between zoning, housing supply, residential location choice, and travel behavior. Solid lines indicate relationships that are included in the microsimulation models in this dissertation, while dashed lines are hypothesized to exist in the world but are not modeled in this dissertation for simplicity.

Housing prices are a function of housing supply and housing demand. Housing prices affect housing construction; as prices rise, new development is more likely to be profitable. However, development is also constrained by zoning codes. As housing prices rise, homeowners become more risk-averse in protecting their investment through means including zoning (Fischel 2001). Zoning also affects housing prices by excluding excluding noxious and obnoxious uses such as landfills or scrapyards.[1] Housing construction of course directly affects housing supply.

Residential location choice is a function of housing prices and other exogenous features of housing and neighborhoods. In the aggregate, residential location choice determines housing demand, which in turn affects prices and residential location choice. Residential location choice affects the access (i.e. what urban opportunities one can reach; Committee of the Transport Access Manual 2020) a particular household experiences by determining whether they live in a high-access or low-access neighborhood. In turn, access affects how residents choose to travel. Finally, these travel choices affect access by affecting levels of congestion and thus travel times.

To model this complex system, I divide it into three models, and remove relationships drawn as dashed lines in Figure 4. First, a construction model that forecasts development based on exogenous initial housing prices and zoning scenarios, shown in green in Figure 4. Next, a residential location choice model that forecasts residential choices based on housing prices and housing supply from the construction model, shown in purple. Last, a travel demand model that forecasts travel behavior based on residential location choices and exogenous sociodemographics, shown in salmon.

The construction model is presented in Chapters 2 and 3. Chapter 2 develops a model of whether it will be profitable to develop a variety of different building types on all single-family-zoned parcels in Southern California. In order for construction to actually occur, however, the parcel must be purchased by a developer. I assume most single-family homeowners are not interested in replacing their homes with an apartment building. Estimates of profitability are adjusted with probabilities of sale to a developer in Chapter 3.

[1] Zoning also has a long history of being used to exclude affordable housing and enforce segregation. While this may at least nominally be related to concerns about housing prices, research suggests actual effects are variable and small (Galster 2004, and following comment by Jill Khadduri).

The travel behavior implications of these new units will depend on who lives in them. If they are occupied primarily by high-income individuals, the denser environment may not have as significant an effect on travel behavior as one might expect; high-income individuals drive more and use transit less than lower-income individuals (Pucher and Renne 2003), to the point that some transit construction projects have actually resulted in *less* transit usage due to rising rents and ensuing displacement of low-income, low-car-ownership households (Pollack, Bluestone, and Billingham 2010). The sociodemographics of residents are a key input into the travel demand model that is described in Chapter 6.

The residential location choice model is described in Chapter 4. It takes the form of an equilibrium sorting model (Kuminoff, Smith, and Timmins 2013). This model structure is a extension of a discrete choice model. Discrete choice models simulate how people make choices with a finite choice set by modeling a "utility"—or value—of each choice, and predicting that individuals probabilistically choose the choice with the highest utility (Ben-Akiva and Lerman 1985). The equilibrium sorting model uses a similar framework, but employs an equilibration procedure to account for the fact that the discrete choices are exclusive—only one household can (successfully) choose a particular home. When the model is used to predict a new housing choice equilibrium, an iterative process is used where the demand is calculated. Prices are then raised on housing units where demand exceeds supply, and lowered on housing units where supply exceeds demand, to bring the market to equilibrium. These types of models have been employed in residential location choice research in the past (Bayer, McMillan, and Rueben 2004; Tra 2010).

I am using a sorting model in a somewhat different way than others have; most previous applications of these models have used them to evaluate how people would re-sort in response to a change in the amenities available at a location (*e.g.*, air quality (Tra 2007, 2010,

2013) or open space (Klaiber and Phaneuf 2010)). This requires some changes to the model structure to account for the variable supply of housing between estimation and simulation.

Furthermore, I extend the equilibrium sorting model to jointly model vehicle ownership, as these decisions are closely related. Since car ownership does not need to be brought into a market equilibrium (because there is no exogenously-imposed distribution of car ownership levels, as there is for housing), I extend the equilibrium sorting model to model a multidimensional choice where some dimensions are equilibrated and some are not.

The results of this residential location choice and vehicle ownership model are used in the activity-based travel demand model, to answer the question of how the original zoning changes affect actual travel behavior (Chapter 6). I repurpose the ActivitySim/Travel Model 1 travel demand model originally developed in the San Francisco Bay Area (Erhardt et al. 2012), since both metropolitan areas are in Calfornia and are relatively similar. I make slight modifications to the model, including removing some model components that are Bay Area-specific, and replace the input data with data from the Los Angeles area, but do not re-estimate any of the models due to data availability.

Using the ActivitySim model requires generating a synthetic population with records for individuals and households. Since the residential location choice model re-sorts households across the region, standard population synthesis methods based on small-area demographic estimates from the Census do not apply. Instead, I develop a synthetic population based on the outputs of the residential location choice model, combined small-area housing type prevalence estimated from the construction model.

A key metric to understand transport effects is the level of traffic congestion. ActivitySim does not estimate traffic congestion levels directly; rather, it produces origin-destination trip flow matrices. I use a static traffic assignment algorithm to assign vehicle trips to the road network and calcuate congestion levels, which is described in Chapter 6.

Chapter 2

PROFITABILITY

Zoning changes are a necessary but not sufficient condition to promote development of multifamily housing. For that multifamily housing to be built, it also has to be profitable for a developer to construct. In this chapter, I develop simulations of developer decisionmaking to understand where and under what circumstances development of multifamily homes on single-family lots could occur, given a change to zoning.

Models of urban land development have been applied for decades. Early models operated at an aggregate level, identifying the zones where development of various types was likely to occur (*e.g.* Waddell et al. 2003). Others have used an abstract grid to develop cellular automata models to predict development based on development of adjacent cells (*e.g.* Clarke, Hoppen, and Gaydos 1997; White, Engelen, and Uljee 2015).

More recent models identify individual parcels for development. Some models use a random-utility formulation for this (e.g. Waddell et al. 2010). However, this formulation does not explicitly model development profitability; it only identifies parcels likely to be developed using arbitrarily-scaled utility functions, rather than the actual costs that developers use to make decisions. Real-world developers generally perform a "pro-forma" analysis before undertaking a project, which is a rough calculation of the profitability of the project based on acquisition costs, construction costs and expected rents (Miles, Netherton, and Schmitz 2015, 178). More recent work on developer location choice explicitly simulates this process (*e.g.* Waddell 2013, 28–29; Johnson et al. 2018; Stiphany and Wegmann 2020).

I also simulate these developer pro-forma processes, using the most detailed information on costs and cash flow available. Previous work has been based on per-square-foot costs (*e.g.*

Johnson et al. 2018). Given the small size of the developments I intend to model, a more scale-appropriate modeling methodology is based on specific buildings, rather than entire developments. While it may be appropriate to characterize a large development solely by type of use and square footage, the same is not true of small developments that might take place on a single-family home lot, where the value of a marginal square foot varies greatly depending on whether it affects the total unit count on the property.

This work most closely follows Monkkonen, Carlton, and Macfarlane (2020), who evaluated the effects of a shift in statewide zoning in California to allow fourplexes on all lots. My analysis differs from that of Monkkonen et al. primarily because I simulate several scenarios for zoning change in my results. My prototype building approach additionally appears to be more detailed than Monkkonen, Carlton, and Macfarlane, particularly in terms of space constraints within parcels, though since they rely on a proprietary software product it is difficult to directly compare their methods to mine.

2.1 Modeling Approach

To address concerns about the appropriateness of square-foot-based models for small-lot development, I introduce the concept of prototype buildings—a small number of building plans, with associated cost estimates, for small multifamily structures. These prototype buildings also reduce the data requirements for running the model, as the costs for the prototype buildings can be calculated with off-the-shelf construction cost estimating manuals.

2.1.1 Building Fit Evaluation

An important aspect of determining which if any prototype buildings will be profitable on a particular parcel is determining which buildings could even be feasibly built on the parcel—as a large multifamily building might simply not fit on some small lots. Moreover, not all of the area of a lot is "buildable," as most municipalities have setback requirements for how close a building can be built to the property line. I assume a 1.524m (5 foot) side and rear setback for all parcels.

Front setbacks are considerably more variable than side and rear setbacks, so assuming a blanket value for them across the region is likely to bias results. For instance, just in the City of Los Angeles, the required front setbacks in residential zones range from 0 to 45 feet (Conway, Owens, and Gomba 2018). Unfortunately, such data are not available for all of Southern California's municipalities. Instead, I estimate the setbacks based on the existing buildings, under the assumption that some buildings in each neighborhood are built up to the setback. This is not perfect, particularly in neighborhoods where setbacks have changed over time, but it should approximate the correct value.

To do this, I retrieved street centerlines from OpenStreetMap (OpenStreetMap contributors 2020), parcels from the Southern California Association of Governments (Southern California Association of Governments 2018), and building footprints from the Microsoft Building Footprints dataset (Microsoft 2018). I computed the distance of each building to every street within 30m. I then grouped the buildings by parcels, and for each street segment (i.e. block of street), I estimated the setback as the 20$^{\text{th}}$ percentile of setbacks observed on each parcel within 30m. I used the 20$^{\text{th}}$ percentile because developers generally are only required to build at least as far back as the setback, not exactly that far back.

To compute buildable area, for each parcel I create a 1.524m interior buffer to account for the side and rear setbacks, and then subtract a buffered street segments layer, where each street segment is buffered by its computed setback. The inputs and results of this operation are shown in Figure 5, for a gridded street pattern in San Gabriel, CA, and a curvilinear pattern in Calabasas, CA. Parcels are shown in gray, buildable areas in green, and buildings and street networks in black. All geoprocessing was implemented in PostGIS 3 in a Postgres 12, running on a dual-core Intel i7 Macbook Pro with 16 GB of RAM and a 1TB solid-state disk, and was orchestrated using an `sqmake` workflow (Conway 2020).

Figure 5. Computation of Front Setbacks Based on Street Geometry and Building Footprints

I then need to determine which of my prototype buildings will fit in the buildable area on each property. Determining if an arbitrary polygon can fit into another polygon is a relatively complex problem (B. Chazelle 1983). I make several simplifying assumptions to make a simple and fast algorithm for determining if each of my prototype buildings will fit in the buildable area of each parcel. First, all of my prototype buildings are rectangular with integer meter dimensions. Secondly, I do not allow arbitrary rotations, but only rotations in 15 degree increments. Third, I do not allow arbitrary translation but quantize translations to the nearest meter. For each irregular parcel polygon, I then rasterize the polygon at a 1-meter resolution. For each pixel within the polygon, I evaluate whether a rectangle the size of the building would fit if its upper-left corner was at that pixel. If the building fits within the polygon, then its upper left corner must also be a pixel within the polygon, and by checking every pixel I can determine if the building fits in any position. I repeat this analyis for 0, 15, 30, 45, 60, and 75 degree rotations of the parcel, and for 0 and 90 degree rotations of the building (since rotating the building is as simple as reversing the dimensions). Together, these yield every possible relative rotation of the building and parcel from 0 to 165 degrees, in 15 degree increments. Since rectangles are symmetrical, there is no need to compute the remaining rotations, 180 to 345, as they are identical to the rotations computed so far.

Figure 6 shows the computation for how each of my prototype buildings did or did not fit on a particular parcel. Each row represents a different rotation of the buildable area of the parcel, and the columns represent different building dimensions and different rotations of the buildings. Single-family homes and duplexes are combined because my protypical single-family home and duplex have the same footprint. The rasterized buildable area is shown in maroon, and the building in yellow, if the building fits within the buildable area of the parcel. When the building can fit in multiple configurations for a particular rotation, only one is shown. When the building does not fit within the buildable area, the buildable area is shown in gray. The single-family home/duplex can fit in any rotation on this parcel, while the threeplex can fit in only three (Figure 6c,β,θ), and the sixplex will not fit at all. Thus this parcel can support the first three building types, but not the sixplex.

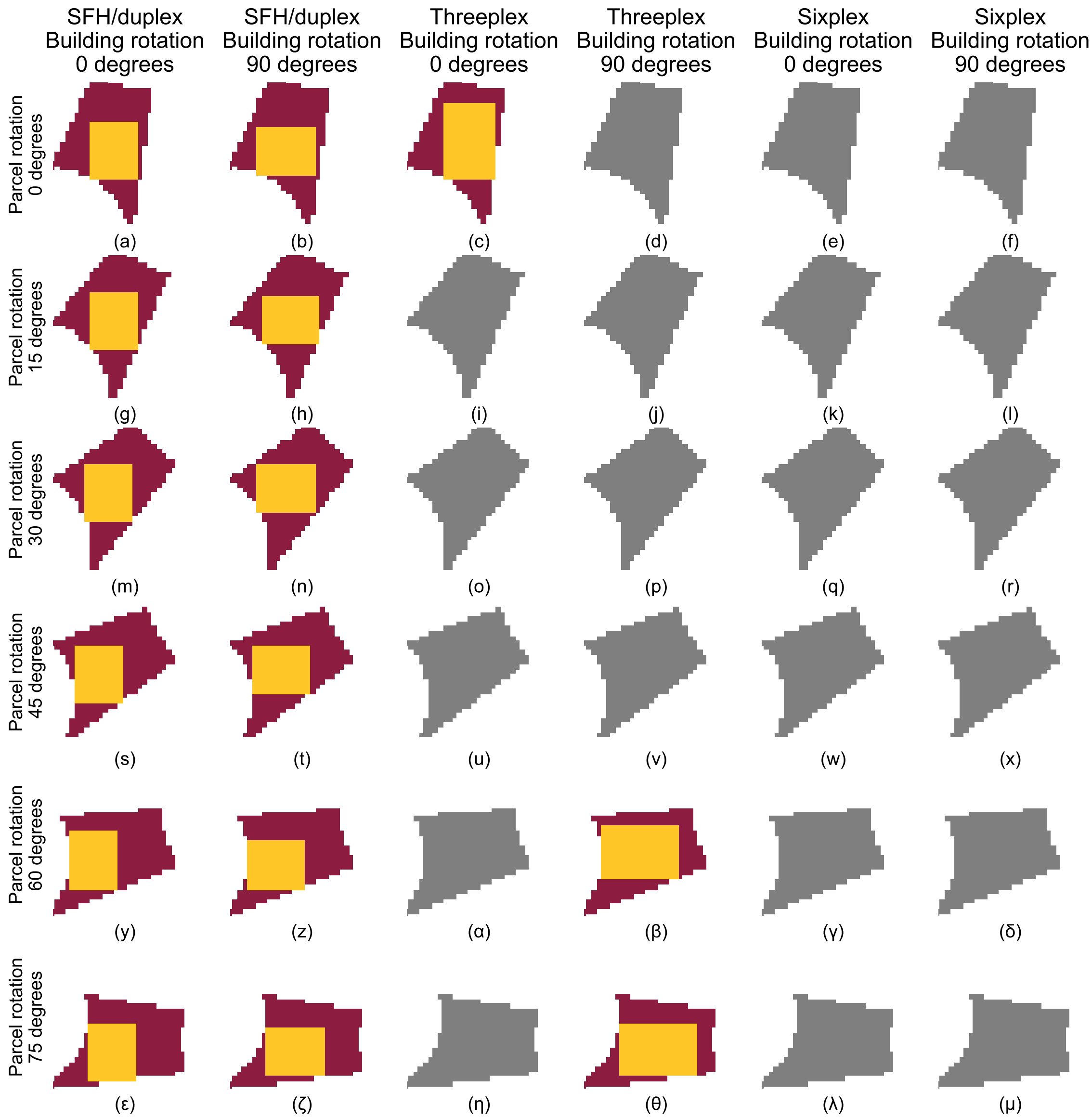


Figure 6. Evaluating Which Buildings Fit on Which Parcels

I implemented this algorithm in Python, using the `shapely` (Gillies 2020), `rasterio` (Mapbox 2018), and `numpy` (Harris et al. 2020) packages. In order to speed computation, the core algorithm was just-in-time compiled from Python to machine code using `numba` (Lam, Pitrou, and Seibert 2015). It took approximately 2.5 hours to compute whether each of the three footprints would fit on each of the 4.75 million parcels in Southern California, on the aforementioned 3.1 GHz, 2-core Intel i7 system. Computations were executed in parallel. Building fit metrics for all four prototype buildings for the area of San Gabriel, CA previously shown are displayed in Figure 7.

This approach is significantly more complex and computationally-intensive than simply comparing the area of the buildable area to the area of the desired building. However, the simpler method may identify parcels that are oddly shaped as being suitable for a building. For instance, the parcel shown in Figure 8 is large enough that, by area, it could accomodate any of the building types in this chapter. However, none of them will actually fit on this irregularly-shaped parcel. Overall, 13% of the parcels considered in this research are large enough in terms of buildable area to accomodate one of my prototype buildings, but do not have a section of buildable area that the prototype building could actually fit in.

That said, many of the parcels that the area-based approach finds feasible and my approach does not are relatively regularly shaped and probably could fit such buildings, if they were reconfigured slightly (for instance, to be longer and narrower). In future research, including more alternative footprints for each type of building is warranted to close this gap.

2.1.2 Prototype Buildings

The core of the approach used in this chapter to evaluate building feasibility is the "prototype building." This is the design of a building, with dimensions and cost of construction. Using a small number of prototype buildings allows more detailed cost estimations than is

(c) Sixplex

Data: SCAG, OpenStreetMap contributors, Microsoft

Figure 7. Which Parcels Can Support Each Prototype Building in San Gabriel, CA

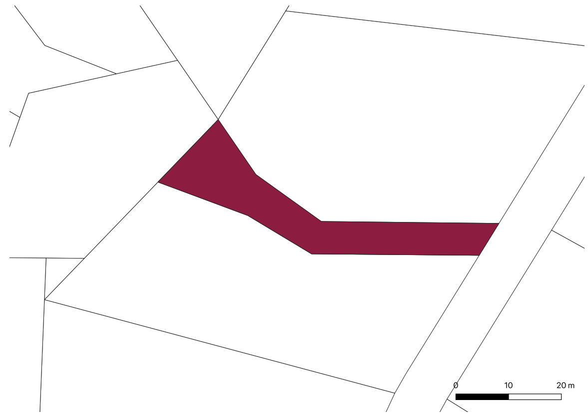


Figure 8. An Irregularly-Shaped Parcel that an Area Comparison Finds Suitable for a Duplex, Threeplex, or Sixplex, but that My Building Fit Algorithm Does Not

possible with square-foot-based formulae, using construction-industry cost estimating manuals. In this project, I estimate the cost of a variety of prototype buildings using the RSMeans Online Square Foot Estimator, with 2020 Q4 data (RSMeans 2020).[2] The prototype buildings I estimate the cost for are shown in Table 1. They include a duplex, a threeplex, and a sixplex. They also include a single-family home, as one reaction to a hot housing market might be to produce larger housing, a trend known as "mansionization" (Monkkonen, Carlton, and Macfarlane 2020).

The single-family home and the duplex are very similar; they have the same number of stories and the same footprint. The key difference is that the duplex has two smaller units in the same footprint, but the interior square footage, garage spaces, and so on, are the same. The threeplex has a larger footprint, understandably. I assume these prototype buildings are all built like the typical single-family home—stickbuilt wood-frame construction covered in stucco. The sixplex is larger and significantly more expensive, as it assumes commercial construction rates, cement siding, and six subterranean parking spaces.

[2] I considered using 2020 base data rather than Q4 in case COVID-19 had caused construction demand and prices to change, but this was not the case in the RSMeans data released in 2020 Q4.

Table 1. Prototype Buildings for Which Construction Costs are Estimated

Building	Units	Footprint (meters)	Stories	Beds	Baths	Garage spaces	Construction
Single-family home	1	12×10	2	4	2.5	2	Wood frame and stucco
Duplex	2	12×10	2	2× 2	2×1.5	2×1	Wood frame and stucco
Threeplex	3	16×11	2	1×1, 2×2	1×1, 2×1.5	3	Wood frame and stucco
Sixplex	6	21×10	3	1×3, 3×2, 2×1	1×2, 3×1.5, 2×1	6 subterranean	Wood and fiber cement

Construction costs were estimated for each of these prototype buildings, using the RSMeans Square Foot Estimator model with costs for construction in Los Angeles.[3] Architectural fees were estimated to add 10% on top of the estimated costs, within the general range of architectural costs for new residential construction (HomeAdvisor 2020), and an additional 10% contingency budget was added. The single-family home, duplex, and triplex were assumed to have standard residential labor construction costs, while the larger sixplex was estimated as commercial construction with union labor. Sometimes zoning relaxations to promote housing construction include "prevailing wage" requirements (*e.g.* California's SB 35; Wiener 2017), which could raise labor costs above these assumptions.

Parking can add significantly to the cost of construction. The single-family and duplex units include the cost of construction of two garage parking spaces, while the threeplex as-

[3]Due to licensing restrictions, the calculated costs cannot be reproduced here.

sumes the construction of three garage parking spaces. The sixplex assumes six underground parking spaces. The cost of these is not estimated by RSMeans, but a large apartment project in Los Angeles spent $21,000 per parking space in 2001 (Shoup 2011, p. 148). The Consumer Price Index for large cities in the West indicates inflation of 52% since 2001 (Bureau of Labor Statistics 2020),[4] so I estimate the cost of subterranean parking as $31,975 per space. This is likely an underestimate due to economies of scale present in the larger project.

In addition to construction costs, developments are subject to a number of permitting and impact fees imposed by municipalities. While they can be efficient if they assign the marginal cost of city services to new development, they can also reduce housing supply (Been 2005). Unfortunately, there is a lack of transparency in how these fees are calculated, and it may not even be possible for developers to estimate these fees in advance (Mawhorter, Garcia, and Raetz 2018). Thus, including them in cost estimates is difficult. However, Mawhorter, Garcia, and Raetz (2018) developed estimates of development fees for prototypical single-family and multifamily developments in the City of Los Angeles, of $32,127 and $11,746 per unit, respectively. While these developments are much larger than the developments I am simulating, these fees represent the best available estimate, and thus I apply them to all projects in the region.

Building a new building on an existing single family home lot requires demolishing any existing single family home on the lot. Demolition costs were estimated for a 1600 square foot single-family home on a 6 inch thick, 1800 square foot (to account for a garage) concrete slab using RSMeans software. Additionally, disposal costs for the debris generated were es-

[4] The producer price index for construction is not used because it was restarted with new categories in 2014 so does not have a continuous series for this time period.

timated based on current rates in Southern California.[5] This estimated value was applied to any lot that had an existing structure.

Landscaping costs represent another cost component in producing a finished development. I estimate the cost of landscaping by creating a cost estimate for a yard 25% covered in hardscaping (driveways, walkways, patios, etc.) and 75% covered in irrigated lawn using the RSMeans software. I convert this estimate to a per-square-foot cost and multiply it by the number of square feet on the parcel not covered by the building.

2.1.3 Building Value

Estimating the value of the current land use as well as future land uses requires estimating both the income from the property (in the form of rents) as well as the expenses associated with the property. Construction costs were detailed in the previous section; in this section, I discuss estimating the rental value (Section 2.1.3.1) and the ongoing operating costs (Section 2.1.3.2) for all land uses. I then combine these estimates into estimations of present value for new structures (Section 2.1.3.3), existing structures (Section 2.1.3.4), and vacant lots (Section 2.1.3.5). Finally, I present sensitivity tests of the value model in Section 2.1.3.6.

[5] A 6 inch thick, 1800 square foot concrete slab has a volume of $1800 \times 0.5 \times 27^{-1} = 33.333$ cubic yards. The edges of the slab are thicker; assuming a 170 foot perimeter and 2 foot thick by 18 inch wide thickened edge, the additional volume is $170 \times 1.5 \times 1.5 \times 27^{-1} = 14.167$ cubic yards, for a total of $33.333 + 14.167 = 47.5$ cubic yards. When broken up, concrete roughly doubles in volume, and weighs approximately one short ton per cubic yard (Dumpsters.com 2021). At current rates at the Simi Valley Landfill, each ton of concrete costs $32.40 for recycling (Waste Management 2020), meaning that foundation debris disposal from this building would cost $3,116. The Federal Emergency Management Agency estimates that an 1800 square foot house (again, including the garage) will generate 360 cubic yard of debris (exclusive of vegatation), weighing 0.5 short tons per cubic yard, translating to 180 tons of debris (Federal Emergency Management Agency 2007). The Simi Valley Landfill charges $74/ton for disposal of mixed construction and demolition waste (Waste Management 2020), costing $13,320 for disposal.

2.1.3.1 Monthly Rental Income

To estimate the value of both new and existing structures for use in determining the most profitable use, a simple hedonic model of monthly rent was developed. I chose to develop a model of monthly rent, rather than sale values, because sales of multifamily properties are likely to be somewhat rare. I thus assume for the purposes of determining profitability that all units are rented. Data to estimate the model came from the 2013–2017 5-year American Community Survey microdata distributed by the Integrated Public Use Microdata Sample (Ruggles et al. 2019). Unfortunately, the data available for estimating a hedonic model in IPUMS is somewhat limited; I used dummy variables for properties built in 2000 or later, a dummy for a single family home, and dummies for 1, 2 and 3 or more bedrooms (relative to studio). Additionally, I included fixed effects for the 124 Public Use Microdata Areas in Southern California, which is the most granular geography available in IPUMS, to control for unobserved neighborhood features. Model estimation results are in Table 2; fixed effects are displayed in Figure 9. The dependent variable is the natural logarithm of monthly rent. Since the model has no constant, the fixed effects for each location represent a location-specific constant, and introduce location-based variation in rent of up to $1,000.

The model fits relatively poorly as hedonic models go, with an R^2 of 0.36, likely due to the small number of attributes available in the IPUMS data. In future work I hope to improve this model with a better dataset.

I use this model to estimate the monthly rent that would be garnered by each prototype building on each parcel, by estimating the monthly rent for each unit in the building and summing. Since this is a log-linear model, the estimated rent is $e^{XB+\frac{s^2}{2}}$, where s^2 is the

Table 2. Hedonic Regression to Predict ln(Monthly Rental Value) for Constructed Buildings

	Coefficient	Std. Err.	t	p
Built in 2000 or later	0.13***	0	26.94	0
Single-family home	0.12***	0	40.5	0
1 bedroom	0.09***	0.01	17.35	0
2 bedroom	0.37***	0.01	72.54	0
3 bedroom	0.57***	0.01	96.45	0
PUMA fixed effects	6.17–7.33 (see Figure 9)			
R^2	0.36			
Adj. R^2	0.36			
Sample size	124,258			

Standard errors are heteroskedasticity-robust (HC3) (Angrist and Pischke 2009, 300)
.: $p < 0.1$, *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$

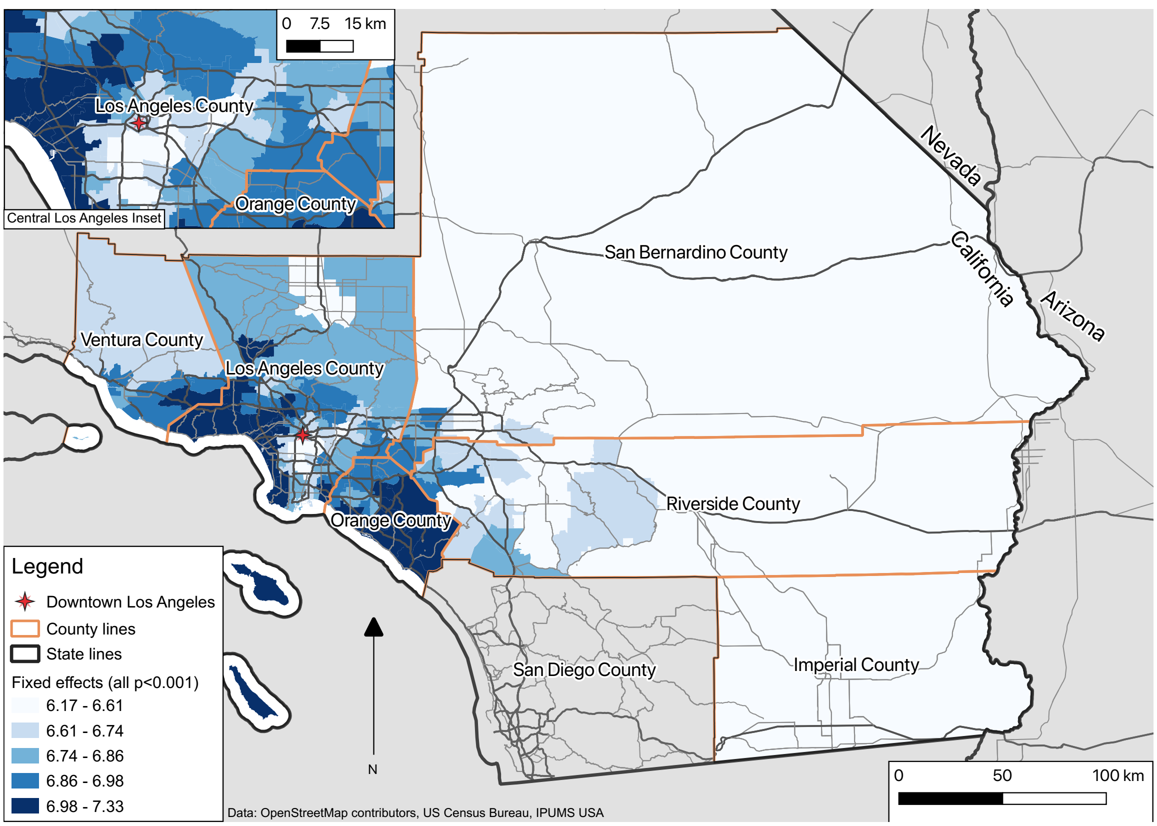


Figure 9. PUMA-Level Fixed Effects from Model

residual variance and corrects for the asymmetry of the lognormally distributed error term (Giles 2013).

The rents estimated by this model are lower than expected for Los Angeles County, because the rents in the IPUMS are lower than one would expect. The median rent in the IPUMS in the Los Angeles—Long Beach—Anaheim metropolitan statistical area is $1,409/month, while Zillow estimates rents in the same area at $2,545/month (Zillow 2020). Intuitively, the Zillow numbers are much closer to lived experience renting in large US metropolitan areas. I estimate a constant scaling factor of 1.8 and multiply all predicted rents from the hedonic model by this value to better represent realistic rents, while preserving the heterogeneity introduced by the hedonic model.[6]

I also estimate the monthly rent of the existing single-family homes, since not demolishing the home is always an option. I retrieve data on existing properties in the SCAG region from the Zillow ZTRAX Assessment database (2018 edition), which aggregates public property assessment data from across the US (Zillow 2019b). I extract single-family homes and vacant residential parcels from this dataset and join them to the aforementioned SCAG land use dataset, which is parcel-based. I use this hedonic model based on the IPUMS to estimate the rents for these existing single-family homes.[7] The vast majority of the parcels join correctly between datasets, but 1.3% do not, with some county-level variation (Table 3). These parcels are discarded.

[6] The Zillow number is not simply the median rent, but the mean of the middle quintile (Zillow 2020); the adjustment factor is computed by calculating the same statistic from metropolitan Los Angeles IPUMS data and comparing to the Zillow estimate.

[7] Assessment data from Imperial County shows almost all single family homes having 0 bedrooms. In future work I intend to impute more realistic bedroom counts. However, since Imperial County is remote and sparsely populated, these missing data do not affect the more urbanized areas of the SCAG region where development is likely to be more profitable.

Table 3. Percent of Properties in ZTRAX Dataset Matching SCAG Dataset

County	Total single-family parcels in ZTRAX	Unmatched single-family parcels in ZTRAX	Percent unmatched
Los Angeles	1,522,112	8,688	0.6%
Orange	597,562	12,844	2.1%
Riverside	596,629	18,280	3.1%
San Bernardino	608,070	3,925	0.6%
Ventura	175,716	429	0.2%
Imperial	34,083	860	2.5%
Total	3,534,172	45,026	1.3%

2.1.3.2 Operating Costs

Developers estimate profitability by comparing annual net income over a certain time period (discussed below) to costs. To convert the estimated monthly rent from the previous section to annual net operating income i, I subtract 45% for maintenance, taxes, etc. (Barron 2012) and multiply by 12 to annualize:

$$i = 12(1 - m)p(1 - v) \tag{2.1}$$

where i is annual net operating income, p is the monthly rent, m is the proportion of monthly rent needed to cover maintenance, taxes, etc., and v is the assumed vacancy rate. All variables used in this chapter are defined in Table 4. In future work I plan to disaggregate this and in particular to investigate the effects of California's Proposition 13, which can cause large tax discontinuities when properties are sold or significantly redeveloped, and may thus disincentivize development.

2.1.3.3 Net Present Value for New Buildings

When investors and developers are considering whether a project is financially feasible, they compute a net present value based on discounted operating income flow for a term, for instance ten years,[8] and discounted expected sale values at the end of that term (Miles, Netherton, and Schmitz 2015, 212).

I compute the present value of the property using a discount rate, an appreciation rate, and a capitalization rate for when the property is eventually sold (212). For existing structures, I sum the discounted net operating income for years 0–9, and the discounted expected sale value in year 10. For existing properties, the full calculation is

$$v_{ex} = \left[\sum_{t=0}^{9} i(1+a)^t(1+r_{ex})^{-t}\right] + i(1+a)^{10}(1-f)(1+r_{ex})^{-10}c^{-1} \qquad (2.2)$$

where v_{ex} is the net present value, a is the annual rate of rent and sale value appreciation, r_{ex} is the discount rate for existing structures, c is the capitalization rate, i is annual rental income from the property, and f is the transaction costs of selling the property.

The sum term in (2.2) represents the discounted net operating income for the first 10 years of ownership, and the remaining portion the sale value. I estimate the annual rate of appreciation to be 4.80%. Using the Zillow home value index for single family homes, I computed the year-on-year appreciation for each year from 2010 to 2017 and took the mean. I compute the capitalization rate by comparing estimated annual income from (2.1) with sale values for properties that sold in 2013–2017 (using the Zillow Home Value Index for single family homes in the Los Angeles metropolitan area to convert values to 2017 dollars (Zillow 2021)). The median value is 4.4%, which I use in my calculations; I use the median rather

[8] Ten years is a common term over which developers estimate income (Miles, Netherton, and Schmitz 2015); while the exact term chosen should not significantly affect the estimated value of the property, in future work I intend to conduct sensitivity tests of this term.

than the mean due to significant outliers. I assume transaction costs are 9% of the sale value (Zillow 2019a).

The final term is the discount rate. This is less well defined. Theoretically, it should be somewhat higher than interest rates for loans for new homes, as the homeowner takes on somewhat more risk than the bank does when investing in a home. Rather than exogenously assuming a discount rate, I consider r_{ex} a free parameter and calibrate (2.2) so that the median net present value to sale value ratio for homes that sold in 2013–2017 is 1. This results in a discount rate of 6.5%, slightly higher than interest rates over the last decade (Freddie Mac 2021).

2.1.3.4 Net Present Value for Existing Buildings

The net present value of redeveloped parcels is computed slightly differently, to account for the time they are under construction. I assume construction takes two years, so net operating income (loss) for the first two years is assumed to be half the cost of constructing the project, with no gross receipts. Net operating income for remaining years, and sale value at the end of the 10-year period, is calculated as with single-family homes. Thus, the calculation is as follows:

$$\begin{aligned} v_{new} = \frac{b}{2} + \frac{b}{2}(1 + r_{new})^{-1} + \left[\sum_{t=2}^{9} i(1+a)^t (1+r_{new})^{-t} \right] \\ + i(1+a)^{10}(1-f)(1+r_{new})^{-10} c^{-1} \end{aligned} \tag{2.3}$$

where b is the (negative) construction cost, v_{new} is net present value, r_{new} is the discount rate for new construction, and all other variables are as used in (2.2).

I use a different discount rate for new construction than was used for existing construction, to account for the increased risk of new construction relative to purchasing an exist-

Table 4. Nomenclature

Variable	Description
v_{ex}	Net present value of an existing home
v_{new}	Net present value of a new structure
t	Year since start of model
i	Monthly rental income
p	Monthly rent estimated by hedonic model (Table 2)
m	Maintenance and taxes as a proportion of rent
v	Assumed vacancy rate
r_{ex}	Discount rate for existing property
r_{new}	Discount rate for newly-constructed properties
a	Annual property appreciation
c	Capitalization rate
f	Transaction costs

ing structure. Construction loans currently have interests rates of 6%–10% (ValuePenguin 2020); I split the difference and assume an interest rate for new construction of 8%. As before, not all risk to construction is likely capitalized into interest rates. Garcia (2019) assumes a 3% "developer fee" on top of interest rates to account for risk and developer overhead, so I add this to the interest rate and assume a discount rate for new construction of 11%.

To determine the most profitable use, I simply take the use that has the highest net present value among existing and all new uses.

2.1.3.5 Net Present Value for Vacant Lots

Vacant lots are handled differently, because they generally do not accrue rent (unless used for parking), and thus (2.2) would evaluate to zero and these properties would always be redeveloped as long as construction costs did not exceed rents. However, the value of vacant lots is defined almost solely by option value, which the net present value calculation does not incorporate. Thus, I compute the net present value for the existing single-family-zoned vacant

Table 5. Hedonic Model for ln(Vacant Property Sale Price)

	Coef	Std. Err.	t-value	p-value
Constant	7.83***	0.06	138.72	0.0
ln(lot area) (square meters)	0.29***	0.01	41.75	0.0
Transaction year: 2008	-0.05	0.04	-1.1	0.27
Transaction year: 2009	-0.47***	0.05	-10.12	0.0
Transaction year: 2010	-0.64***	0.04	-14.35	0.0
Transaction year: 2011	-0.79***	0.05	-17.14	0.0
Transaction year: 2012	-0.68***	0.04	-16.66	0.0
Transaction year: 2013	-0.41***	0.04	-11.56	0.0
Transaction year: 2014	-0.39***	0.03	-11.1	0.0
Transaction year: 2015	-0.41***	0.03	-11.89	0.0
Transaction year: 2016	-0.34***	0.03	-11.71	0.0
Transaction year: 2017	base			
Not shown: PUMA fixed effects (see Figure 10)				
Sample size	19,094			
R^2	0.53			
Adj. R^2	0.52			

property using a log-linear hedonic model based on Zillow ZTRAX transaction data (Zillow 2019b), which is presented in Table 5. It includes property area, the year of the transaction (to account for inflation), and fixed effects for most PUMAs (some adjacent PUMAs are merged to ensure all PUMAs have five or more observations, using a Queen's case adjacency matrix computed with PySAL (Rey and Anselin 2007)). The PUMA-level fixed effects are presented in Figure 10. This model is used to estimate the value of each single-family-zoned vacant property if it were to sell in 2017. The net present value for potential redevelopments of vacant property are calculated the same as for non-vacant properties using (2.3).

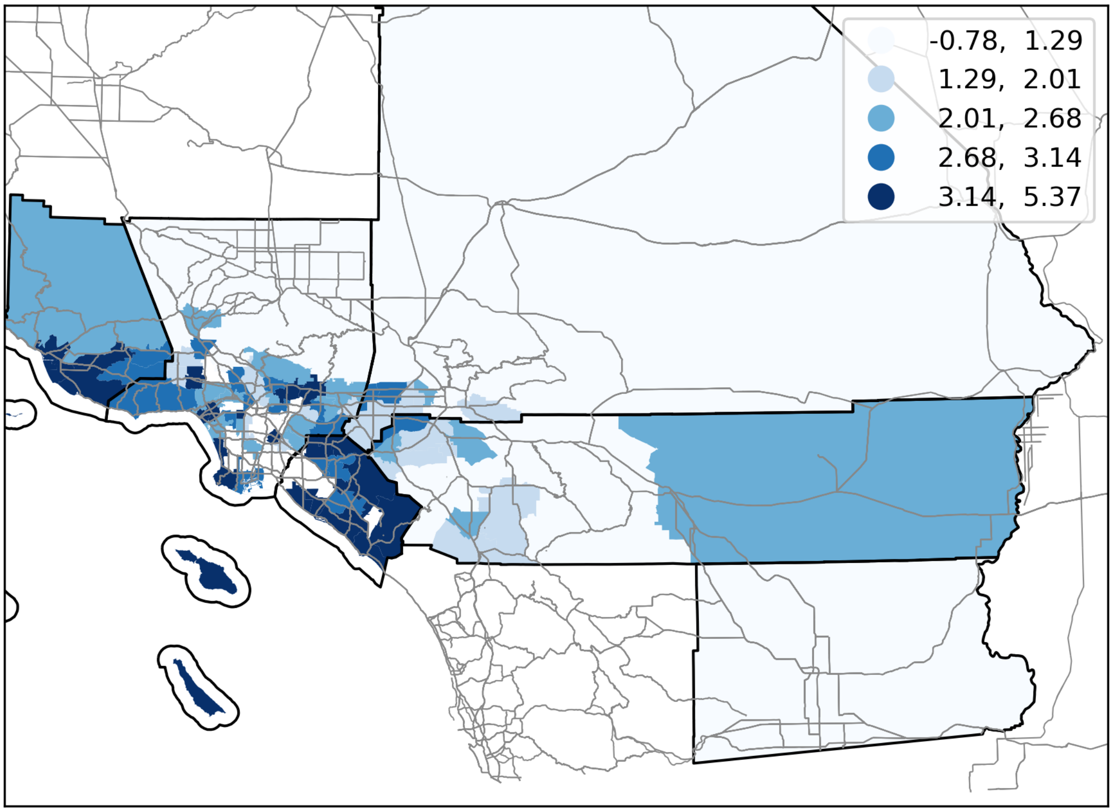


Figure 10. Fixed Effects for Hedonic Model of Vacant Properties

2.1.3.6 Sensitivity Tests

Since the assumed and estimated parameters are somewhat uncertain, I performed sensitivity tests where I varied the values of these parameters to see how the results vary when developers operate under different plausible assumptions (Table 6). The current appreciation scenario uses the values described above, and is the one that I believe to most closely represent the development environment in Southern California, but the other scenarios are plausible. A low appreciation scenario lowers property appreciation to 1.4%, the current Consumer Price Index (Bureau of Labor Statistics 2021), since an increase in housing supply construction may cause rents and thus appreciation to fall. The low discount rate scenario

Table 6. Parameters for Sensitivity Tests of Net Present Value Model

	Discount rate					
	New construction	Existing home	Capitalization rate	Appreciation rate	Operating cost	Contingency
Current appreciation	11.0%	6.0%	4.424%	4.803%	45%	10%
Low appreciation	11.0%	6.813%	4.424%	1.4%	45%	10%
Low discount rate	8.0%	4.125%	4.424%	1.4%	45%	10%
Equal discount rate (8% existing and new)	8.0%	8.0%	4.424%	1.4%	45%	10%
Low operating cost (25%)	11.0%	6.813%	4.424%	4.803%	25%	10%
High construction cost	11.0%	6.813%	4.424%	1.4%	45%	40%

assumes that discount rates closely match current interest rates for new construction loans and consumer loans to purchase existing homes, and could be plausible if the Federal Reserve keeps interest rates low.[9] The equal discount rate scenario assumes that the discount rates for new construction and existing homes are the same, which would be plausible if the government were to provide subsidized loans to encourage housing construction. The low operating cost scenario assumes that maintenance, taxes, etc. only cost 25% of rent, rather than 45% as in other scenarios. Finally, the high construction cost scenario assumes a 30% increase in the cost of construction, plausible if the recent astronomical increases in the costs of building materials (Dezember and Quiros-Guiterrez 2021) continue, or if zoning changes contain "prevailing wage" requirements that result in higher labor costs.

As a check on the results, I compare the estimated net present values for the current appreciation scenario with the sale values of properties that transacted betwen 2013 and 2017, again using Zillow ZTRAX (Zillow 2019b) data adjusted to 2017 dollars with the Zillow Home Value Index for single-family homes (Zillow 2021). The results are shown in Figure

[9] 4.125% is the median interest rate on for home purchase loans in Los Angeles County, computed using the 2019 Home Mortgage Disclosure Act (HMDA) data (Consumer Financial Protection Bureau 2019).

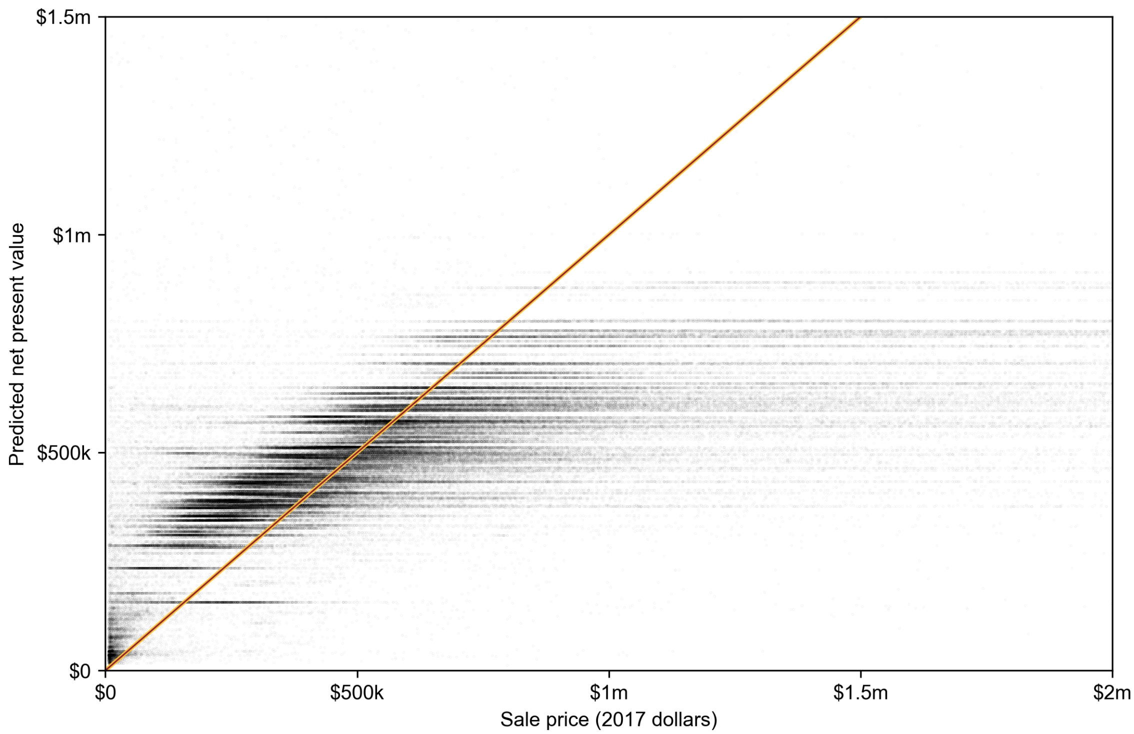


Figure 11. Comparisons of Net Present Values and Observed Sale Prices

11. Net present values track fairly well with observed sale values. The R^2 of the predicted vs actual values is only 0.054, however this is primarily due to many outliers in the sales data. If sales over $2 million are removed, the R^2 is a respectable 0.377—indicating the net present value model does a reasonable job of reproducing sale values, in the same ballpark as the hedonic model of rents it is based on (Table 7).

2.2 Results

Table 7 shows the results of running this model under each of the scenarios described in Table 6.[10] In the current appreciation case (shown in the first line of the table), 96% of single-family-zoned parcels are not redeveloped. The remaining 4% are redeveloped primarily into sixplexes. To accomodate this redevelopment, 58,000 single-family homes are torn down, leading to a net increase of 397,000 housing units.

45% of the redevelopment occurs on vacant lots, rather than lots with existing single-family homes. Therefore, the net increase may be a high estimate, for two reasons. First, I don't account for the possibility that property owners will hold onto land to redevelop later when it may be more profitable; others have addressed this with dynamic models of redevelopment (Murphy 2018). Second, a small amount of the single-family zoned vacant properties are common areas or flood control features that are unlikely to be redeveloped.

Growth is not evenly spread across the region, but is concentrated in two primary regions. First, the west side of Los Angeles and much of Orange County see significant redevelopment. There is also significant redevelopment in relatively less-developed areas in the Antelope Valley north of Los Angeles, and in outlying areas around Palm Springs east of Los Angeles (Figure 12).

The remaining rows of Table 7 indicate that the net present value model is highly sensitive to the assumptions used, with reasonable variations in assumptions leading to forecasts of as low as 5,000 marginal units, or as many as 4,333,000. The geographic distribution of

[10] Due to a technical error, some parcels sizes were underestimated by up to 1 meter. Correcting this error changes the number of profitable marginal units in Tables 7 and 8 by no more than 1.5%, and will be corrected in future research.

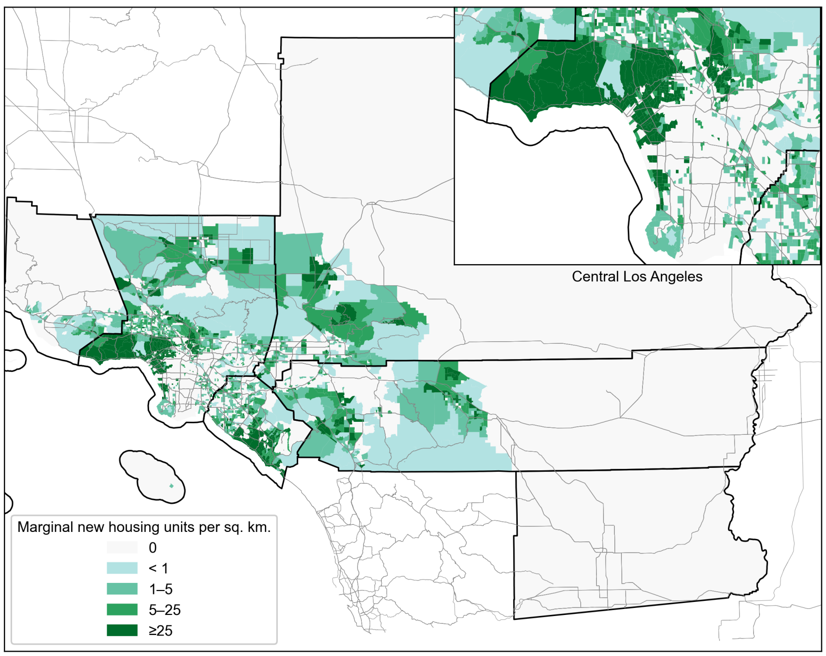


Figure 12. Percentage Growth in Units in the Current Appreciation Scenario

Table 7. Number of New Units Developed Under Various Development Scenarios, in Thousands

	Non-redeveloped parcels	New units: Single-family home	Duplex	Threeplex	Sixplex	Total	Teardowns	Marginal units
Current appreciation	96.2%	0	6	168	282	455	58	397
Low appreciation	99.5%	0	1	39	0	40	2	38
Low operating cost (25%)	66.3%	0	10	415	4,758	5,183	849	4,333
High construction cost	99.9%	0	0	5	0	5	0	5
Low discount rate	96.3%	0	6	162	268	435	54	381
Equal discount rate (8% existing and new)	80.5%	0	6	1,175	884	2,065	495	1,570

development is similar in all scenarios, though development is more intense in the scenarios that produce more units (maps for each scenario are presented in Appendix A).

Increased housing supply does not result in lowered rents in this model, *ceteris paribus*. However, apartments are cheaper than single family homes, so an increase in apartment construction can bring aggregate rents down as more people are able to choose this less expensive option. Housing construction can reduce rents through two mechanisms: a market equilibrium mechanism and a unit mix mechanism. I only model the unit mix mechanism.

2.2.1 Redevelopment Near Transit Stations

Often, upzoning proposals do not constitute a blanket elimination of single-family zoning, but rather upzoning only in specific areas. Often, the locations that are upzoned are proximate to transit infrastructure (*e.g.* Wiener et al. 2019). Thus, in addition to the wholesale elimination of single-family zoning presented above, I also evaluate the effects of restricting zoning changes to only those areas within the High Quality Transit Area as defined by SCAG (Southern California Association of Goverments 2019). These are areas within 1/2 mile of a rail or ferry stop, or a high-frequency bus line. Redeveloping in these areas may promote sustainable transportation, and ease concerns about parking. If single-family zoning were only eliminated in the High Quality Transit Area, the net present value model forecasts a net change of 69,000 units across the SCAG region—much smaller than the number that would be developed if single-family zoning were eliminated everywhere. As with the previous scenarios, the range of possible outcomes varies significantly based on the assumptions of the model (Table 8). Additional units are now concentrated almost exclusively on the west side of Los Angeles and in Orange County, as other areas that were redeveloped in the base model are outside the High Quality Transit Area (Figure 13). This eases gentrifica-

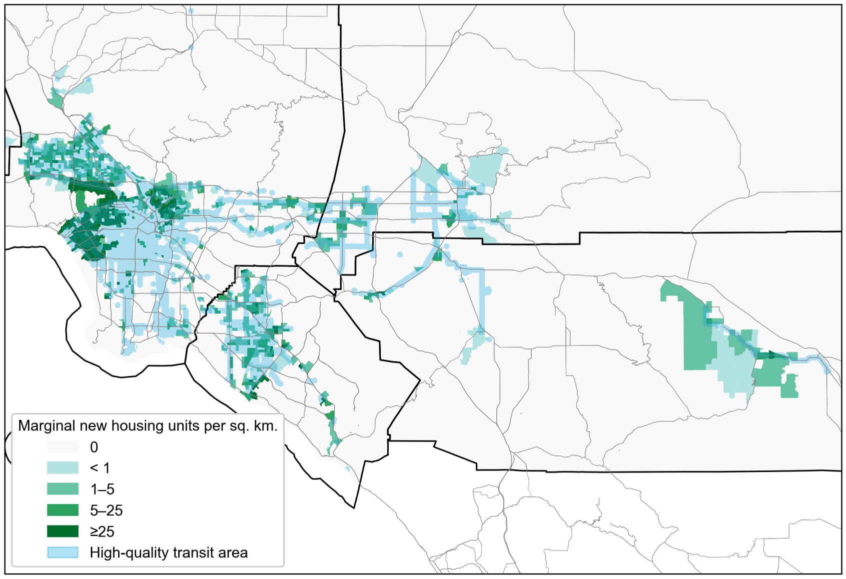


Figure 13. Percentage Growth in Units in the Current Appreciation Scenario, High Quality Transit Area Only

tion concerns, as the lower-income areas that saw significant redevelopment in the previous models are excluded from zoning changes. As before, geographic patterns are similar in all scenarios, although they vary in intensity; maps for all scenarios are presented in Appendix A.

2.3 Conclusion

In this chapter, I detailed a method for using construction industry cost estimating manuals as a data source for land development models. When evaluating the effects of relaxing single-family zoning, this is a useful approach, as it is based on actual buildings, rather than abstract costs and values per square foot. While such aggregate metrics are appropriate when evaluating large developments, my more detailed approach is much more appropriate for small redevelopments of single-family home lots with only a few units.

Table 8. Number of New Units in the High Quality Transit Area Developed Under Various Scenarios, in Thousands

		New units						
	Non-redeveloped parcels	Single-family home	Duplex	Threeplex	Sixplex	Total	Teardowns	Marginal units
Current appreciation	99.3%	0	1	27	56	84	15	69
Low appreciation	99.9%	0	0	4	0	4	0	4
Low operating cost (25%)	91.0%	0	1	80	1,340	1,422	245	1,177
High construction cost	100.0%	0	0	1	0	1	0	1
Low discount rate	99.3%	0	1	26	56	82	14	68
Equal discount rate (8% existing and new)	94.2%	0	1	351	257	609	156	453

While the model presented here is among the most detailed simulations of development in the literature, it still glosses over significant details. The model does not evaluate the topography of the parcel, and may recommend building on slopes in excess of what is allowed by zoning—in particular, this may explain some of the redevelopment the model proposes in the Malibu area. Such lots would also have significant site preparation costs not included in the RSMeans cost estimate. Similarly, the model does not account for lot coverage requirements, which require a certain portion of the lot to be left empty, or restrictions on development in the wildland-urban interaface. These regulations are not likely to be lifted when single-family zoning is relaxed. Restrictions on building on hillsides are important to life safety in Southern California, where earthquakes, mudslides, and other slope failures are somewhat regular occurrences. While lot coverage requirements may be used to exclude housing from a neighborhood, they may also serve a flood-control purpose by limiting the impervious surfaces in an area, and again may not change if single-family zoning is eliminated. The model does not consider the road, electricity, water, or sewer access of a parcel, and may recommend building on parcels that lack these amenities; while this is not impossible, it would increase the cost of development in ways not modeled here.

My choice to use a static model means I cannot forecast the timing of redevelopment, and cannot account for property owners holding off on profitable projects because they believe they will be more profitable in the future. Murphy (2018) uses a dynamic model that accounts for the time component of redevelopment. In future work, considering such a dynamic framework could improve the accuracy of the forecasts, particularly over long time horizons.

California is famous for regulatory delay of housing projects. These delays can add significantly to project costs (Shannon 2015) and are largely not accounted for in this model, except to the extent that they may fall within the two-year development window. Future work could better account for these delays. Some upzoning bills require that complying development be approved through a ministerial process rather than review by a zoning board or public comment, which could reduce the cost and length of delays.

Chapter 3

FORECASTING OF HOME SALES TO HOMEOWNERS AND DEVELOPERS

The previous chapter shows that there could be significant opportunity to cost-effectively redevelop properties in Southern California. However, just because a property can be profitably redeveloped doesn't mean it will be. Properties are generally redeveloped after being purchased by a developer. Thus, for a property to be redeveloped, it first must come up for sale, and then the winning bid must be placed by a developer. In this chapter, I simulate that process. Previous models of redevelopment have largely stopped with evaluating feasibility, without going to the next step of the redevelopment "funnel" of estimating which properties will actually be sold to developers (Monkkonen, Carlton, and Macfarlane 2020).

I simulate the sale process for land in two steps. First, I model the probability that each home will sell in a five-year period using a logistic regression model. Then, I assume the probability that a developer will enter the winning bid for a piece of property. Most property is not sold to developers, but to homeowners or investors who do not plan to significantly redevelop the property. These simulations are detailed in the next two sections.

I assume that changes to zoning will not affect how often homes go on the market, for two reasons. First, transaction costs of moving are very high, and the literature on the decision to move suggests that it is generally not due to changes in land prices (Pyle 1985; Rossi 1980). Second, the increase in value brought by a change in zoning is likely to be stable, as increased development capacity on the land will remain as long as the policy remains in force, meaning that homeowners can sell on their timeline without fear of losing the additional value of the properties. At the margin, owners may decided to sell in order to capitalize on their windfall profit more quickly, but since there is little risk of losing that windfall profit later, this effect seems likely to be close to negligible. Owners of investment properties may respond more elastically to market changes, which is a limitation of this approach. However, only one-quarter of single family homes in the Los Angeles-Long Beach CSA (all counties in the SCAG region except rural Imperial County) are not owner-occupied (US Census Bureau 2019). Owners of vacant property may also be willing to sell to realize profits from a policy change.

While profitability is estimated in the previous chapter without a time component, the probability that a property will sell increases with time. Since I assume a five-year time horizon in the logistic regression model of property sales, the results presented in this chapter and used in further modeling represent expected development over a five-year period after regulations are relaxed.

3.1 Modeling of Single-Family Home Sales

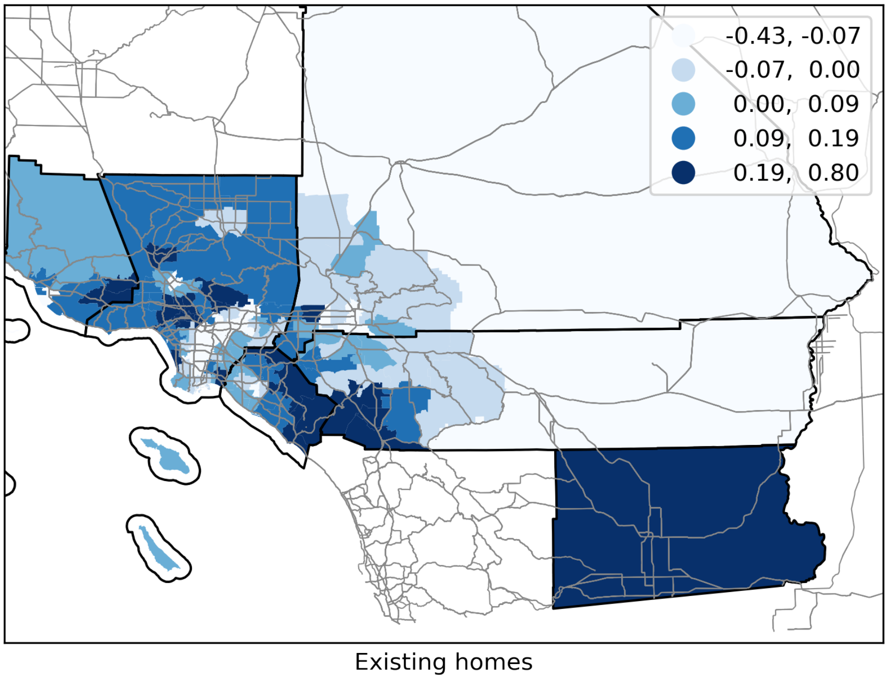
To understand the probability that a property will sell, I build two logistic regression models, one for existing single-family homes and one for vacant properties, based on Zillow ZTrans transaction data, as well as assessment data gathered by Zillow (Zillow 2019b). These models forecast the probability that a parcel will sell, based on the building that is on the property, the size of the property, and its location at the Public Use Microdata Area level. The dependent variable is whether the property sold between 2013 and 2017, meaning the models have a five-year time horizon. This model is not expected to fit particularly well, as many property sales are not motivated by aspects of the homes, but rather of the sellers, but it does provide a baseline probability that accounts for geographic differences in home sales.

The models are shown in Table 9; the fixed effects at the Public Use Microdata Area

(PUMA) level are shown in Figure 14. For the vacant property model, some PUMA's do not contain sufficient properties for estimation, and some PUMA's suffer from perfect prediction (i.e. all vacant properties either did or did not sell), which prevents estimation. To solve these issues, I merged adjacent PUMA's in the vacant property model (using a Queen's case adjacency matrix calculated by PySAL (Rey and Anselin 2007)) so that every merged PUMA contains 5 properties or more and none have perfect prediction. The Pseudo-R^2 of the existing home sale model is very low, 0.01, but this likely represents that there is significant randomness in this process; the model does introduce spatial variation and variation by house size and year built, which provides reasonable baseline estimates for the probability of sale. A probability of sale is all that is needed, since I am not simulating which specific homes will sell, only the probability that a given home will sell and be purchased by a developer. These probabilities are then summed over a neighborhood to produce estimates of redevelopment, without forecasting redevelopment of any one specific home.

Properties with missing data for any of the variables used in the model in Table 9 were dropped from the simulation at this point, with one exception. All property assessment data from Imperial County is missing information on number of bedrooms, year built, and number of bathrooms. Rather than dropping the entire county, these variables are assumed to be zero in the model; the average effect of these omitted variables will be captured in the PUMA fixed effect for the single PUMA that is coterminous with Imperial County. There are no vacant residential-zoned properties in Imperial County present in the SCAG land-use dataset. Imperial County is the most remote, rural, and least populous county in the SCAG region, so this should not materially affect results in the more urbanized parts of the region.

These models are then used to forecast a predicted probability of each parcel being sold in



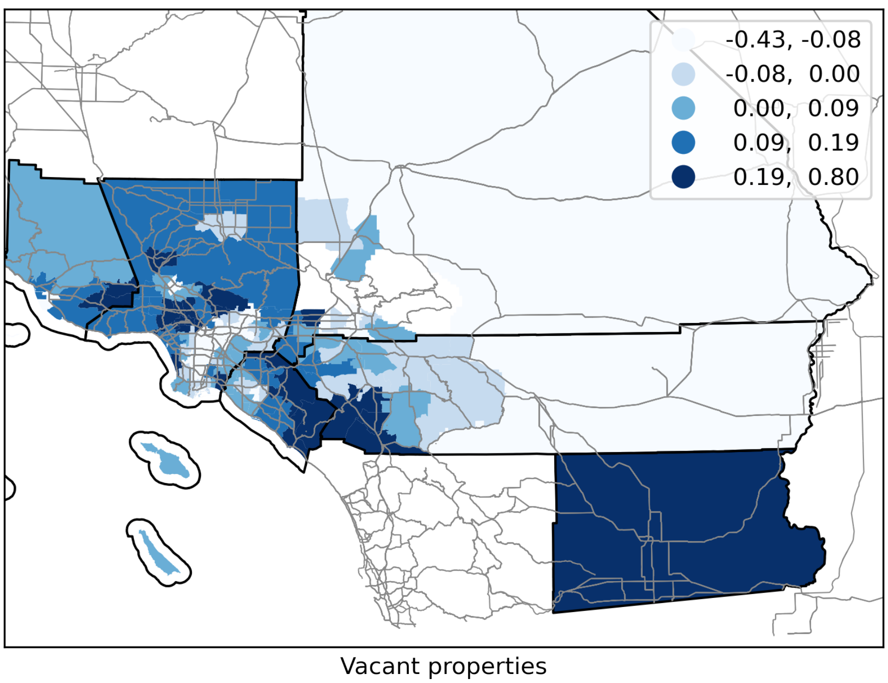


Figure 14. Fixed Effects from Logistic Regression Models of Property Sales

Table 9. Logit Models for Probability that a Property Sold in the Last 5 Years

	Existing homes				
	Odds ratio	Coef.	Std. err.	t-value	p-value
Constant		-0.16	0.01	-15.43	0.0
Bathrooms	1.09***	0.09	0.0	44.31	0.0
Lot area (square meters)	1.0***	-0.0	0.0	-12.72	0.0
Year built (base: 1940 or earlier)					
Built 1941–1960	0.99.	-0.01	0.01	-1.79	0.07
Built 1961–1980	0.96***	-0.04	0.01	-6.21	0.0
Built 1981–2000	1.06***	0.06	0.01	8.92	0.0
Built 2001–present	1.51***	0.41	0.01	49.16	0.0
Bedrooms (base: two bedrooms)					
No bedrooms	0.55***	-0.6	0.03	-23.74	0.0
One bedroom	0.81***	-0.21	0.01	-15.52	0.0
Three bedrooms	1.08***	0.07	0.0	17.31	0.0
Four or more bedrooms	1.09***	0.09	0.01	16.75	0.0
Sample size	2321732				
Pseudo R^2	0.01				
	Vacant properties				
	Odds ratio	Coef.	Std. err.	t-value	p-value
Constant		-1.15	0.04	-28.98	0.0
Lot area (square meters)	1.0.	0.0	0.0	1.65	0.1
Sample size	10932				
Pseudo-R^2	0.11				

****: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$, .: $p < 0.1$*
Not shown: PUMA-level fixed effects

a five-year period, which is combined with the probability that a developer wins the bidding process described in the next section.

3.2 Simulation of the Residential Bid Process

Prospective homeowners and developers compete in the property sale market, and for a property to be redeveloped it generally must be sold to a developer. When properties are

Table 10. Assumed Probabilities of Redevelopment Given Sale, by Profitability

Ratio of net present value of most profitable redevelopment option to net present value of existing use	Assumed probability of redevelopment given sale
<0.95	0
0.95–1.0	0.05
1–1.1	0.1
1.1–1.25	0.2
1.25–1.5	0.3
1.5–2	0.4
≥2	0.5

sold, they are usually not sold to developers, but to other homeowners. In housing markets where redevelopment is highly profitable, however, sales to developers are likely to be much more common. Therefore, I assume that the probability that a property is redeveloped given that it is sold is a step function of the ratio of the net present value of the most profitable redevelopment of the parcel to the net present value of the existing home on the parcel (or, in the case of vacant properties, the option value of the property as estimated by the hedonic model in Table 5 in the previous chapter). The assumed probabilities are shown in Table 10. I assume a small probability of redevelopment when redevelopment is less than 5% below profitability, as developers may face different assumed costs or errors in the net present value estimation may cause profitable developments to appear unprofitable.

While simply assuming probabilities of redevelopment may seem arbitrary and capricious, it is still preferable to not including such a probability in the model at all. Not including the probability implicitly assumes that *all* profitable parcels are redeveloped, something that is clearly not true in the world. Waddell (2013) considers a random sample of candidate properties for redevelopment in each year of his simulation, akin to what I am doing here.

In future work, I plan to perform sensitivity tests of these redevelopment probabilities. In the meantime, I have performed sensitivity tests of many other variables in the model.

If these assumed redevelopment probabilities are too high, for example, results might more closely approximate one of the less aggressive redevelopment outcomes.

3.3 Results

The results of the sales model applied to the different scenarios used in the construction model are shown in Table 11. As expected, the total number of units constructed over a five year period is significantly lower than the number of units that would be profitable to redevelop. However, I believe these values are more realistic. Under the current appreciation scenario, 45,000 marginal units are forecast, but if developers perceived additional risk due to market saturation and assume that property appreciation will follow the consumer price index, only 2,000 would be built. The more optimistic low operating cost scenario forecasts 644,000 marginal units.

To put these numbers in perspective, the six counties in Southern California permitted 47,000 new units in 2019 (US Census Bureau 2018, and author calculations). The current appreciation scenario, then, is forecast to create a shock of approximately one years' worth of development. Since the model of property sale uses data from five years, this estimate is implicitly a five-year estimate, or a 20% increase in housing production in Southern California over the next five years, a significant increase.[11] If the forecast of the low operating cost scenario were to come to fruition, it would more than triple annual housing production, a truly astronomical number.

If development is constrained to only High-Quality Transit Areas, the number of marginal units is significantly reduced, with expected development ranging from <1,000 to

[11] This is a conservative estimate; the 45,000 units from my model represent net production, accounting for property teardowns to provide land for development. In contrast, the building permits survey represents gross production.

Table 11. Redeveloped Parcels Forecast by the Sales model, with Number of Profitable Units from the Construction Model for Comparison

	Non-redeveloped parcels	Single-family home	Duplex	Three-plex	Sixplex	Total	Tear-downs	Marg-inal units	Profitable marginal units
Current appreciation	99.6%	0	0	14	38	52	8	45	397
Low appreciation	100.0%	0	0	3	0	3	0	2	38
Equal discount rate (8% existing and new)	97.8%	0	0	112	134	246	57	190	1,570
High construction cost	100.0%	0	0	0	0	0	0	0	5
Low discount rate	99.7%	0	0	12	29	41	6	36	381
Low operating cost (25%)	95.1%	0	1	36	734	771	128	644	4,333
High-Quality Transit Area only									
Current appreciation	99.9%	0	0	3	10	12	2	10	69
Low appreciation	100.0%	0	0	0	0	0	0	0	4
Equal discount rate (8% existing and new)	99.4%	0	0	33	36	69	17	52	453
High construction cost	100.0%	0	0	0	0	0	0	0	1
Low discount rate	99.9%	0	0	2	6	8	2	7	68
Low operating cost (25%)	98.6%	0	0	7	212	220	37	182	1,177

182,000 units, indicating that much of the forecast redevelopment occurs outside the high-quality transit area. This suggests that broad-brush zoning changes may not affect transport mode choice as much as some have hoped.

3.3.1 The Geography of Redevelopment

As with the results in the previous chapter, development is not evenly spread across Southern California, as shown in Figure 15 (which, for brevity, presents only a sample of the scenarios). The pattern of redevelopment is broadly similar to the patterns of where profitable redevelopment is possible presented in the previous chapter, although new unit densities are understandably lower. Whereas the net present value model showed significant redevelopment in the Antelope and Coachella Valleys north and east of Los Angeles, redevelopment densities are now more varied across the region, with the heaviest redevelopment occurring in wealthy areas such as Santa Monica west of Los Angeles, and in Orange County. Properties in these locations are more likely to sell (Figure 14). While this increases the relative amount of development in these areas as opposed to others, it is not the only or even the primary contributor to the spatial pattern, as properties in these areas are also more likely to be profitable to redevelop (Figure 12 in previous chapter).

This eases some concerns about gentrification in the most urban areas, as development is largely concentrated in higher income areas, such as Santa Monica, Malibu, Orange County, where gentrification is less of a concern. This geographic distribution is consistent with the findings of Monkkonen, Lens, and Manville (2020), who found that both zoned capacity and high rents were required to spur development.

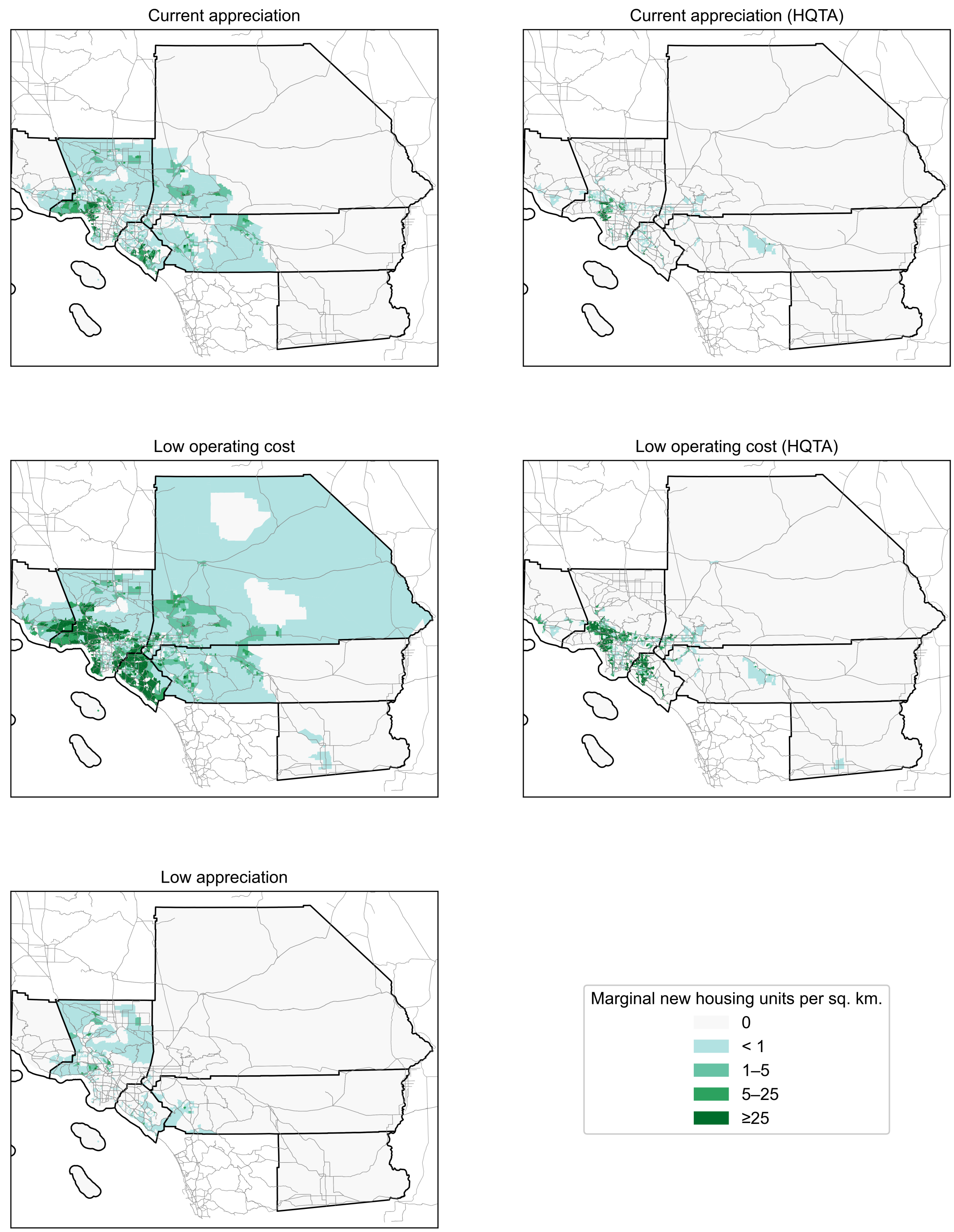


Figure 15. Redevelopment by Location, Including Sales and Redevelopment Probabilities

3.4 Access Implications

Access metrics are gaining ground as a useful way to measure the performance of a transportation system. These metrics measure what the transportation system allows one to access, for instance how many jobs are within a 30 minute transit ride (Committee of the Transport Access Manual 2020). Access measures the quality of a transport system, and is correlated with mode share (Owen and Levinson 2015). Thus, building homes in high-access areas is likely to result in more convenient public transport use for residents of those new units. If, however, development is concentrated on the fringes of the region, significant automobile travel may be induced.

Figure 16 compares the distribution of the number of jobs reachable within 30 minutes, for each of the scenarios, and overlays this on the distribution of access over the existing housing stock. In most scenarios, the new development is slightly more likely to occur that are *less* accessible, suggesting that while a blanket upzoning may promote affordability, such a broad-brush approach may not prompt people to switch to alternative modes of transportation.

However, many zoning changes do not take such a broad-brush approach, but rather target upzonings to specific areas around public transport or other amenities. When upzoning and simulated development is restricted to only the High Quality Transit Area, unsurprisingly, the new development occurs in more accessible locations than the existing Southern California housing stock (Figure 17). This may lead to more positive transport outcomes. However, remember that restricting development to only High Quality Transit Areas results in far fewer homes being built overall, so trying to serve this transport goal may undermine affordable housing goals.

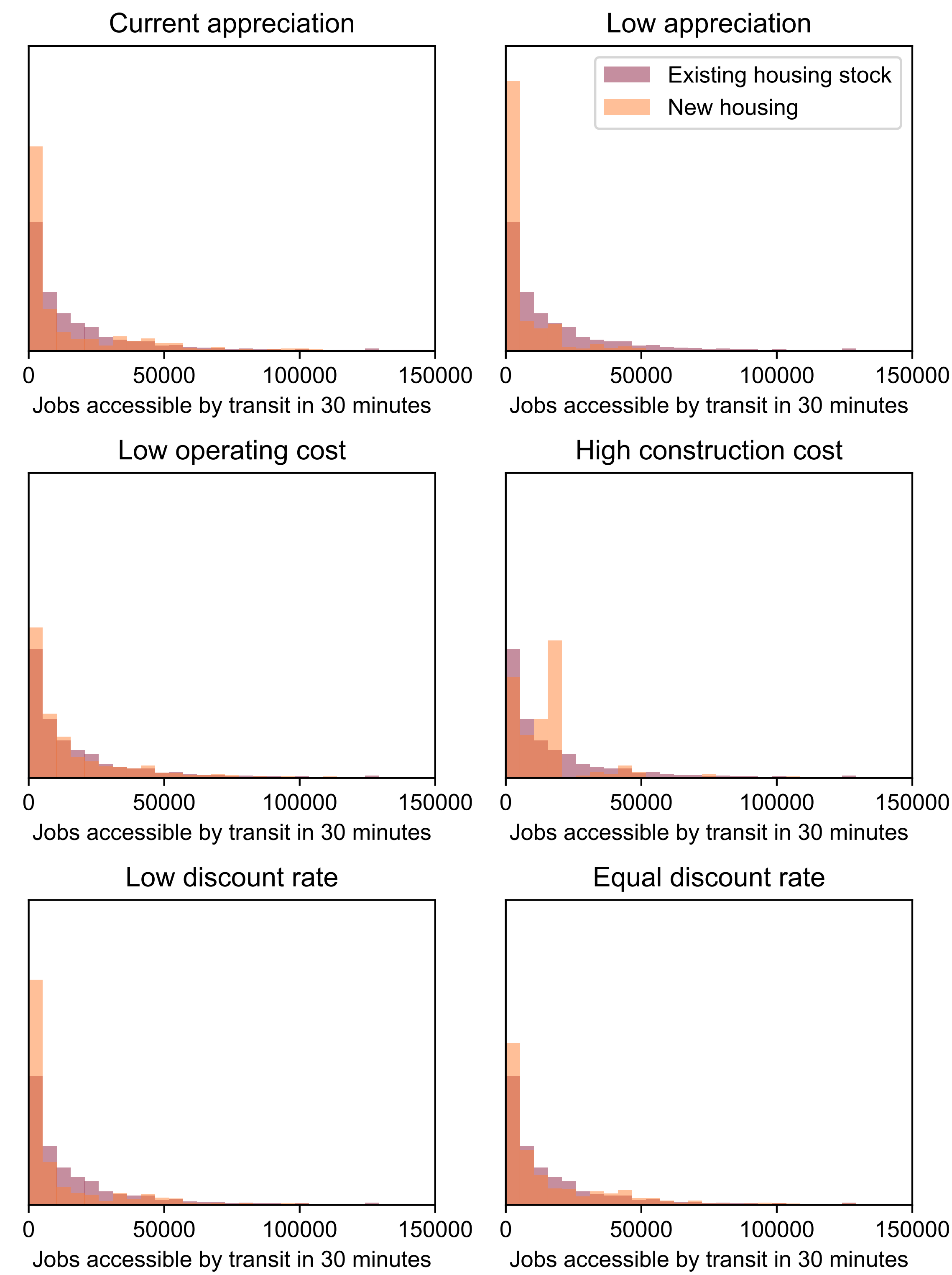


Figure 16. Access Levels around New Development for Various Development Scenarios

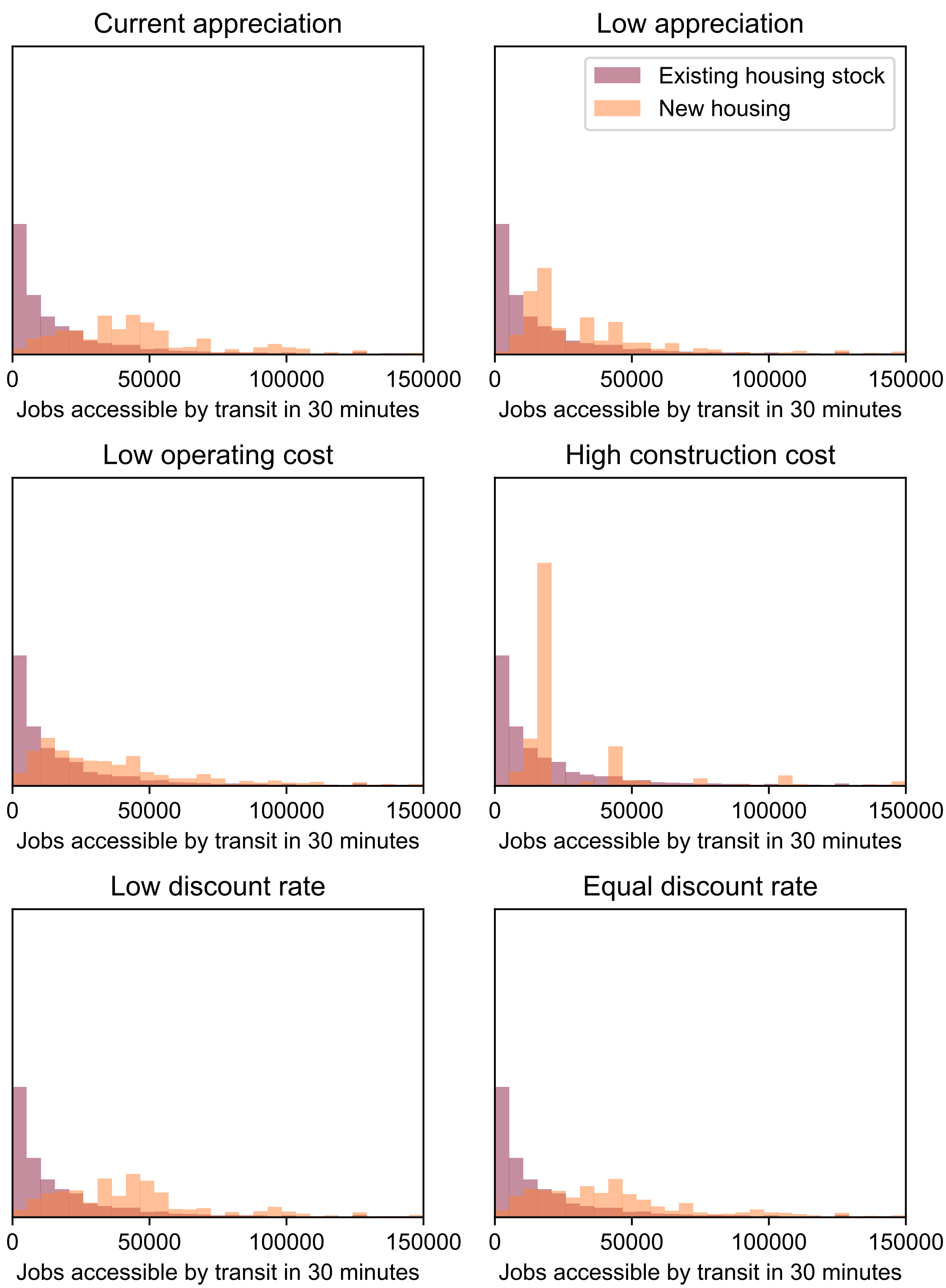


Figure 17. Access Levels Around New Development Restricted to the High Quality Transit Area, for Various Development Scenarios

The job access figures used are from the Access Across America 2017 transit access dataset (Owen and Murphy 2018). The block-level results are aggregated to tract level using population weighted weights from the Missouri Census Data Center's Geocorr platform (Missouri Census Data Center 2014).

3.5 Conclusions

As part of the RHNA process, the California Department of Housing and Community Development estimated that there is a shortfall of 1.3 million housing units in the SCAG region (McCauley 2019). Even the more optimistic low operating cost scenario does not result in enough units to meet the estimated shortfall. However, small multifamily units are not the only development occurring in Southern California; in 2019, 91% of permitted multifamily units in the SCAG region were in projects with five or more units. These larger developments are beyond the scope of this dissertation, but their economies of scale likely make them feasible in additional locations. The results of single-family upzoning simulated in this dissertation could be a significant contributor to a broader solution.

Government spending could also help meet the unit goal. In the aftermath of World War II, facing a similar housing crisis, the US federal government intervened to subsidize the housing market through readily-available loans and federal mortgage insurance that spurred significant suburban development (Hanchett 2000). While these policies were extremely racially unjust (e.g. Rothstein 2017), they demonstrate the ability of government spending to tilt the tables towards more development. The RHNA process in California demonstrates that enabling additional housing construction is a current policy goal.

Furthermore, estimates of housing construction are *highly* variable depending on the assumptions made by developers when evaluating profitability. This uncertainty makes understanding the effects of supply-side housing policies very difficult. The cyclical nature of housing construction suggests that this volatility in whether projects are profitable is real rather than an effect of the modeling herein. This indicates that researchers should always evaluate multiple scenarios with different assumptions when forecasting the effects of these policies.

This research also demonstrates that when evaluating large-scale estimates of market feasibility of redevelopment, it is important to account for the probability that a property will actually be redeveloped, even if it is profitable to do so. Generally, properties are not redeveloped until they are sold, and even then they often do not transact to a developer, even if redevelopment would be profitable. Many other evaluations of the profitability of redevelopment do not account for this effect (Monkkonen, Carlton, and Macfarlane 2020, e.g. Johnson et al. 2018). These other models give an indication about whether development is possible under different policy scenarios, and forecast where development is likely to occur geographically. However, without estimating the probability of redevelopment given profitability, any forecasts about future land use are likely to be overstated.

Chapter 4

SIMULATING RESIDENTIAL LOCATION CHOICE AND VEHICLE OWNERSHIP DECISIONS RESULTING FROM CHANGES IN THE BUILT ENVIRONMENT

4.1 Introduction

In Chapters 2 and 3, I developed and applied a model of what multifamily construction would occur, given exogenous changes in zoning codes. Before I can apply a travel demand model, I need to simulate where people will choose to live given changes in housing supply, as residential location is an important input to travel demand models. Since the vehicle ownership decision is closely related to the residential location choice decision, I model these two phenomena jointly.

This research uses an equilibrium sorting model framework (Bayer, McMillan, and Rueben 2004; Tra 2007, 2010, 2013; Klaiber and Kuminoff 2014; Kuminoff, Smith, and Timmins 2013) to understand how changes in the distribution of housing resulting from a shock to the regulatory system will likely affect the distribution of the population and car ownership choices. The equilibrium sorting model framework is unique in that it has a market clearing or equilibration step in which prices are adjusted to bring the market into equilibrium. This allows modeling how households re-sort across the region in response to a shock of some sort. In a classical discrete choice model, it is not possible to understand how the population will sort after a shock to the alternatives and their amenities, because the prices of the various options are taken as exogenous. Thus, classic discrete choice models can only measure *marginal* changes in well-being or sorting, but cannot account for the complex market dynamics that occur with a non-marginal change such as a significant increase in apartment construction. In most previous research, this shock is a change to neighborhood amenities or sociodemographics (*e.g.*, Tra 2007; Klaiber and Phaneuf 2010; Bayer, McMillan, and Rueben 2004; van Duijn and Rouwendal 2013). In this research, the shock is an exogenous change to the housing supply, as estimated by the construction and development models in the previous two chapters.

4.2 Literature Review

4.2.1 Residential Location Choice

Mathematical models of residential location choice are prevalent in the literature. Most often, these models use a discrete choice formulation, with examples dating at least to McFadden (1978). They model the choice between housing alternatives—which may be individual homes or aggregate zones—by modeling the "utility" of each alternative and finding the alternative that (probabilistically) has the maximum utility. These models may use a multinomial logit, or a more complex nested or mixed logit model.

Some authors extend the discrete choice framework to include an equilibration step. These are known as sorting models because this equilibration or market-clearing step simulates changes in residential location choices that result from changes in the attributes of housing. Whereas traditional discrete choice models that do not have an equilibration step can only evaluate the effects of marginal changes to the independent variables, sorting models allow the evaluation of non-marginal effects.

Sorting models were described by Bayer, McMillan, and Rueben (2004). Since then, they have been applied in a number of contexts, primarily for the valuation of unpriced amenities such as open space (Klaiber and Phaneuf 2010), air quality (Tra 2007, 2010, 2013), and cultural amenities (van Duijn and Rouwendal 2013). In this chapter, I apply such a model to examine the effects of a change in the housing supply.

Several authors in the transport literature have used equilibrated discrete choice models to forecast residential location choice. The UrbanSim framework adjusts prices of housing in an iterative process to bring the market to equilibrium (Waddell 2010; Waddell et al. 2018). The CT-RAMP and related ActivitySim activity based travel demand model (and possibly others) use a technique known as "shadow pricing" in order to doubly constrain destination choice models; effectively, a "price" representing unobserved disutility is added to certain choices (for instance, schools) to ensure capacity constraints are not exceeded (Atlanta Regional Commission 2019; Association of Metropolitan Planning Organizations 2019). de Palma, de Lapparent, and Picard (2015) also equilibrates a complex model of residential and workplace choice.

The key difference between this prior work and my current work based on the equilibrium sorting models used in economics is how they treat unobserved heterogeneity. The general equilibrium sorting model I use is able to differentiate unobserved between-housing-type heterogeneity from observed variables that do not vary within a housing type (i.e., grouping of houses by observable characteristics such as location or size). For instance, suppose the housing types used in the travel demand model are single-family or multifamily, as well as a PUMA. Within each housing type, there will be no variation in whether properties are single-family or multifamily. In a standard discrete choice model, it would be possible to estimate an alternative-specific constaint for a specific housing type, or to estimate the effect of single family homes on utility, but not both, because one implies the other. Through the two-stage estimation process described below, I am able to recover both the effects of observable variables that do not vary across housing types, as well as unobserved heterogeneity in the utility of different housing types. This estimate of unobserved heterogeneity is preserved in simulation, assuming that unobserved heterogeneity between housing types does not change as a result of the policy scenario.

4.2.2 Joint Models of Residential Location and Car Ownership

While standalone models of car ownership are somewhat rare, joint models of car ownership and residential location choice are more common. Car ownership is a choice that is perforce closely associated with residential location choice. Residents of dense, transit-served neighborhoods may be able to own fewer or even no cars, while residents of far-flung suburbs likely will require more cars to serve their daily transportation needs. Furthermore, households likely consider the need for car ownership when making their residential location choice, with some preferring to choose a neighborhood where they can own no or few cars, and others preferring automobility and a suburban neighborhood where driving is less constrained and car ownership levels are higher.

Salon (2009) considered the joint choice of residential location (by Census tract), car ownership, and commute mode for residents of New York City. She found that car ownership and use would be effectively reduced by increasing travel times by car relative to those by transit. Because of the lack of an equilibration step in her model, she was unable to consider the non-marginal effects of a housing shock, however. Guerra (2015) studied the joint choice of car ownership and residential location in Mexico City with a mixed logit model, and found that they were surprisingly unrelated, or perhaps even negatively related, due to the tendency of wealthy reseidents of the city center to own vehicles. Bhat et al. (2013) jointly models residential location choice and car ownership in the San Francisco Bay Area, with an eye towards understanding the travel behavior of immigrants. They find that over time immigrants assimilate to a car-oriented lifestyle, but more importantly for this paper, Bhat et al. find that the effect of residential location on car ownership is rather strong, with households choosing lower density neighborhoods owning more cars than their higher-density counterparts.

4.3 Data

Microdata on households is retrieved from the Integrated Public Use Microdata Sample (IPUMS) for the years 2012–2017 (Ruggles et al. 2019), for the SCAG region. The IPUMS contains information about household sociodemographics as well as about the homes these households occupy. These data are used to describe the population making residential location choices as well as the choices available to those residents.

The IPUMS contains individual responses to the American Community Survey. To protect respondent privacy, geographic data is only provided at the coarse Public Use Microdata Area (PUMA) level, geographies that contain at least 100,000 residents. In the densest parts of the SCAG region, these are relatively small, but in the less dense parts they can be quite large. This represents a limitation of this work; locational preferences can only be modelled at this geographic aggregation, and neighborhood-level attributes must be summarized to this level. There are 124 PUMAs in the SCAG region (Figure 18).

The IPUMS data are augmented with neighborhood information. School quality is calculated based on the median proportion of fifth-grade students considered proficient in math for elementary schools in the PUMA, based on 2012 STAR test data from the California Department of Education (California Department of Education 2012).

Intersection density and the number of jobs accessible within a 45 minute driving commute are obtained from the EPA Smart Location Database (US EPA 2013). Jobs accessible by

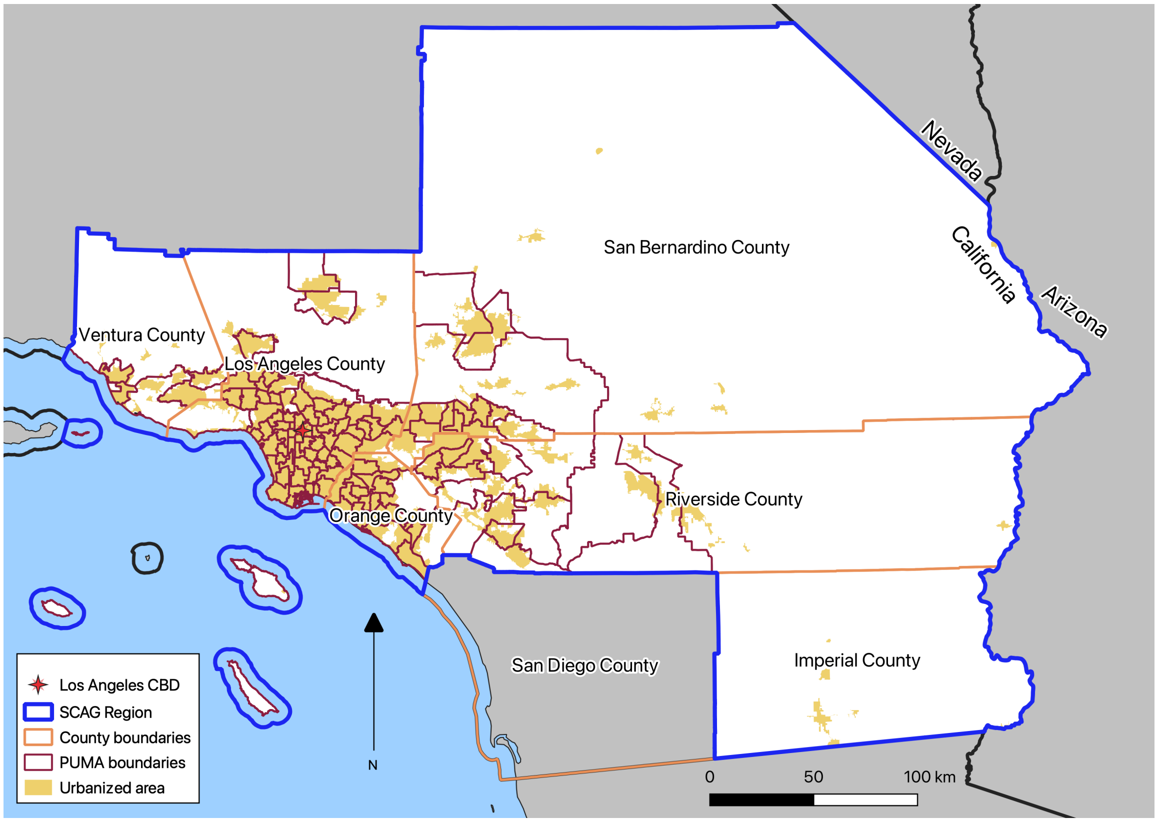


Figure 18. PUMAs in Southern California

public transit within 30 minutes are obtained from the University of Minnesota Accessibility Observatory (Owen and Murphy 2018). These datasets are provided at much finer levels of geography than the PUMA (block groups in the case of the Smart Location Database, and blocks in the case of the Accessibility Observatory data). These are summarized to the PUMA level by averaging all constituent geographies, weighted by population in the case of SLD data, and housing units in the case of Accessibility Observatory data.

Retail job density comes from the US Census Bureau LODES dataset (US Census Bureau 2017). The number of retail jobs within 1.5km of every Census tract is computed and density is calculated. That density is then averaged to the PUMA level, weighted by housing units. Crosswalks from tracts and blocks to PUMAs weighted by housing units were calculated using the Geocorr 2014 tool (Missouri Census Data Center 2014).

Table 12. Factor Analysis of PUMA-Level Density Variables, Varimax Rotation

	Regional access	PUMA access
Retail job density near PUMA	0.57	0.75
Access to jobs via transit within 30 minutes		0.70
Access to jobs via auto within 45 minutes	0.76	0.48
Intersection density	0.72	0.32

A challenge with including density and accessibility variables in regression models is that they are very often collinear; for instance, retail job density has a correlation with job access by auto of 0.79. To reduce this collinearity, I conducted a exploratory factor analysis of these four density variables, and replaced them with two factors, which I term "regional access" and "PUMA access." The loadings for this factor analysis are shown in Table 12. A varimax rotation was used, which produces factors that are orthogonal (i.e. uncorrelated). The factor analysis was estimated using `factor_analyzer` in Python (Biggs 2019).

There are two largely distinct submarkets of the housing market: the rental market and the homeownership market. Most households specifically look in one of these markets or the other, and most landowners list their properties either for rent or for sale. The model presented herein models both of these submarkets simultaneously, by defining a separate utility function for rented and owned homes, and jointly modeling the choice of each household between the rental and ownership markets. This is done by parameterizing the utility functions with interaction terms between household sociodemographics and tenure choice. All components of the utility function for a particular home, except budget, are also allowed to vary between rental and ownership tenures. Housing units, however, are exogenously assigned to either the rental or ownership market, and this is not adjusted within the model to reflect consumer demand. This is roughly reflective of the real world, where most housing units are in one market or the other, so this is not expected to materially affect the results.

Since the formulation of the equilibrium sorting model I am using requires that households cannot choose any residential choice that costs more than their income, I exclude households that report making less than $15,000 per year, or that report currently paying more in rent or in (assumed) owner costs annually than their income. Many of these households may be students who have outside income, or retirees who are living on savings, and thus their reported income is not representative of the housing units they can actually afford. After this filtering, I fit the model using data for 262,924 households, which are weighted to expand out to the full population of the SCAG region.

4.4 Methods

4.4.1 Modeling Approach

The primary modeling approach used is an equilibrium sorting model. The formulation I use is an extension of the one described by Tra (2007; 2010; 2013), as implemented in Python in my open-source package `eqsormo`.[12] Tra's model represented the choice of a housing unit; my model extends his framework to jointly estimate the choice of housing unit and car ownership. Several high-performance computational approaches are used in `eqsormo` to allow the models to be estimated in a reasonable amount of time with reasonable computation resources; these are described in detail in Appendix B.

This sorting model uses a random utility framework to determine where households will choose to locate and how many cars they will choose to own, assuming that each household

[12] https://github.com/mattwigway/eqsormo; I used version 0.8.5 for this work. `eqsormo` in turn depends on SciPy (SciPy Developers 2019), `statsmodels` (Seabold and Perktold 2010), and `dill` (McKerns et al. 2012).

Table 13. Nomenclature

Symbol	Definition
i	Index of households
h	Index of housing types (location, age category, and single-family/multifamily)
c	Index of car ownership levels
x_i	Sociodemographics of household i
a_h	Observable attributes of housing type h (age category, single-family/multi-family, neighborhood features, etc.)
U_{ihc}	Utility received by household i by choosing housing type h and car ownership level c
V_{ihc}	Systematic portion of utility received by household i by choosing housing type h and car ownership level c
P_{ihc}	Probability household i chooses housing choice h and car ownership level c
$\bar{V}_{ih}$	Average utility of the housing choices within housing type h to household i
p_h	Price of housing type h
z_i	Income of household i
β_a	Coefficients for utility housing attributes
β_{ax}	Coefficients for utility of interactions between housing attributes and demographics
β_{cx}	Coefficient for sociodemographic characteristic x for level of car ownership c, generally repeated for each level of car ownership
β_b	Budget coefficient
γ_c	Alternative specific constant for level of car ownership c
ϵ_{ihc}	random error for utility of household i choosing housing type h and car ownership level c
Θ_h	Utility of housing type h that does not vary over households
ξ_h	Type-specific unobserved utility of housing type h that does not vary over households
μ'	Scale parameter for nested logit derivation of aggregate logit model, assumed equal to 1
S_h	Supply of housing type h
B_{ihc}	Measure of variability of utilities to household i of housing choices contained in housing type h given car ownership level c
$d_h(p)$	Demand for housing type h given price vector p
W_i	Survey weight for household i
ν	Constant estimated in second stage which estimates location of Θ_h. Since utility is arbitrarily located, this parameter has no interpretation.
α	Scalar used to scale price changes during sorting

i chooses housing unit h and level of car ownership c to maximize its utility

$$U_{ihc} = \beta_{rent}x_i + \beta_{a,rent}a_h + \beta_{ax,rent}a_h x_i + \beta_{a,own}a_h + \beta_{ax,own}a_h x_i + \beta_b \ln (z_i - p_h) \\ + \xi_h + \gamma_c + \beta_{cx}x_i + \beta_{ca}a_h + \ln S_h + \epsilon_{ihc} \quad (4.1)$$

This is a linear-in-parameters utility specification common in discrete choice modeling. a_h is observable housing attributes of housing type h, p_h is the price of housing type h, z_i is the income of household i, and x_i are sociodemographic features of household i. γ_c is an alternative-specific constant for car ownership. γ_c and all of the β parameters are estimated. ξ_h is the unobserved utility of housing unit h—that is, a fixed effect. S_h is the supply of housing units of type h. Table 13 lists all nomenclature used in this chapter. While (4.1) shows only a single housing attribute and sociodemographic characteristic for simplicity, it is customary to use multiple housing attributes and sociodemographic characteristics in the model. By interacting housing attributes with sociodemographics, different households can exhibit heterogenous preferences—for example, large families likely prefer homes with additional bedrooms, and higher-income families are less price sensitive.

By assuming that ϵ_{ihc} is iid extreme value distributed, the model becomes a multinomial logit model (Ben-Akiva and Lerman 1985). What distinguishes the equilibrium sorting model from the multinomial logit model is the equilibration step. While the vector of prices p_h is initially observed, a market-clearing step is undertaken wherein prices are adjusted so that demand equals supply when the model is used for simulations. This equilibration step is what allows me to evaluate the equilibrium after a change to the supply of different types of housing.

Estimating the model is a two-stage process, because the variables a_h that do not vary over households within a particular housing type h are perfectly collinear with the fixed effects ξ_h. In the first stage, the parameters for interaction terms and car ownership (β_{ax}, β_{cx}, β_{ca}, and γ_c) in (4.1) are recovered, and in the second stage the base effects of housing attributes β_a

as well as the type-specific unobserved utilities ξ_h are recovered. To facilitate this estimation process, I rewrite (4.1) as (Klaiber and Phaneuf 2010).

$$U_{ihc} = \beta_{ax} a_h x_i + \beta_b \ln(z_i - p_h) + \ln S_h + \gamma_c + \beta_{cx} x_i + \beta_{ca} a_h + \Theta_h + \epsilon_{ihc} \tag{4.2}$$

with

$$\Theta_h = \nu + \beta_a a_h + \xi_h \tag{4.3}$$

The first stage (Equation 4.2) recovers the coefficients from the interaction terms β_{ax}, β_{cx}, β_{ca}, and γ_c, as well as as an intermediate value Θ_h, which represents the portion of utility attributable to housing choice h that does not vary over households. The second stage decomposes Θ_h into its constituent parts, and recovers β_a and ξ_h (Tra 2007).

The first stage can be estimated as a multinomial logit model. Θ_h can be treated as an alternative specific constant. One complication is that there are many potential housing types, and estimating a full set of alternative-specific constants may be computationally challenging. However, since a full set of alternative specific constants means that market shares will be predicted perfectly, any set of γ_c and β values implies a set of alternative specific constants that reproduce market shares. Bayer, McMillan, and Rueben (2004) proposes a simple and fast contraction mapping to find the optimal Θ_h values from arbitrary starting values Θ_h^0 is

$$\Theta_h^{t+1} = \Theta_h^t - \ln \frac{\sum_i \sum_c P_{ihc}}{S_h} \tag{4.4}$$

By applying (4.4) repeatedly, the values of Θ_h can be brought arbitrarily close to values that reproduce the observed market shares.

The logic behind this contraction mapping is that the ASC (Θ_h) is reduced at each step for alternatives that are overconsumed (i.e. $\frac{\sum_i \sum_c P_{ihc}}{S_h} > 1$; summing P_{ihc} over i and c produces the predicted demand for housing type h), and increased for alternatives that are underconsumed. When demand is equal to supply, the fraction is equal to 1 and the natural logarithm is equal to zero, and thus the process converges.

(4.4) finds only a single set of ASCs, however in this model formulation there are two sets of ASCs—one (Θ_h) for housing choices, and one (γ_c) for car ownership levels. Similar to how a complete set of ASCs for all choice alternatives perfectly reproduces market shares, complete ASCs for the margins of a joint choice model perfectly reproduce marginal market shares. It is straightforward to extend the contraction mapping to multiple sets of ASCs. I simply alternate between (4.4) and a similar formula for car ownership

$$\gamma_c^{t+1} = \gamma_c^t - \ln \frac{\sum_i \sum_h P_{ihc}}{S_c} \tag{4.5}$$

where S_c represents the "supply" or observed market share of car ownership level c. Each iteration moves both sets of ASCs closer to reproducing market shares until convergence is achieved.

Another complication with estimating the first stage is that there are frequently a large number of alternatives. However, since the first stage is essentially a multinomial logit model, the parameters can be recovered by randomly sampling from the alternatives available to each household (Bayer, McMillan, and Rueben 2004; Ben-Akiva and Lerman 1985, ch. 9). I use a sample size of 10 housing alternatives for each household, and retain all car ownership choice alternatives for each sampled housing alternative, resulting in 40 joint alternatives per household. However, I recalculate Θ_h using the contraction mapping procedure above using the full set of alternatives once the other parameters have been identified, and before fitting the second stage.

In most applications of equilibrium sorting models, a constant marginal utility of price is estimated in the second stage—that is, there is a term $\beta_p p_h$ in (4.1) and (4.3) (e.g. van Duijn and Rouwendal 2013; Klaiber and Phaneuf 2010; Bayer, McMillan, and Rueben 2004). This requires estimating β_p via instrumental variables, since price is almost certainly correlated with unobserved aspects of housing (Bayer, McMillan, and Rueben 2004).

In the Tra (2007; 2010; 2013) formulation, a nonlinear budget (income minus housing price) term is included in the first stage. This allows households with different incomes to have different marginal utilities of price, which is a more accurate representation of the real world. Additionally, since the first stage contains a full set of alternative-specific constants (i.e. the fixed effects Θ_h) for each housing type, all unobserved features of housing are controlled for. Assuming that preferences for the unobserved attributes of housing do not vary over households, this means that instrumental variables estimation is not indicated (Tra 2007, 85).

In order to guarantee consistency of the second stage, it is necessary for the number of choices to be large relative to the size of the choice set (Klaiber and Phaneuf 2010, 62n3). Thus, like other authors employing this framework, I have aggregated homes into housing types—in my case, by Public Use Microdata Area, single-family/multi-family housing type, rental/ownership tenure, and approximate age less than or greater than 15 years.

Age is approximate because it is derived from the difference of the year of the survey in the five-year IPUMS sample, and the midpoint of the categorized year built variable. 15 years of age strikes a balance between including many properties that do not carry the "new home" price premium, while still including reasonable numbers of properties in each category. Due to the aggregation of construction, this means that all homes built in 2000 or later will be included in the "new" category.

The aggregation of alternatives requires the addition of the natural logarithm of housing supply to the utility function, as described in Ben-Akiva and Lerman (1985, ch. 9). When alternatives are aggregated and represented by an average utility,[13] Ben-Akiva and Lerman propose adding two terms to the utility function. With $\bar{V}_{ihc}$ representing the systematic utility to household i of the average home of housing type h and car ownership level c, they add two terms:

$$V_{ihc} = \bar{V}_{ihc} + \mu' \ln S_h + \mu' \ln B_{ihc} \tag{4.6}$$

where S_h is the supply of housing type h and B_{ihc} is a measure of the variability of the utility of the different housing choices contained within housing type h, given car ownership level c since car ownership levels are not aggregated. B_{ihc} is defined as (Ben-Akiva and Lerman 1985, 257)

$$B_{ihc} = \frac{1}{S_h} \sum_{i}^{I} e^{V_{ihc} - \bar{V}_{ihc}} \tag{4.7}$$

For the purposes of this analysis, I assume there is no variation in housing options within a housing type, which makes B_{ihc} equal to 1 and $\ln B_{ihc}$ equal to zero, so this term drops out of the utility function. Waddell (2000) used a similar specification including the log of supply in his aggregate utility function.

Estimating μ' is not possible because the log of supply is perfectly collinear with the alternative-specific constants Θ_h. However, under certain conditions μ' can be assumed to be equal to 1. To understand these conditions, it is instructive to note that the utility function for aggregate alternatives given in (4.6) can be derived as the marginal choice probability

[13]For technical reasons, the average utility computed in this project is not a true average, because the utility function (4.2) is computed based on average price, rather than being computed separately for each home within the aggregate choice. Since price enters the utility function non-linearly, computing utility using average price does not give exactly the same answer as computing the average of all utilities. However, price variation *within* an aggregate alternative is assumed to be small, so budgets only vary over a relatively small range within each alternative, and within a small range the natural logarithm approximates a linear function, meaning that the average utility used herein approximates a true average of the utility of each home within the aggregate alternative.

of a nested logit model, with μ' equal to the ratio of the scales of the Gumbel-distributed error terms (Ben-Akiva and Lerman 1985, 259). The ratio of the error terms can be assumed to equal 1 when there is no unobserved variation at the upper level of the housing type but rather all unobserved variation is at the level of the individual housing unit within the aggregate housing types (289). Since the housing-unit-level ASCs can be seen as fixed effects that control for the utility of any unobserved housing attributes, this is a reasonable assumption.[14] Thus, I assume that $\mu' = 1$, which allows me to estimate the first stage.

No such adjustment is needed for the ASCs for car ownership (γ_c), for two reasons. Firstly, these are not aggregate alternatives. Secondly, since the ASCs for car ownership are not decomposed by a second stage model, any needed adjustment for the relative availability of larger numbers of cars in a household will simply be subsumed by the ASC.

Other authors applying equilibrium sorting models have generally not included the log of supply in their first stage (e.g. Klaiber and Phaneuf 2010; Tra 2007, 2010, 2013; van Duijn and Rouwendal 2013). However, the problem created by not including this term was less severe in these cases, as these authors did not simulate the effects of changes in housing supply. Thus, the log of supply term is subsumed into the unobserved portion of utility ξ_h. The main concern for these models is then that the size of the alternatives would bias the second-stage estimation that decomposes the ASCs to constituent parts; the ASCs will partially represent the sizes of the alternatives. Since I am simulating a change in the supply of housing, properly accounting for the size of the aggregate alternatives is critical.

The second stage is straightforward to estimate; (4.3) is simply estimated using OLS. One

[14] If tastes for the unobserved attributes vary over different households, this assumption may not be perfectly correct. For instance, neighborhood income composition is not a covariate in my model; if neighborhood income composition affects the taste of high income and low income households differently, there may be error at the housing type level, meaning that μ' should be less than 1.

downside to this sequential estimation is that the standard errors from OLS will not account for the error in Θ_h from the first stage.

The IPUMS data includes household weights. While weights are necessary to calculate descriptive statistics, it is less clear whether they are needed when estimating a regression (Solon, Haider, and Wooldridge 2015). For all of the regressions in this paper, I do not use the weights in initial estimation. However, once the first-stage parameters have been identified, I recalulate the alternative-specific constants Θ_h and γ_c using the contraction mapping procedure described in Bayer, McMillan, and Rueben (2004) to find the values that will clear the market given the weights, but holding the other estimated parameters constant. To do this, the log of supply term is replaced with the log of the supply given the weights, to properly account for the availability of alternatives in the weighted data. These alternative-specific constants are the ones that are decomposed by the second stage. I also use the weights when solving for a new equilibrium, and for all descriptive statistics.

4.4.2 Solving for a New Equilibrium

The key feature that differentiates sorting models from other discrete choice models of residential location is their equilibration step. While other discrete choice models can simulate the effects of changes in the covariates on choice outcomes, they do not account for the fact that changes in amenities or the supply of various choice alternatives will have effects on prices. As demand exceeds supply in certain locations in simulation, standard discrete choice models do not simulate the effect that will have on prices to bring the market into equilibrium. Modeling changing prices is critical to sort households correctly when there is a supply change.

To solve for a new equilibrium, I first adjust the utility (4.6) with the new supply of each housing type implied by the land use scenario. I then find the price vector p_h such that the market shares of each housing type are equal to the demand for that housing type. An iterative process described in Tra (2007, eq. 7.7), which Tra attributes to Anas (1982), is used. Starting from the initial price vector p_0, additional price vectors are calculated as:

$$P_{t+1} = P_t - \alpha \left[\frac{\delta D_{P_t}}{\delta P_t} \right]^{-1} (D_{P_t} - S) \tag{4.8}$$

where P_t is the vector of prices for all housing types at time t, D_{P_t} is the vector of demand for all housing types given prices P_t, and S is the vector of total supply of each housing type. At each iteration, the derivative of demand for each housing type with regard to price of all housing types is computed (i.e., the Jacobian matrix). This is the term $\frac{\delta D_{P_t}}{\delta P_t}$ in (4.8). The inverse of this matrix is multiplied by the difference between demand and supply to compute how far price would have to move to bring the market to equilibrium, if demand and price had a linear relationship. This process is repeated until the demand vector D_{P_t} is arbitrarily close to the supply vector S.

Like Tra (2007), I hold a single price constant, so that the system of prices is identified and so that any starting value P_0 converges to the same equilibrium price vector. I implement this by setting $p_{h,t+1} = p_h$ at each iteration, for a predetermined single value of h, rather than using (4.8) for this single housing type. Bringing demand into equilibrium with supply for all other housing types perforce equilibrates supply for this housing type with a fixed price, since total demand for all housing types equals total supply in this model form by construction.

From a theoretical standpoint, this constant price reflects demand for housing from outside the region, consistent with an open-city model of urban economics. In the open-city model, prices in the center of an idealized one-dimensional city are assumed to be determined by external demand, and all other prices in the city are determined based on that price and their relative location. If this price were raised (lowered), people would flee (flock to) the city (O'Flaherty 2005, 122). Since the sorting model does not have any other modeling of external demand, fixing this price is necessary. To be consistent with the theory, I fix the price in downtown Los Angeles, despite the polycentricity of the region—specifically, I fix the price in PUMA 03744, which covers the Los Angeles central business district. However, I can only fix the price for one housing type within this PUMA. Given that I am simulating construction scenarios, I fix the price of older housing, rather than newer housing. I fix the price for pre-2000 multifamily rental housing because it is the most prevalent type of housing in this PUMA.

Fixing one price is also needed for a technical reason. (4.8) adjusts the prices in each iteration by computing the derivative of demand for housing type h with respect to price at the current price of that housing type, and then using a linear approximation to adjust prices closer to equilibrium. However, the linear approximation of the demand curve implied by using the Jacobian matrix in (4.8) is likely to overshoot the market-clearing prices that are closest to the existing prices, and find an equilibrium far from present prices. In practice, without the fixed price, prices may become significantly negative—which, given that I do not model household migration, can produce an equilibrium outcome, but clearly not a realistic one.

Estimating the full Jacobian matrix $\frac{\delta D_{P_t}}{\delta P_t}$ is computationally costly; it takes approximately 3 hours to evaluate for each iteration on the machine I use for simulation (16 cores and 128 GB of RAM). Tra also presents a alternate form of (4.8) in which only the diagonal of the Jacobian matrix is computed, and off-diagonal elements are assumed to be zero (Tra 2007, eq. 7.7). Estimating only the diagonal takes only about 40 minutes, so this is a much faster process. However, ignoring the off-diagonal elements means that the system may not reach convergence. To take advantage of the speed gains of using the diagonal Jacobian, while retaining guaranteed convergence, I run 3 iterations of (4.8) with the diagonal Jacobian before switching to computing the full Jacobian. Since with the price of one alternative fixed, there is only a single equilibrium, any process that converges will converge to the same equilibrium, so starting with the diagonal of the Jacobian does not affect the final solution, only the speed with which it is achieved.

The parameter α in (4.8) is not presented in Tra (2007), but makes the convergence of the price system faster and more robust.[15] The Jacobian represents a linear approximation to the demand curve, and the portion of (4.8) to the right of α is the *search direction* of a *line search* towards a minimum. Moving exactly the amount suggested by the linear approximation will not necessarily produce the maximum absolute reduction in excess demand. The line search process is the process of choosing the scalar α which one can multiply the search direction by in order to produce the largest reduction in the objective function. Tra implicitly assumed α was 1, which is not guaranteed to converge for all problems, but is not incorrect if it does converge as it did for Tra.

In most line search applications, α is loosely approximated, to avoid spending computational time in the line search that could be spent in additional iterations of the algorithm (Nocedal and Wright 2006, 37). However, since estimating the Jacobian takes several hours, while computing the excess demand for a given α takes only a few minutes, in this particular instance it is worthwhile to spend additional computation time to find an optimal α. I use Brent's method as implemented in SciPy (SciPy Developers 2019) to find an optimal α, using the sum of squared excess demand as the scalar objective function. I limit Brent's method five iterations and a tolerance of 0.001 to control calculation time while still producing an alpha sufficiently close to optimal. The values of α as well as the sum of squared excess demand

[15] Many thanks to Sam Zhang for his assistance with numerical optimization.

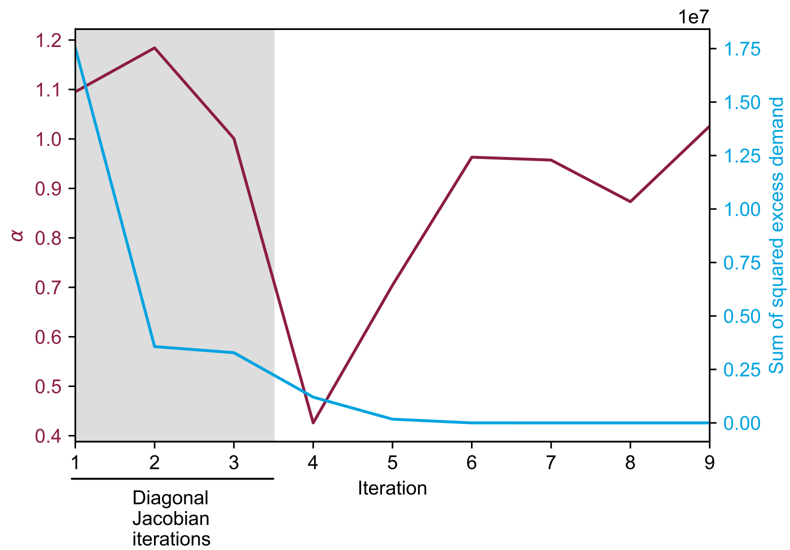


Figure 19. Evolution of α and Sum of Squared Excess Demand Over 9 Iterations of Price Clearing

over the 9 iterations required to clear the market for a representative scenario are shown in Figure 19.

4.4.3 Scenarios

I simulate six scenarios of housing construction in the SCAG region. Five come from the simulations of development described in Chapters 2 and 3. One is the "Current appreciation" scenario, which is the best guess at the outcome of a zoning change. In addition to the expected scenario, I model a high development and a low development scenario, which correspond to the low operating cost and low appreciation scenarios described in Chapters 2 and 3. I also test versions of these current appreciation and low operating cost scenarios where redevelopment is only allowed in the High-Quality Transit Area as defined by SCAG, since some upzoning proposals have focused on areas near transit in an effort to promote sustainable transportation and transportation affordability (Wiener et al. 2019).

The seventh scenario is based on the 2020–2021 draft Regional Housing Needs Allocation (RHNA), to compare the simulated zoning scenarios to existing plans. In the RHNA process, the California Department of Housing and Community Development projects the need for new housing in each of California's regions; regional governments then divide that allocation among the local governments within their jurisdiction. Enforcement has historically been fairly weak, but this program does attempt to encourage housing production across the state (California Department of Housing and Community Development 2019; Fulton and Shigley 2005, 280–283). Importantly, the law only requires that governments plan for housing; it is left to the private sector to build (or not build) the planned housing. In August 2019, the need for new housing in the SCAG region was estimated to be 1.3 million units (Dillon 2019). The task then fell to SCAG and its member cities to plan for this many units, and allocate them among the different municipalities in the region; I use draft allocations published in September of 2020 (Southern California Association of Governments 2020).

Constructing scenarios from the RHNA allocations requires some additional assumptions. First, I assume that all housing planned for as part of the RHNA process will actually be built—which is unlikely to happen. Even if zoning allows housing construction, it may not be built. However, recent changes to the guidelines for the RHNA process encourage cities to more strongly consider the feasibility of housing development (Monkkonen, Carlton, and Macfarlane 2020), making this assumption less problematic.

The remaining assumptions come from the fact that the current RHNA scenarios specify the total number of housing units in each municipality in the SCAG region. They do not separate out housing units that will be rented versus those which will be sold, and the plan does not currently specify where in each city the housing units will be built, nor does it specify whether they will be single-family or multifamily units—all salient variables in my model.

To estimate the proportion of the units that will be single-family or multi-family, I use the proportion of housing units permitted in the jurisdiction from 2010–2018 that were multifamily, from the Census Building Permits Survey (US Census Bureau 2018). For four cities that did not produce any housing during this time period or were not in this survey (Vernon, City of Industry, Laguna Woods, and Calipatria), I used proportion of all existing housing stock that is multifamily from the current American Community Survey. I use this to calculate single and multi-family growth rates for each municipality, relative to the existing housing stock. I treat unincorporated areas as separate municipalities.

To move from city-level geography to the PUMAs used as the geographic unit in the analysis, I assign every Census tract in the SCAG region to a municipality and assign the growth rates for that municipality to that tract. I then take a population-weighted average of all tracts in each PUMA to compute growth rates for single and multi-family housing, using tables computed by the Geocorr 2014 tool (Missouri Census Data Center 2014). The growth rates across the SCAG region are shown in Figure 20; single-family growth occurs in exurban and rural areas, while multi-family growth is concentrated in the core of the region. I apply these growth rates to the existing rental and ownership housing stock in each PUMA, implicitly assuming that the proportion of single-family or multi-family housing that is rented within each PUMA will stay the same in the future.

Since the equilibrium sorting model I use requires that housing demand exactly equal supply, I scale the overall housing stock resulting from each scenario to be equal to the number of units in the Census microdata. Therefore, I am simulating the effects of a change in the distribution of housing, but not the total number of units. Future research could

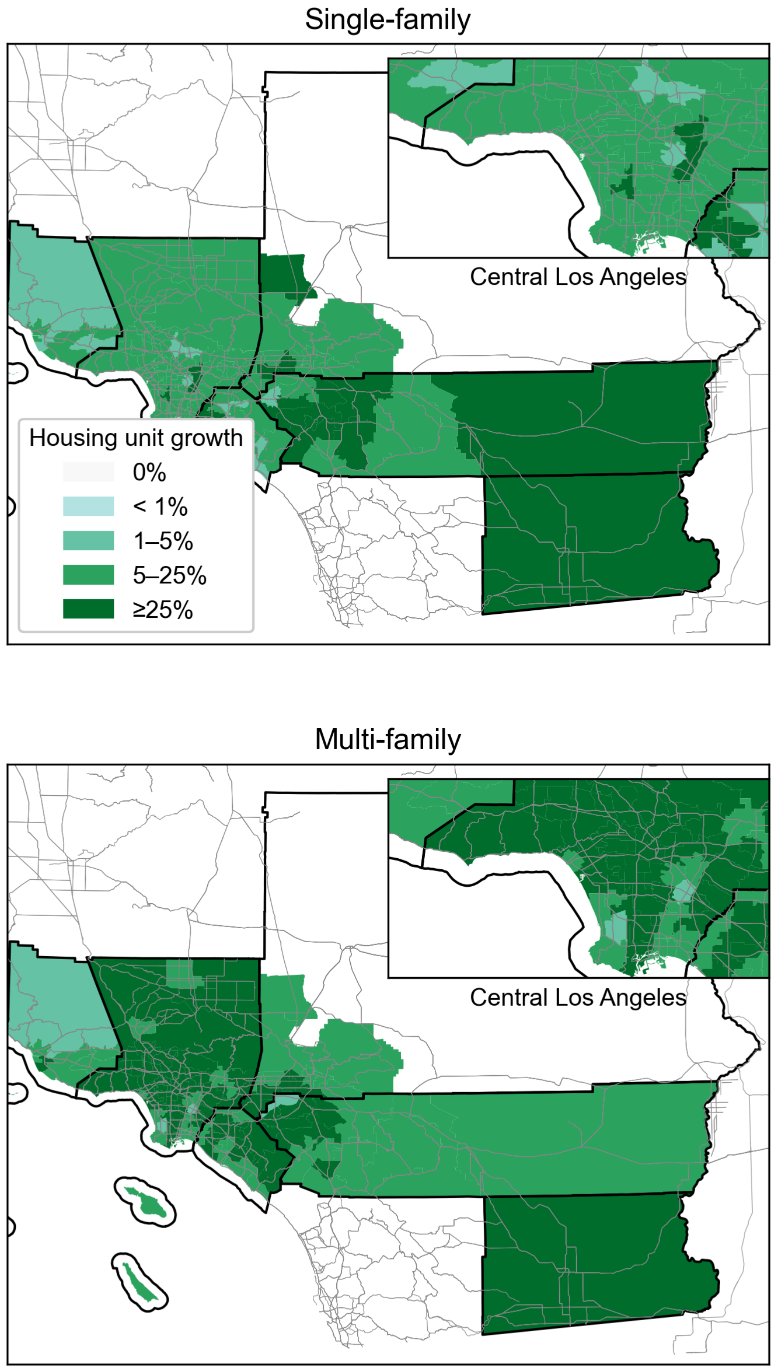


Figure 20. Growth Rates for Single- and Multi-Family Housing Across the SCAG Region, Derived from RHNA Allocations

simulate household migration and formation to more realistically represent the effects of an increase in housing supply.

4.5 Results

4.5.1 Model Estimation

The estimated coefficients for the equilibrium sorting model are shown in Appendix C. Estimating the model took 14 hours on a server rented from Amazon Web Services, with 128GB of memory and a 16-core AMD EPYC processor.

Appendix C is divided into several sections, for different conceptual components of the model. However, all of the "first-stage" components are jointly estimated through interaction terms between tenure, household characteristics, housing characteristics, and vehicle ownership levels. The first section describes the tenure choice component—that is, whether a particular household chooses to own or rent. Higher income households and households with senior sitizens, university educated members, or immigrants are more likely to own, while larger households, households with more workers, and households with children are more likely to rent, holding all else equal.

The next section describes the residential location choice model, which is jointly estimated with the tenure choice model. The first coefficient represents the budget, which, as expected, is positive—people prefer housing that leaves them more remaining budget (recall that the budget parameter is estimated as $\ln y - p$ where y is income and p is annual price). Currently, a single coefficient is estimated for both renters and owners, with annual owner costs estimated to be 4.4% of home value. In future work I intend to estimate separate budget coefficients for owners and renters—something that is conceptually straightforward, but technically rather difficult due to the existing architecture of the `eqsormo` package.

Parameters are estimated separately for renters and owners for the remainder of the residential location choice model. Among renters, college-educated individuals are more likely to choose neighborhoods with higher test scores, more accessibility both at the PUMA and regional level, and are less likely to live in a single-family home, all else equal. Households with more workers are similarly likely to live in more accessible areas, suggesting a desire for access, but less likely to live in areas with PUMA access above the 90th percentile—possibly due to the size of units in these areas. Immigrants are also likely to live in more accessible areas, but not the most accessible areas, and less likely to live in single-family homes. Households with children are less likely to live in higher accessibility areas. Seniors are more likely to live in single-family homes and somewhat less likely to live in accessible areas. Larger households, unsurprisingly, are more likely to live in single-family homes. Higher-income households are somewhat more likely to live in areas with higher test scores, and much more if there are children in the household. They are also more likely to live in single-family homes and dense PUMAs. Surprisingly, households with children are less likely to live in areas with high test scores, except for households with both university-educated members making over $100,000 per year. This counterintuitive finding likely has to do with the relatively higher housing costs in these areas being less affordable to households with children than other households in the same income bracket, due to the costs of caring for a child and the higher housing requirements of a family. An alternate explanation could be related to omitted variables; this is discussed below.

Results for owners are similar to those for renters. College-educated individuals are more likely to live in areas with higher test scores and higher access, although less likely to live in the areas with the highest regional access. They are again less likely to live in a single-family home. Again, households with workers tend to live in more accessible areas, but not the areas with the highest PUMA access. Immigrants are less likely to live in single-family homes, and more likely to live in accessible areas, though not the most accessible areas. Households with children are less likely to live in areas with high regional access, but there is no significant relationship with PUMA access. Senior homeowners are more likely to live in single-family homes in areas with higher test scores—possibly because they considered test scores when they made their residential location choice in an earlier life stage, or they tend to live in wealthier neighborhoods. Results around accessibility are mixed for senior homeowners. Again, larger households are likely to live in single-family homes. High income households are more likely to live in high-test-score areas. The highest income households are more likely to reside in single family homes than the lowest-income households, but middle income households are less likely to. Results around accessibility are mixed. Like renters, households with children are less likely to like in areas with high test scores, except for the highest-income college-educated households.

The vehicle ownership choice model is presented next. Households with children are likely to own vehicles, although less likely to own three or more—possibly due to a lack of drivers. Higher income households, households with more workers, and households with a college-educated member are likely to own more vehicles. All the access variables are negatively correlated with vehicle ownership, as expected, and all but one are significant. Since the vehicle ownership model is estimated jointly with the residential location choice model, the tradeoffs between location and vehicle ownership decisions are explicitly modeled by these parameters.

The final section details the second-stage model, which decomposes the alternative-specific constants Θ_h from the first stage into effects that do not vary over households. This is the only portion of the sorting model that is not jointly estimated with the remainder of the model. As with the residential location choice model, parameters are estimated separately for owners and renters. These coefficients represent baselines that are then modified by the interaction terms in the first stage. Single-family homes are relatively less preferred in this stage, while for both renters and owners, correlations with access are mixed. Test scores have a negative coefficient. The residuals from the second-stage model ξ_h represent the type shocks that are assumed to represent unobserved utility of each housing type that persists into the sorting process.

Neither race, nor median neighborhood income, nor any other aspects of the potential neighbors in a new neighborhood are included in the sorting model, though they often would be included in non-equilibrated models of residential location choice. The reason for this is that these aspects are endogenous—they change when households re-sort across space. In other applications of sorting models, authors have included these endogenous variables and used predicted values for them during simulation (Bayer, McMillan, and Rueben 2004). However, since the model is fit using observed values for these variables, and since the predicted and the observed variables do not exactly match, significant re-sorting occurs even when no changes to housing supply are imposed, simply because replacing the observed values with predicted values places the market out of equilibrium; in testing, this re-sorting due to the mechanics of the model was often larger than the re-sorting due to the development scenario, so these variables have been excluded. The danger of course is that they may bias the coefficients of other variables that are in the model, for instance math test score; this could explain some of the counterintuitive results for the relationship between presence of children and test scores, for example.

4.5.1.1 Model Fit

The first stage of the model had a McFadden's adjusted pseudo-R^2 of 0.14, reasonable for a joint choice model of residential location and car ownership. The second stage had an R^2 of 0.94, suggesting that the components of utility that do not vary across households are well explained by this model. While second-stage R^2 statistics are often not presented in examples of equilibrium sorting models, Tra (2010) reports second-stage R^2 values less than 0.2 across several models—presumably because he did not include the log of supply in his utility function, and it was captured in ξ_h.

4.5.2 Simulation Results

The model developed above was used to simulate the effects of the four land-use scenarios described in Section 4.4.3 on car ownership and residential location choice. Importantly, in computing the statistics that follow, at no point did I assign each household to a particular location or level of car ownership; rather, I used the predicted probability of each choice outcome for each household to compute expected values for each of the values shown below. For instance, the proportion of households owning c cars in PUMA j is computed as $\sum_{h \in j} \sum_i P_{ihc} W_i / \sum_{h \in j} \sum_c \sum_i P_{ihc} W_i$.

Each scenario took 13–20 hours to simulate on Amazon Web Services machines with 128GB of RAM and 16-core AMD EPYC processors.

The first travel-related outcomes of this dissertation are presented in Table 14, which summarizes household vehicle ownership for the 6 scenarios. Vehicle ownership is quite stable across all scenarios, ranging from 2.029 to 2.033 vehicles per household. The lowest vehicle ownership occurs in the RHNA scenario, suggesting that SCAG has somewhat successfully guided growth planning into areas where car ownership is less necessary. However, the changes are so small between scenarios that is difficult to say with certainty whether they represent true differences between the scenarios, or amplified noise in the models.

The second section of Table 14 again shows vehicle ownership levels, but only for the

Table 14. Predicted Household Vehicle Ownership for Different Scenarios

	Percent of households				
	0 cars	1 car	2 cars	3+ cars	Average
Existing (model estimation sample)	4.4%	28.6%	39.9%	27.1%	2.033
Fitted values	4.4%	28.6%	39.9%	27.1%	2.033
RHNA	4.5%	28.8%	39.8%	27.0%	2.027
Current appreciation	4.4%	28.6%	39.9%	27.1%	2.033
Low appreciation	4.4%	28.6%	39.9%	27.1%	2.033
Low operating cost	4.4%	28.6%	39.9%	27.1%	2.032
Current appreciation (HQTA)	4.4%	28.6%	39.9%	27.1%	2.032
Low operating cost (HQTA)	4.5%	28.7%	39.8%	27.0%	2.029
Residents of new buildings only					
Current appreciation	2.9%	24.5%	41.8%	30.8%	2.159
Low appreciation	2.5%	22.5%	42.4%	32.7%	2.217
Low operating cost	3.6%	27.1%	41.1%	28.2%	2.080
RHNA	4.8%	29.7%	39.4%	26.1%	1.999
Current appreciation (HQTA)	5.1%	31.3%	39.7%	23.9%	1.945
Low operating cost (HQTA)	4.7%	30.5%	40.0%	24.7%	1.972

households choosing to live in the new buildings.[16] Here, there is much more variation. The lowest car ownership levels among new building residents occur in the two High Quality Transit Area scenarios—suggesting that these developments may have positive transport outcomes, or at least attract people predisposed to lower car ownership (which is the stronger effect remains an open topic of research; see Guan, Wang, and Cao (2020) for a recent review). The highest car ownership occurs in the two "business as usual" scenarios, where the only

[16] The model does not specifically forecast which households will live in the new housing, because the new housing does not exist in the baseline case and thus its utility cannot be estimated. Instead, the model forecasts which households are likely to choose housing built since 2000. Within the homes built since 2000 in each PUMA and multifamily/single-family aggregate choice, a certain percentage of the homes are provided by the scenario. The probability of choosing a single or multifamily home built since 2000 in each PUMA is multiplied by the percentage of such homes that are provided by the scenario to estimate the probability that each household chooses one of the homes provided by the scenario. These probabilities are used as weights to calculate car ownership statistics for households residing in the new housing.

policy change is to allow multifamily homes throughout the region; by not directing these homes geographically, their residents own more cars.

A key concern with redevelopment is equity. One of the outputs of the equilibrium sorting model is a set of new prices for each housing option that represent the simulated demand at market equilibrium. These prices are displayed for renters in Figure 21. Rents remain relatively constant in most scenarios. In the aggressive low operating cost scenario, rents fall in some areas that see redevelopment and rise in others. In part, this is because newer homes tend to cost more than older homes, all else equal, so even though there is some reduction in prices due to unit mix effects, all scenarios make the housing stock newer overall. This is a difficult problem to solve—simply not building is not a solution, because for older affordable homes to exist in the future, they need to be built now. Rent increases in outlying areas of Riverside, San Bernardino, and Imperial counties likely result from the mechanics of the estimation process. Since the overall supply of housing is held constant and only the spatial distribution is adjusted, these areas that see little development have their housing supply *reduced* by this scaling process. This is economically consistent with the open city theory where building new buildings in the center attracts migration, making these outlying areas have relatively less of the regional housing stock, but a critical assumption of that theory is that the region functions completely as a signle housing market, whereas in actuality few if any households consider housing both in Imperial and Los Angeles counties, for example.

Another lens to evaluate equity in the outputs of the model is the income distribution across space. US cities are often segregated by income; one possible goal for zoning reform is to reduce this segregation. In order to evaluate the changes to the spatial distribution of income brought on by these scenarios, a baseline must be established. Since the residential location choice model does not fit perfectly, the predicted spatial distribution of income is flatter

Figure 21. Changes in Monthly Average Rents Under Different Scenarios, Weighted Average of Single- and Multi-Family Rents

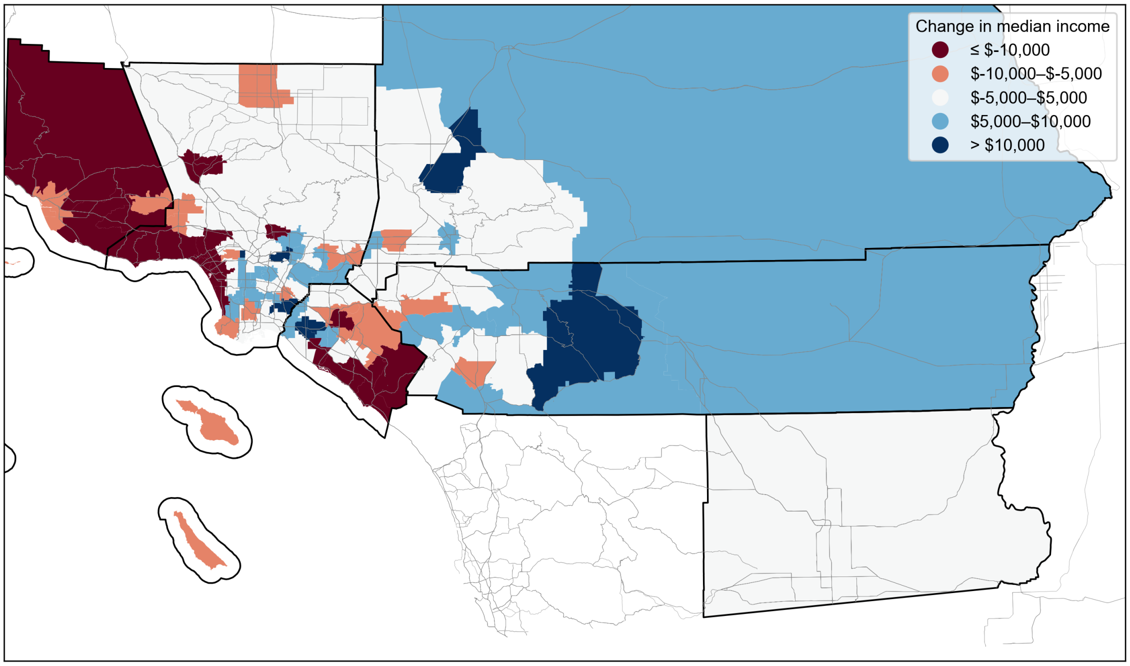


Figure 22. PUMA-Level Median Income, Fitted vs. Observed

than the observed distribution even before any changes have been made to housing supply; the standard deviation of PUMA median income is $17,508 for fitted values vs. $22,800 for observed values. These differences between fitted and observed values for PUMA median income are shown in Figure 22.

Using the fitted values as a baseline, I then compare all scenarios to that baseline; results are shown in Figure 23. Most scenarios show no large changes in median income—the number of residents in the new units simply is not enough to move the area median income. However, again in the aggresive low operating cost scenario, median incomes fall throughout the areas that are heavily redeveloped. Since I do not model changes to income at the household level, all changes to median income are due to higher or lower-income populations moving to particular locations.

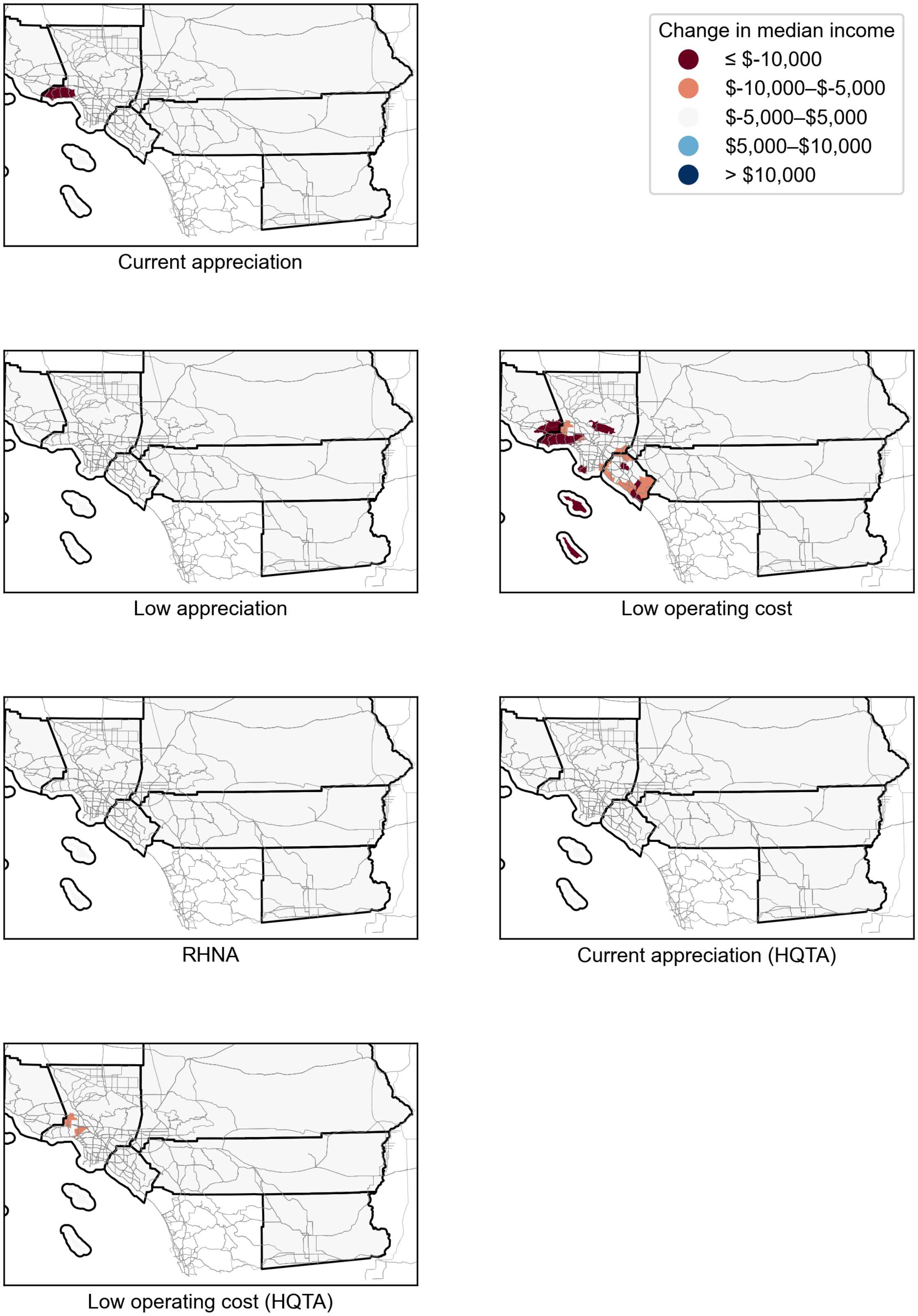


Figure 23. PUMA-Level Median Income, Scenarios vs. Fitted

Another important question is who these new buildings are serving. Figure 24 shows the median income of the residents of new buildings under each of the development scenarios. In many parts of the region, especially the heavily redeveloped areas west of Los Angeles and in Orange County, the median income of the residents of the new buildings is above $75,000. This is concerning from an equity standpoint, although perhaps not surprising; new housing tends to be occupied by wealthier households. Recall, though, that these areas are already quite wealthy; in some areas, these new residents with median incomes over $75,000 actually lower the area median income (Figure 23).

A related question is whether housing policy is creating diverse neighborhoods. Figure 25 shows the changes to the interquartile range of income after the scenario. A larger interquartile range means more income diversity. Many of the areas where redevelopment is concentrated see a reduction in the interquartile range of income after simulation—even those that also see a reduction in their median income. This is likely because the newcomers have slightly below-median income for the area, but their income is still above the 25th percentile. Since they are similar to a large swath of the existing population, the income diversity declines even as the median income declines.

4.6 Limitations

The key concern anytime a statistical model is used for forecasting is that correlation does not imply causation; just because the model has estimated particular relationships does not mean that those are causal. Causality could be reverse, or due to an additional variable not included in the model. While causal inference techniques are widespread in economics, most travel demand models use regression models and try to defend against issues of causality

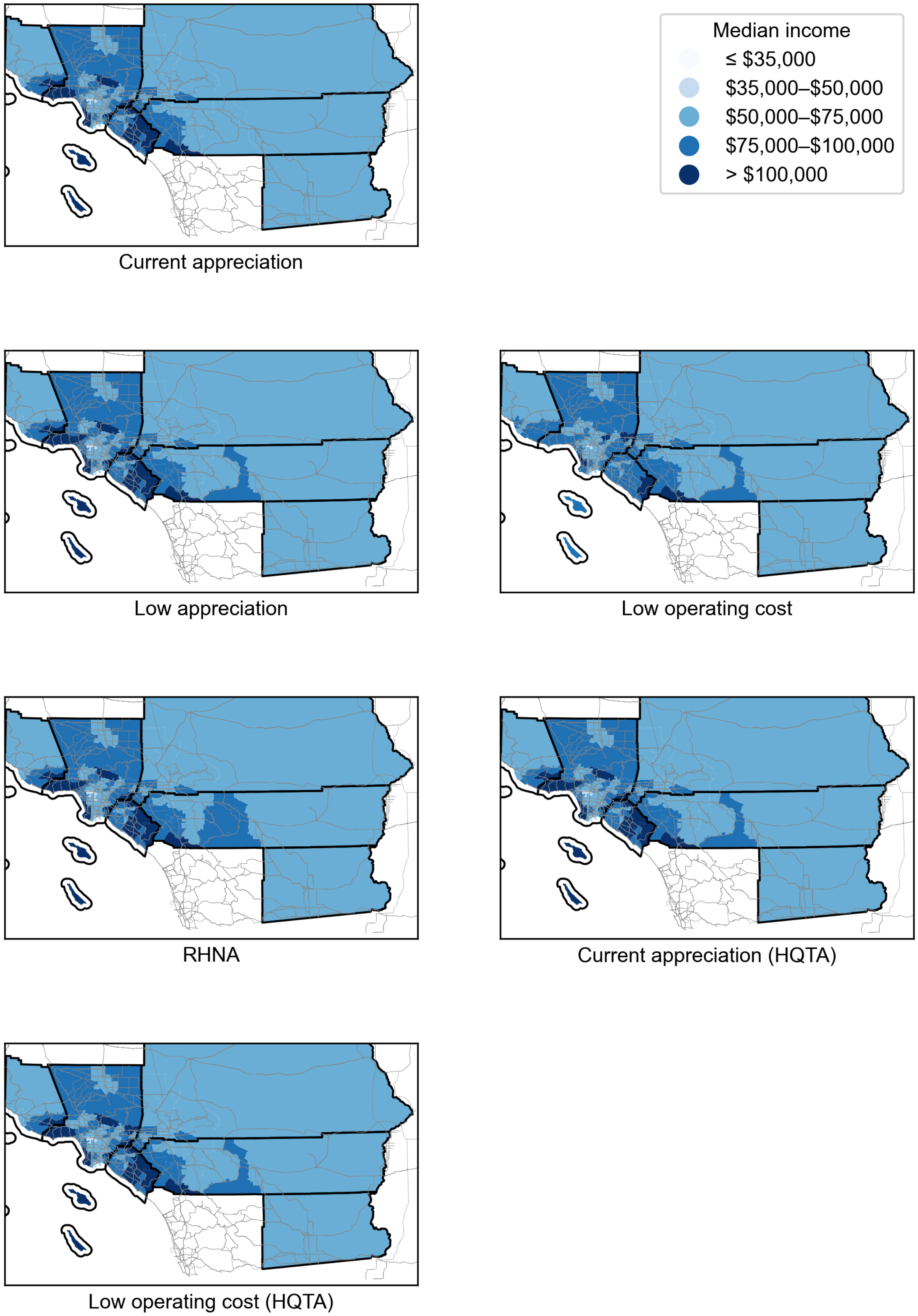


Figure 24. PUMA-Level Median Income, Residents of New Buildings Only

Figure 25. Change in Interquartile Range of PUMA Income

by including a broad range of covariates that are theoretically connected to the outcomes of interest. While this model does include many covariates, there are additional ones that could be included to improve the predictions.

The car ownership model assumes that there is a relationship between the built environment and car ownership. While this has repeatedly been shown to be true, much research also shows that estimates of such a relationship can be biased due to self selection. Households predisposed to own fewer cars, for instance, may preferentially live in denser, walkable neighborhoods. To the extent that those preferences are not reflected in the sociodemographic features in the sorting model, the effect of moving to a different neighborhood on car ownership may be overestimated. Cao, Mokhtarian, and Handy (2009) reviewed a number of studies and concluded that there is an effect of the built environment on transport outcomes exists, there is also an effect of self-selection. They describe several strategies for accounting for this omitted variable bias. In future research, these strategies could be employed to improve the estimation of car ownership.

Furthermore, the car ownership model does not consider parking availability, which has a significant causal relationship with car ownership (Millard-Ball et al. 2020). This is a rather small concern as the vast majority of housing in Southern California does have off-street parking, but is a limitation of the model. The development scenarios also do not vary the amount of parking available per unit. If future research did vary parking levels, accounting for parking in the residential location choice model would be critical.

Additionally, this model assumed away any dynamics of migration or household formation. Additional housing construction would likely result in both of these processes, as people from outside the region migrate in and households within the region split up (for instance, roommates or adult children living at home move out). This model scaled the number of new units to match the existing household population, which is defensible if one assumes the future population will be similar to the existing population in terms of sociodemographic distribution. Immigration and household formation could, however, change sociodemographic distributions.

4.7 Discussion and Conclusion

This chapter has demonstrated the application of an equilibrium sorting model to simulate a land-use change and its effects on transport outcomes. While equilibrium sorting models have been used in a (small) number of papers in economics, they have rarely been used in land use-transport interaction research. Thus, an important contribution of this chapter is simply to demonstrate these models in this type of application. Their ability to simulate changes to outcomes that result from residential relocation in response to a shock to the housing supply is something that other models of residential location choice cannot easily do.

Additionally, I have applied the model to a class of problem I have not seen it applied to in the past—an exogenous change in housing supply. This requires appropriately accounting for the size of aggregate alternatives, something that many authors thus far have not addressed. Given the current policy discussion around increasing housing supply to ease affordability problems, models that can evaluate such changes are valuable. Equilibrium sorting models are one option.

The sorting model suggests that the residents of the new development will be relatively high-income, although in some cases they are still lower-income than the existing residents of the area they are moving in to. While gentrification is less of a concern since redevelopment largely occurs in relatively high-income areas, it is also true that the new housing is not directly easing housing housing supply for the lower-cost end of the market. It could still benefit this market segment indirectly, in the short term by reducing pressure to gentrify lower-income neighborhoods, and in the long term through housing filtering, but direct provision of housing for lower-income households in general would require government subsidy.

Chapter 5

POPULATION SYNTHESIS

Activity-based travel demand models simulate the travel demand of a synthetic population of households with associated sociodemographic and geographic characteristics. Many models create synthetic populations for small areas based on marginal distributions of demographic characteristics from small areas combined with public use Census microdata from larger areas (e.g. Konduri et al. 2016). This approach is not possible in this research because the small-area marginal control distributions are no longer valid once households have resorted in response to the housing supply shock.

The output of the sorting model cannot be used directly, because the geographic specificity is too coarse; while there are only just over 100 PUMAs, most travel demand models have thousands of zones. Within-PUMA travel distances are much too variable for meaningful modeling of mode choice, destination choice, etc., and the majority of trips likely remain within a single PUMA. These results must be disaggregated for use in the travel demand model.

While sociodemographic control totals are not possible due to household re-sorting, I do have exogenous marginal distributions of housing types in each tract, since the development model presented in Chapters 2 and 3 are disaggregate. I use these distributions to disaggregate the PUMA-level results from the sorting model to the tract level. Up until this point, I have not assigned households to specific residential locations, tenure choices, or vehicle ownership levels, but have rather retained the probabilities that each household would would make each possible choice. For the travel demand model, I need to assign households to specific choices, so I also do this as part of the population synthesis process.

5.1 Methods

The output of the sorting model is the probability that every household will choose each combination of PUMA, age of housing, multi- or single-family housing, tenure choice, and vehicle ownership level. I control for age of housing and multi- or single-family housing at the tract level. I calculate the number of housing units in every tract in each of four categories, pre-2000 and 2000 or later multifamily and single family homes, using 2013–2017 American Community Survey data. I then adjust these unit totals based on the scenarios developed in the previous chapters.

To disaggregate to the tract level, I draw households from the population to match the number of housing units implied by the scenario in a particular tract/age/multifamily-single family bin. I weight this draw by the probability a particular household chooses to live in that tract's PUMA and the age/multifamily-single family bin in question, multiplied by the household's weight in the PUMS. This way, the demographics forecast at the PUMA level are disaggregated to the tract level based on housing characteristics, which may not be homogenous across tracts. It is impossible to use demographic control totals as would typically be used in population synthesis (Konduri et al. 2016, e.g.), since the demographics are expected to change as a result of the sorting model.

Using the probability that a household chooses a particular tenure choice and level of vehicle ownership given that they have chosen the PUMA/age/multifamily-single family bin they have been drawn into, I assign each drawn household a vehicle ownership level and a tenure choice. Since the sorting model only forecasts vehicle ownership in 4 categories (0–3+), and the travel demand model uses 7 (0–6+), I disaggregate the 3+ category in simulation to 3, 4, 5, or 6+ vehicles randomly according to the proportions of 3+ vehicle households with each of these levels of vehicle ownership in the IPUMS. I repeat this population synthesis process for each scenario described in the previous chapter, since the number of housing units in each bin varies by scenario.

While most of this dissertation was implemented in Python, the core of the population synthesis algorithm is implemented in Julia (Bezanson et al. 2017) due to its superior performance on very large datasets (in this case, probabilities that each household chooses every possible combination of tract, single/multi-family home, pre- or post-2000 construction, and vehicle ownership), using SQLite to manipulate larger-than-memory data. The population synthesis process takes approximately two hours per alternative, on a quad-core Intel i7 with 16GB RAM and a 1 TB magnetic SATA disk.

5.2 Results

In order to evaluate the quality of the population synthesis process, I compare the characteristics of the synthetic population generated using housing characteristics and forecasted choices from the baseline scenario (i.e. before the change to housing supply). This comparison is particularly important in this case, since demographic controls have not been used as they would be in a traditional population synthesis. Therefore, it is not guaranteed that the synthetic population will be demographically similar to the actual population.

Figure 26 compares the synthetic population to 2012–2017 ACS proportions for age and sex within each county.[17] By and large, the synthetic population is quite close to the control totals, indicating that both age and sex distributions are well controlled by the synthesis pro-

[17] Control totals are the portion of people in a particular age/sex category out of the entire population, so this graphic also indicates a balance between males and females.

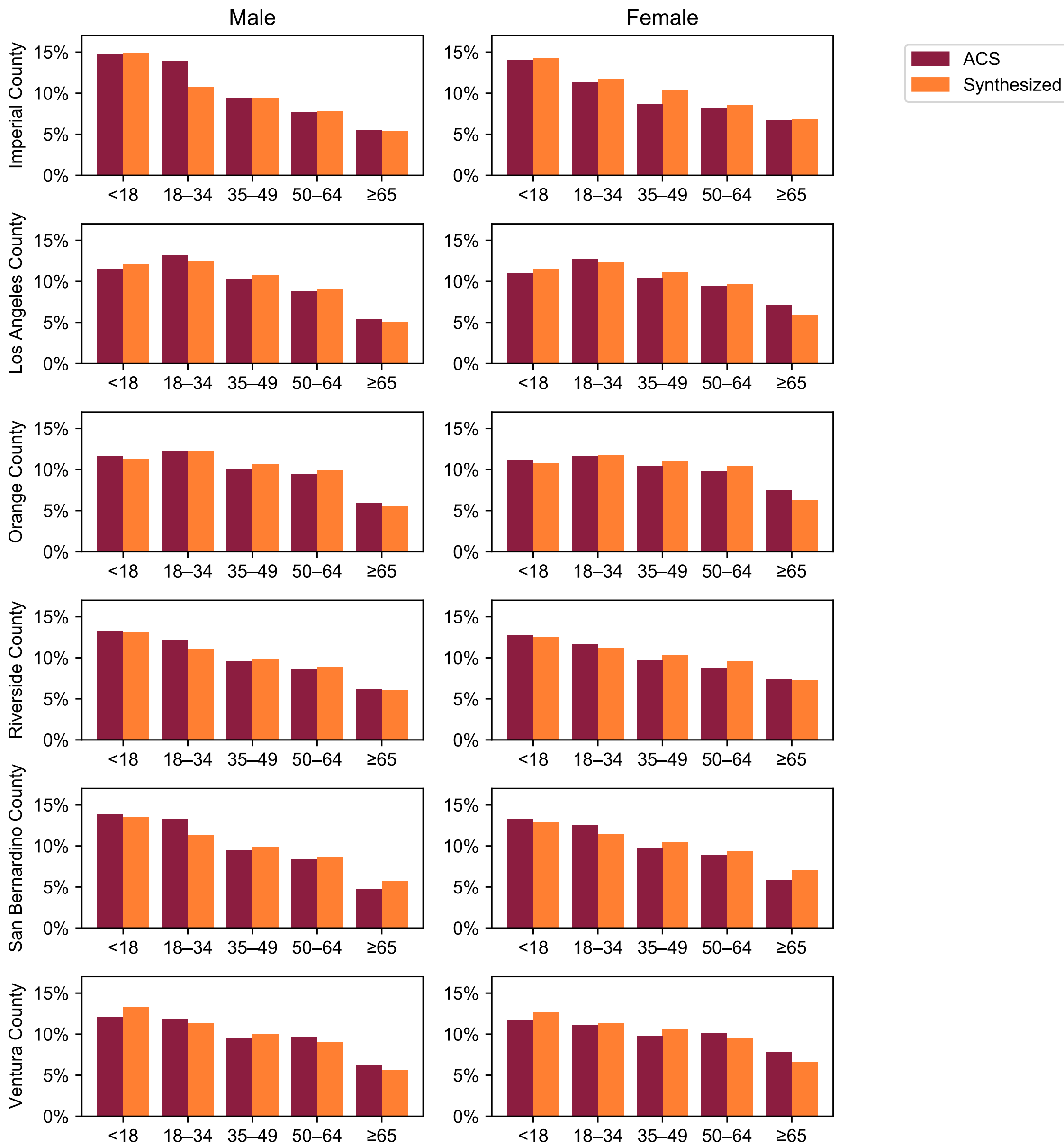


Figure 26. Comparisons of Actual and Synthetic Populations by Age and Sex

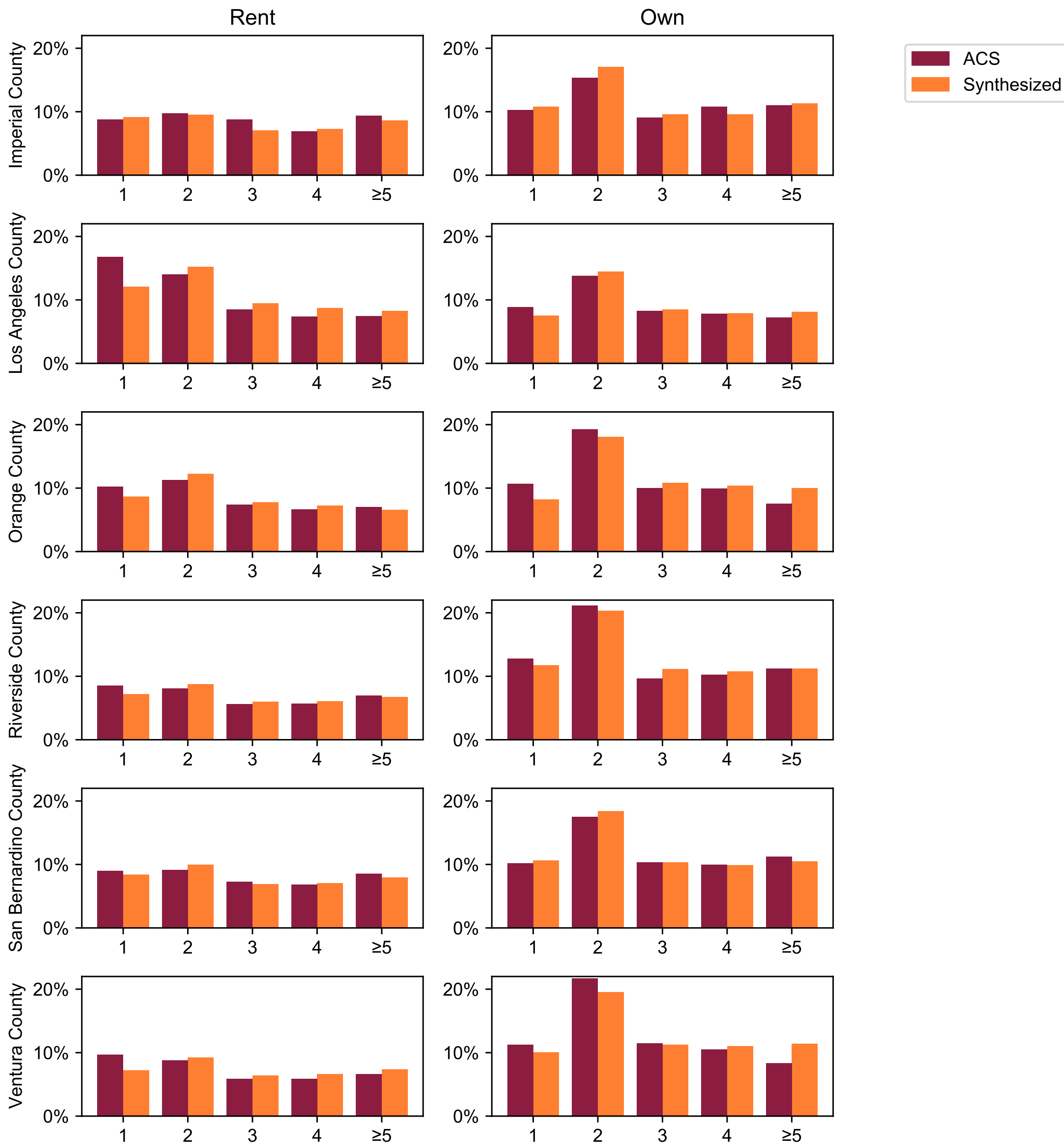


Figure 27. Comparisons of Actual and Synthetic Populations by Tenure and Household Size

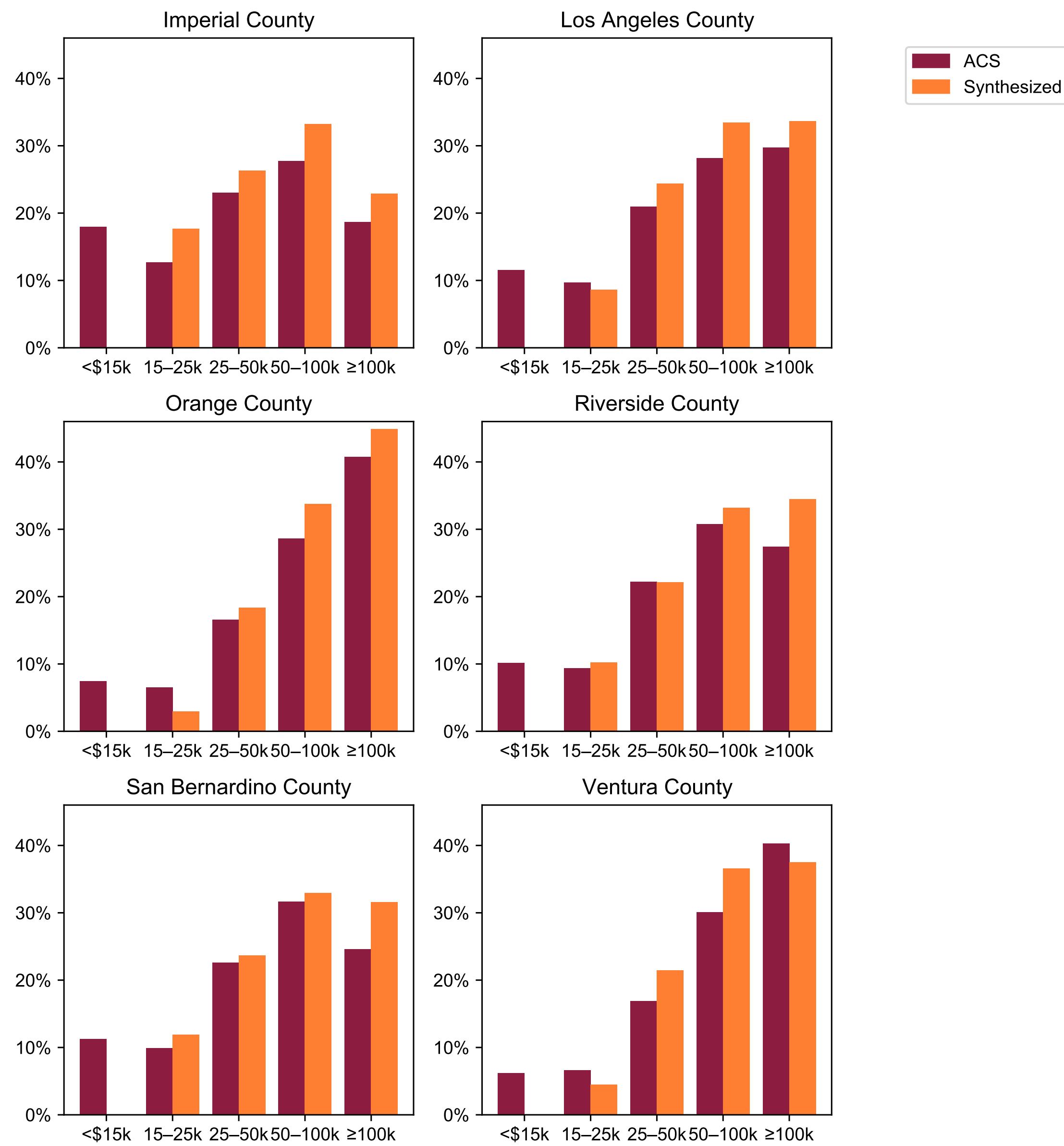


Figure 28. Comparisons of Actual and Synthetic Populations by Income

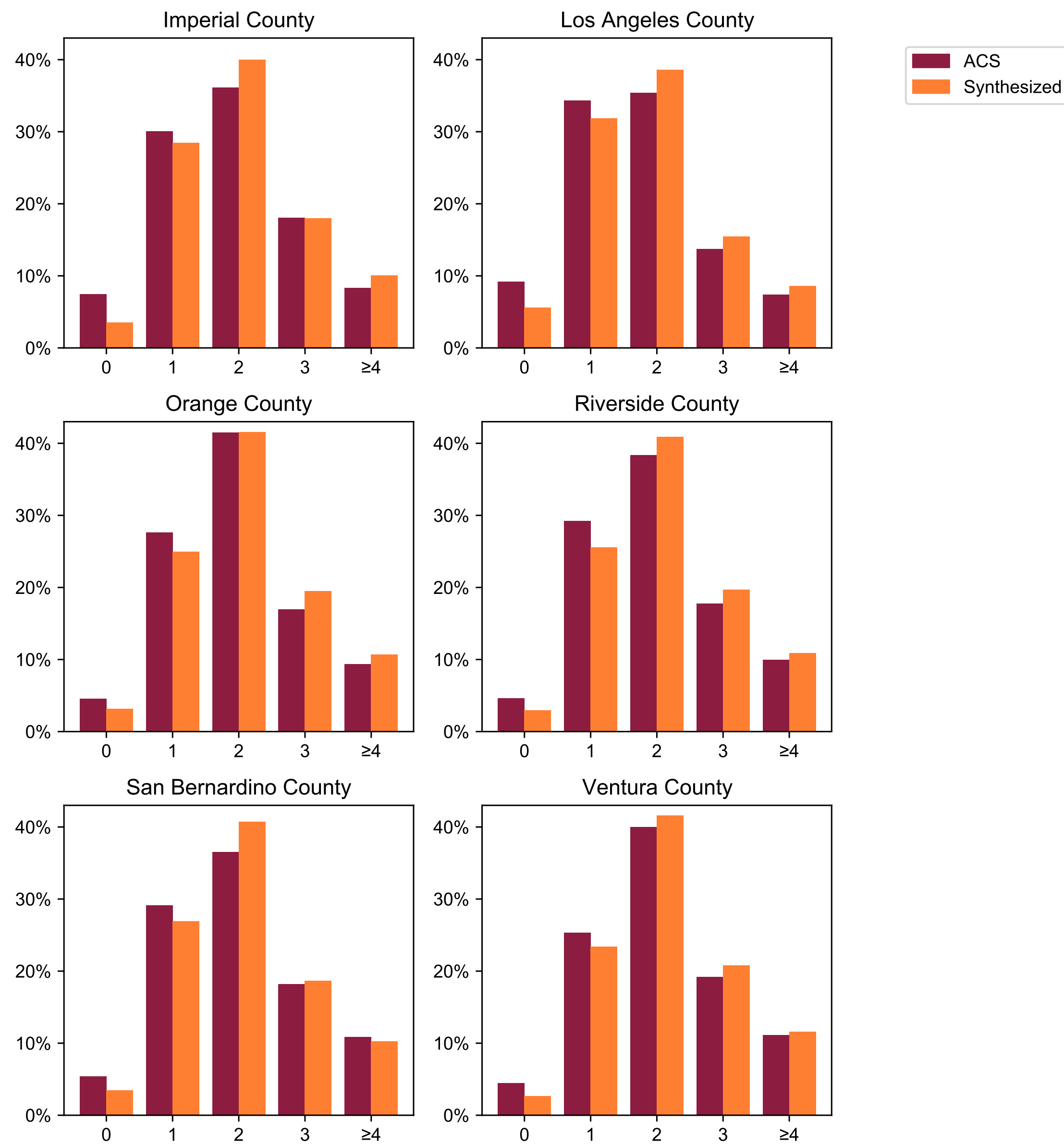


Figure 29. Comparisons of Actual and Synthetic Populations by Vehicle Ownership

cess. Figure 27 compares distributions by tenure and household size, and again the synthetic population does a good job of replicating the observed population.

Figure 28 compares the incomes of the synthetic population with those of the overall population. The major deviation from representativeness here is that there are no households making under $15,000 in the synthetic population, because they were removed from the simulation process in order to make the sorting model possible. This is a limitation of this dissertation, and may result in overstated congestion and driving results because households in this income group likely travel less, and less by car, than other groups. This is likely also the explanation for why household car ownership is somewhat higher in the synthetic population vs. the actual population (Figure 29); the lowest decile of income represents 41% of carless households in the US (King, Smart, and Manville 2019).

The discussion thus far has focused on county-level demographic control totals, which are satisfactorily matched by the synthetic population. However, since I am disaggregating to the tract level, county-level demographics may dilute representativeness at smaller scales. Figure 30 plots tract-level median income from the synthetic population and the actual population. As discussed above, the synthetic population generally has higher tract-level median incomes than the actual population, and there is less variation in median incomes in the synthetic population than there is in the actual population, since the model does not perfectly predict household location choices. However, the higher income tracts tend to align in both the actual and synthetic populations, and the tract median income of the synthetic population predicts observed tract median income with an R^2 of 0.42.

Overall, the synthetic population generator does a relatively good job of drawing households that are similar to the actual population in the baseline scenario, and thus it can be used to generate synthetic populations resulting from the sorting process under each of the supply changes examined in the previous chapter.

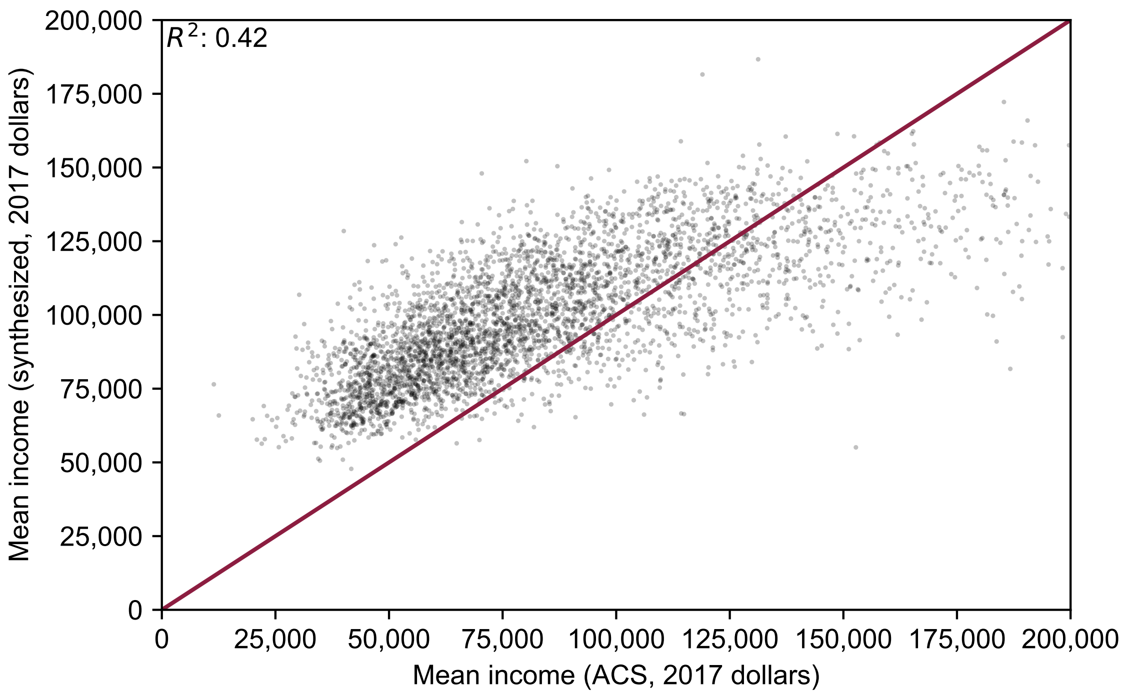


Figure 30. Comparison of Tract Median Income, Actual and Synthetic Populations

Chapter 6

ACTIVITY-BASED TRAVEL DEMAND MODEL

To more fully understand the impacts of the modeled changes to land use on travel, I use an activity-based travel demand model to simulate travel across Southern California, using the residential location and vehicle ownership choices described in Chapter 4. I use the ActivitySim framework to implement this model (Association of Metropolitan Planning Organizations 2019). ActivitySim comes with an implementation of the Metropolitan Transportation Commission (San Francisco Bay Area)'s Travel Model 1 (Erhardt et al. 2012). I put the outputs of the sorting model into the format required by this model, and apply it to understand the granular transportation effects of the scenarios simulated in Chapter 4. Due to household travel survey data availability, however, I do not reestimate any of the models for Southern California. I use static traffic assignment algorithm to derive congestion metrics from the model output.

Specifically, I evaluate shifts in mode choice, vehicle kilometers traveled, and traffic congestion. The discrete-choice models in ActivitySim simulate the school, work, and other destination choices for all tours taken by households in the synthetic population. ActivitySim models work and school location, as well as primary and secondary trip destinations for non-mandatory tours (ActivitySim team 2020). Due to cost and time constraints, I only model the most aggressive development scenario, the low operating cost scenario.

6.1 Model Inputs

Travel Model 1 expects four types of inputs: households, persons, land use, and skims. Households and persons come from the synthetic population generation process described in Chapter 5. Land use data are derived from existing land use and outputs from the profitability and sales models in Chapters 2 and 3, and are described in detail in Section 6.1.1. Skims, or zone-to-zone travel times, are estimated using various data sources and are described in Section 6.1.2.

6.1.1 Land Use

ActivitySim requires a broad array of land use variables for each analysis zone, Census tracts in my case. Some of these, such as total households, total population, and population in each age band, are calculated based on the synthetic population. Total land acreage comes from the Census Bureau, and total residential and commercial acreage are derived from the SCAG land-use dataset. Residential vacant properties are additionally included in the total residential acreage, as they may be redeveloped under the scenarios developed above. Employment statistics are derived from the US Census Bureau LEHD LODES dataset, and are assumed constant in all scenarios. High school enrollment is derived from 2016–2017 enrollment numbers from the California Department of Education (California Department of Education 2017) and geocoded using the California School Campus Database (California School Campus Database 2021). College enrollment comes from the Integrated Postsecondary Education Data System (National Center for Education Statistics 2017).

There is an urban area type category which is documented to be based on a linear combination of population and housing density (Ory 2016). However, the urban area types in the example dataset do not match the documented density categories. Furthermore, there is no clear density-based division of the example dataset into urban area zones. Some suburban zones, for example, are more dense than other zones classified as urban. Since area types are almost always contiguous across multiple nearby analysis zones in the example dataset, I use an ad-hoc smoothing procedure where each analysis zone's area type is first assigned based on density, and then replaced with the most common area type among itself and all adjacent area types, using a Queen's case adjacency matrix calculated with PySAL (Rey and Anselin 2007). The topography of the tract is present in Travel Model 1 in three categories from steep to flat. Using 1/3-arc-second digital elevation models from the US Geological Survey 3D Elevation Program (US Geological Survey 2021), I calculated the slope of each cell using GDAL (GDAL/OGR contributors 2021), and used Zonal Statistics in QGIS (QGIS Project 2021) to compute the median slope in each tract. While the exact definitions of the three slope categories in Travel Model 1 are not, to my knowledge, documented, using the median slope in the analysis zone with cutpoints of 1.68% and 2.77%, correctly classifies 91% of the analysis zones in the example Travel Model 1 input data. These cutpoints were determined using a trivial decision tree model estimated using `scikit-learn` (Pedregosa et al. 2011).

I assume that all land uses except housing stay constant between the baseline and the scenario. This means that any forecast travel changes are due to people moving to more accessible areas, not areas becoming more accessible as land use changes—for instance, due to stores opening in a now more densely populated area. In particular, this is problematic for high school enrollment, as school districts will likely adapt fairly quickly to changes in population distribution; distances to high school may be overestimated by the model. While ActivitySim supports capacity restrictions on school and work locations through "shadow pricing," which adjusts utility of these destinations to meet capacity, I do not use this feature. This somewhat ameliorates these concerns about land use and school enrollment in particular staying constant between the baseline and the scenario.

6.1.2 Skims

Skim matrices are zone-to-zone travel time matrices. This model uses the 3956 Census tracts in the SCAG region as the zones of analysis. There are skims for walking, bicycling, driving, and transit, which represent travel times and distances between tract centroids. For intrazonal trips, I assume travel times and distances are 1/2 of the travel time to the nearest neighboring tract, as was done for highway skims in the Bay Area deployment of Travel Model 1 (Ory, Tsang, and Zorn 2019, line 114).

6.1.2.1 Walking and Bicycling

Walking skims are calculated by computing the distance on foot from the centroid of each zone to every other zone, using data from OpenStreetMap and the Open-Source Routing Machine (OSRM) routing software,[18] (Luxen and Vetter 2011) with default walking settings as shipped with OSRM. Bicycling distance was assumed to be equal to walking distance, although different travel speeds are of course assumed by the model.

6.1.2.2 Driving

Driving skims are also calculated using OSRM. Since data on segment-level traffic congestion are not available at a price within this researcher's means, only free-flow travel times

[18] Version 5.24.0 was used for all computations herein.

can be estimated by OSRM. However, Los Angeles is infamous for its traffic congestion, to the point that it is a key element of popular culture (Jackson 1991; D. Chazelle 2016). Traffic congestion indubitably plays a role in travel decisionmaking, and accounting for it is critical. Thus, I adjust free-flow skims to account for congestion in each time period using data from Uber Movement (Uber 2020). This dataset provides average weekday tract-to-tract travel times for the immediate environs of Los Angeles, at each hour of the day. However, it does not provide spatial coverage across the full SCAG region. Thus, I use this dataset to create a model to predict the level of traffic congestion for a particular O-D pair of tracts at a particular time of day.

The dependent variable in this analysis is the ratio of congested travel time to free-flow travel time, which I assume is equivalent to travel time in the overnight hours. The covariates I use to predict the congested travel time ratio are the hour of the day, the network distance between the tracts as computed by OSRM, and population and employment density for several relevant areas. Specifically, I consider population and employment density in the origin and destination Census tracts, and the 25th, 50th, 75th, and 95th percentile density of tracts in 2-km bands from 0–8 km around each origin and destination tract, and in 2-km buffers from 0–8 km around the straight line connecting the origin and destination tract centroids (to account for possible congestion along the way). To avoid including a tract in multiple bands, and to speed computation, a tract is included in a band iff its centroid lies in that band. The process of determining band covariates is computationally intensive, given the number of possible combinations of tracts and the complexity of the spatial mathematics, so it is run using multiprocessing on a quad-core Intel i7 with 16GB of RAM.

This results in 126 covariates, a large number for an econometric model. However, since the goal of this model is to predict rather than to explain, I turn to machine learning and use a random forest model to identify relationships between these covariates and congestion.

Table 15. Time Windows Used in Activity-Based Model

Name	Start time	End time
Early morning	3:00 AM	6:00 AM
Morning	6:00 AM	10:00 AM
Midday	10:00 AM	3:00 PM
Afternon	3:00 PM	7:00 PM
Evening	7:00 PM	3:00 AM

Random forests estimate a large number of decision trees on random subsamples of the data, and then average the results to create a prediction (James et al. 2013, ch. 8). In this case, I use 100 trees as this provides good results in reasonable computation time. For tractability, I fit the model on a sample of 100,000 Census tract-hour of day pairs from the Uber Movement dataset; the actual estimation sample is slightly smaller, as I exclude 200 randomly selected tracts from the estimation in order to use them to evaluate model fit. I used `scikit-learn` to fit the model (Pedregosa et al. 2011).

The test R^2 of this model is 0.67. Test R^2 is computed on the tract-hour pairs that were not were not included in the model estimation, to avoid reporting a biased value due to possible overfitting. Since congestion levels are indubitably correlated for trips leaving from or going to the same Census tract, I excluded 200 randomly-selected Census tracts from the training data entirely, and computed the test R^2 based only on congestion levels among these tracts.

The ActivitySim/TM1 model estimates travel in 5 time windows (Table 15). I use the model to forecast the congested travel time ratio for all tract pairs in the region for each time window, using a point near the middle of the time window. I multiply the free-flow travel time estimated by OSRM by this predicted value to produce an estimate of congested travel time for each pair of Census tracts in the region.

It is possible that the test R^2 overstates the predictive power of the model, for two reasons.

First, the observations are not independent, so the training process may have seen similar observations (for the same Census tract pair at a different time, for example). Second, the model is extrapolating information from Los Angeles County to other Southern California counties. This is particularly a concern, since the highest-importance feature in the model by far is the hour of the day, so the model may assume Los Angeles County levels of congestion in remote Imperial County.

ActivitySim requires estimates for HOV and toll lane travel times in addition to single-occupant vehicle travel times. I assume that HOV and toll lane travel times are 10% below the estimated SOV travel time in the AM and PM peak periods, and equivalent during the off-peak. Most toll facilities in the SCAG region are toll lanes rather than toll roads, so off-peak there is little differentiation in travel time between tolled and untolled facilities.

I assume that tolls are charged at \$1.57/mile during the peak, and \$1.48/mile offpeak, based on the cost to traverse the entirety of the California Route 73 toll road (The Toll Roads of Orange County 2021). I assume that ¼ of the distance of each tolled trip takes place on tolled facilities, since these facilities are relatively rare within a significant network of untolled freeways.

6.1.2.3 Public Transport

To compute tract-to-tract travel times, I use my open-source public transport routing package TransitRouter.jl[19] combined with General Transit Feed Specification data from across Southern California, representing public transport schedules in early 2018—roughly

[19] https://github.com/mattwigway/TransitRouter.jl

in line with the vintage of other data in the model.[20] I compute travel times by public transport at a time near the middle of each time window (Table 15), for walk and drive access and egress from public transport (walk and drive access and egress times are computed with OSRM). I allow walking up to 1km to the first stop, from the last stop to the destination, and at each transfer. For driving access and egress, I allow travel up to 20km. Due to technical limitations, it is not possible to account for traffic in driving access and egress, so they are assumed to have free-flow travel times.

I derive several metrics about every trip, including in-vehicle time, wait time, walk time, driving time, and number of transfers. Since different public transport modes may attract different segments of the population, Travel Model 1 requires trip information for trips by several public transport modes. Since many trips combine multiple modes, I allow access to a primary mode by other modes; which modes I allow to be used to access each primary mode are shown in Table 16.

ActivitySim/Travel Model 1 requires an estimate of fares for each skim. I simply assume a flat fare of \$1.75 for local bus, light rail, and heavy rail; \$2.50 for express bus, and \$5.00 for commuter rail, based on current LA Metro and Metrolink pricing. In reality, fares are much more complicated and there may be multiple optimal tradeoffs between travel times and fares (Conway and Stewart 2019), but this detail is not represented in this model. Furthermore, some residents may choose to purchase a public transport pass allowing unlimited rides, which changes the marginal costs of choosing transit; this detail is not modeled by ActivitySim.

[20]Many thanks to Evan Siroky and Sasha Aickin for providing me with access to this curated dataset of schedule data.

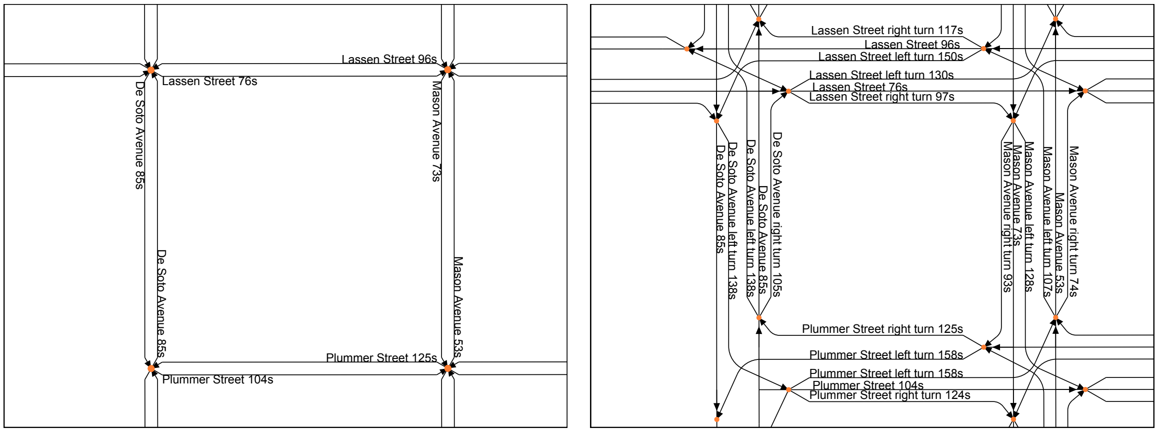
Table 16. Allowed Access Modes for Each Primary Public Transport Mode

	Access modes				
Primary mode	Local bus	Light rail	Heavy rail	Commuter rail	Express bus
Local bus	✓				
Light rail	✓	✓			
Heavy rail	✓	✓	✓		
Commuter rail	✓	✓	✓	✓	
Express bus	✓	✓	✓		✓
All public transport	✓	✓	✓	✓	✓

6.2 Methods

6.2.1 Travel Behavior

ActivitySim and Travel Model 1 are a series of discrete choice models that build on each other to microsimulate many household and personal decisions, including mode choice, departure time choice, and destination choice. ActivitySim coordinates travel decisions at the household level for additional realism; more details are available in (Waddell et al. 2018). I make two slight modifications to Travel Model 1 before applying it to the SCAG region. First, I remove the auto ownership model, as auto ownership is estimated by the sorting model. Secondly, I remove the availability of free parking model and simply assume all trips have free parking—true for the vast majority of trips in the US. I ran travel model simulations on a 96-core AMD EPYC system with 768 GB RAM rented from Amazon Web Services; each run took approximately 6 hours and cost approximately $32 for computation.



a. b.

Figure 31. Normal (Left) and Turn-Based (Right) Graph Representations of the Street Network

6.2.2 Congested Network Assignment

One of the outputs of ActivitySim is a trip matrix; that is, the forecast number of trips between each pair of zones during each time period of the simulation. While I used data from Uber Movement to create skim matrices to use in model estimation, I cannot use that data forecast congestion under a future transport scenario. Instead, I extract the number of vehicle trips for each O-D pair at each time of day, and perform a static traffic assignment. To find the congested equilibrium, I use a Frank-Wolfe iterative algorithm, which iteratively updates link flows and costs until an equilibrium is found (Ortúzar and Willumsen 2011, 398; Rambha 2020; Boyles 2019). I compute congestion metrics for both the base scenario and the low operating cost scenario using this algorithm, so that results are comparable.

The network for the Franke-Wolfe algorithm is derived from OpenStreetMap (OpenStreetMap contributors 2020). Like Waddell et al. (2018), I retain all tertiary roads in my network, as well as unclassified roads[21]. I use my open-source OSMPBF.jl[22] Julia library to read the OpenStreetMap data and convert it into a routable network.

The most obvious representation of a road network in graph format is to make each intersection a vertex and the roads between them edges (Figure 31a). However, since most routing algorithms for weighted graphs do not support costs assigned to verices, only edges, this does not lend itself to representing the time required to make turns. Instead, I represent the street network as a turn-based or *dual graph*, where each street segment is represented by a vertex, and connections/turns (via intersections) to other street segments are represented as edges (Figure 31b). This allows me to attach both the traversal time of the segment and the turning time at the end of the segment to each edge, because each edge implies a particular turn (Winter 2002).

Any traffic assignment algorithm requires some estimation of congested travel times on the links in the network, which is generally a function of their free-flow travel time multiplied by some factor based on the volume/capacity ratio of the link. I derive free-flow travel times by roughly duplicating the travel time computations used by OSRM, which was used to derive the skims. I adjusted travel times for congestion using a simplified version of cost function used in the existing SCAG four-step model (Southern California Association of Governments 2012); most notably, I did not differentiate capacities based on area type or widths of intersecting streets. I cannot use Uber Movement data here because it only represents origin-destination travel times, not the link-level travel times needed for traffic assignment.

I use a slight variation of the typical Frank-Wolfe assignment process due to my use of a

[21]In OpenStreetMap parlance, "unclassified" is, confusingly, a roadway classification that is below tertiary roads but above residential streets.

[22]https://github.com/mattwigway/OSMPBF.jl

turn-based graph. In a turn-based graph, each street segment is represented by a vertex, and the set of edges that connect that vertex to others models traversal of that street segment as well as any turn or intersection costs. Therefore, one physical street segment is typically represented by multiple edges in the turn-based graph. To address this, I calculate the congestion on each edge using the flows from all of the edges making up one physical street segment.

I implemented the traffic assignment in Julia using the LightGraphs package (Bromberger, Fairbanks, and contributors 2017), as well as the ForwardDiff package for automatic differentiation (Revels, Lubin, and Papamarkou 2016). I ran it on a 64-core, 128 GB arm64 server rented from Amazon Web Services, although such a large machine is not required. The most computationally intensive part of the assignment process is using a routing algorithm to derive the fastest trips from each origin to every destination after every iteration of the algorithm. This process is easily parallelized, so speed scales roughly linearly with the number of cores. As suggested by Boyce, Ralevic-Dekic, and Bar-Gera (2004), I run the process until the relative gap (which is an estimate of how far aggregate observed travel times are above optimal travel times) is less than 0.01%. Results converged after 730–740 iterations in about 13 hours, costing approximately $28 for computation. I use the vehicle trip matrices produced by ActivitySim for trips between 6 and 10 AM. Since the cost functions are based on hourly flows, I multiply these vehicle flows by 0.35 to estimate the hourly trip demand for the peak hour (assuming that travel demand is somewhat peaked even within the 6–10 AM time window).

I only perform congested assignment for trips by private vehicle; I assume that public transport, walking, and biking have sufficient capacity that they do not become congested. Furthermore, I assume road congestion has already been taken into account in the public transport schedules used to create the public transport skims.

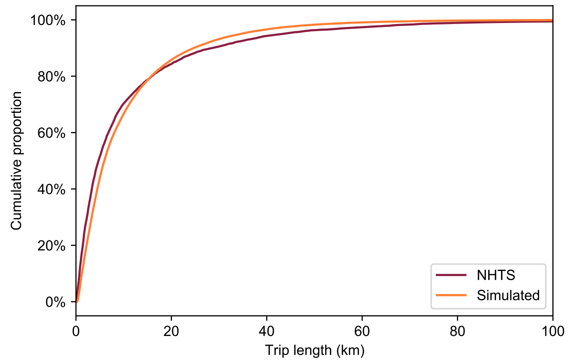


Figure 32. Observed and Simulated Trip Lengths in Southern California

6.3 Validation

To validate the travel demand model, I first ran the model using base conditions, and compared the forecast travel patterns with the mode splits, trip length distributions, and time-of-day distributions to the 2017 National Household Travel Survey. I used the California add-on sample available from the Transportation Secure Data Center (Transportation Secure Data Center 2019), and filtered it to only trips made with the SCAG region by households residing in the SCAG region, to match the households and travel that was simulated. I used 5-day weights to estimate observed behavior on a weekday, to match the target day of the ActivitySim model.

The model does a good job of reproducing observed conditions in Southern California. As Figure 32 shows, the model closely replicates the observed distibution of trip lengths, slightly underrepresenting shorter trips relative to longer trips. Departure times also match closely, although the simulated values show a slightly later AM peak as well as more somewhat evening trips than the observed values (Figure 33)—possibly due to different cultures around working hours and the influence of the tech industry in the San Francisco area where

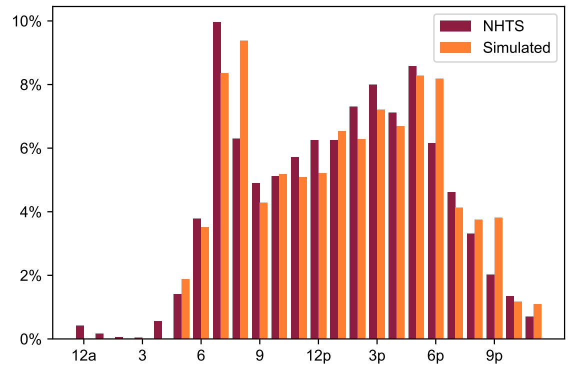


Figure 33. Observed and Simulated Hour of Departure for Trips in Southern California

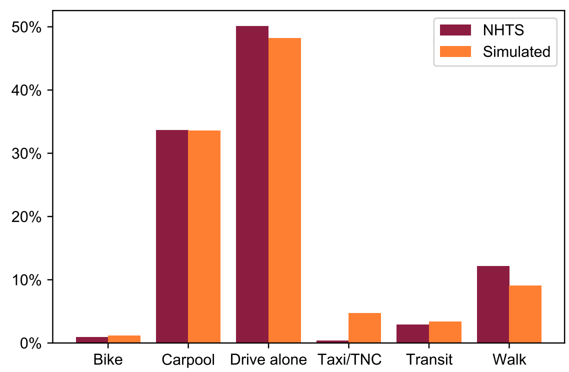


Figure 34. Observed and Simulated Trip Mode Choice in Southern California

the model was originally estimated. The model also reproduces observed mode choice fairly well (Figure 34), although it significantly overrepresents the use of taxis and ridehailing; most of the additional ridehailing trips appear to be drawn from walking and, to a lesser extent, driving alone. This overrepresentation is likely due to the model originally being estimated in the San Francisco area, where ridehailing began and which still has the highest share of ridehailing users in the country (Conway, Salon, and King 2018).

All in all, the model does quite a good job of reproducing baseline conditions even when transferred from one city to another. The main concern when applying a model from one urban area in another one is that the model will not reproduce local conditions; this is not the

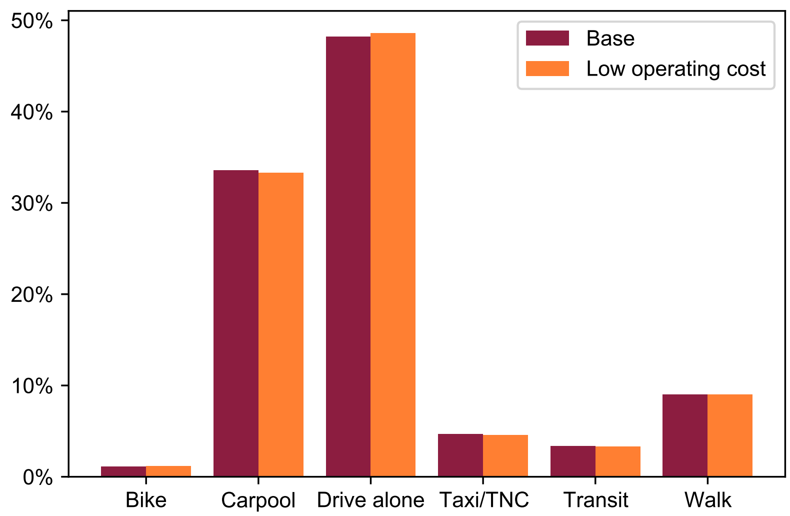


Figure 35. Person-Trip Mode Choice, Low Operating Cost Scenario vs. Baseline

case with the ActivitySim model when transferred to Los Angeles, even without any model re-estimation.

6.4 Results

The travel demand model forecasts minimal changes in travel due to implementation of the low operating cost scenario (the most aggressive redevelopment scenario). Per-capita VKT is forecast to increase very slightly due to the change in land use, from 29.79 to 29.87 km/day. This difference is small enough that it could simply be due to accumulated uncertainty in the models, although a slight increase in VMT is consistent with the slight decrease in job accessibility due to the new units (see Section 3.4). Mode choice changes are essentially nonexistent between the low operating cost and baseline scenarios (Figure 35).

A key concern with any large-scale redevelopment is equity. Even if aggregate travel behavior is stable, if some demographic groups experience significant changes in their travel

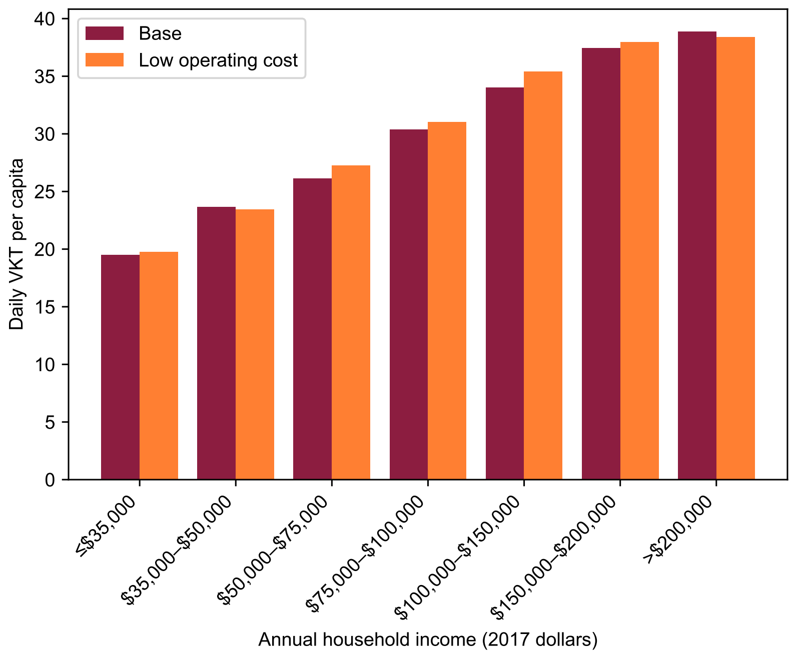


Figure 36. VKT Per Capita, by Annual Household Income

behavior, the outcome could be inequitable. Figure 36 shows the VKT per capita for several income groups. As expected, VKT is significantly higher for higher-income households, but there is relatively little variation in VKT for particular income groups between the baseline and scenario simulations.

6.4.1 Congestion

Using the Frank-Wolfe algorithm described above, I loaded vehicle flows onto the Southern California road network. Baseline levels of congestion forecast by the traffic assignment are shown in Figure 37. As expected, there is heavy traffic congestion the Los Angeles area, especially on the freeway network around downtown Los Angeles (left of center in the figure), with travel times on many freeways exceeding free-flow travel times by more than 50%.

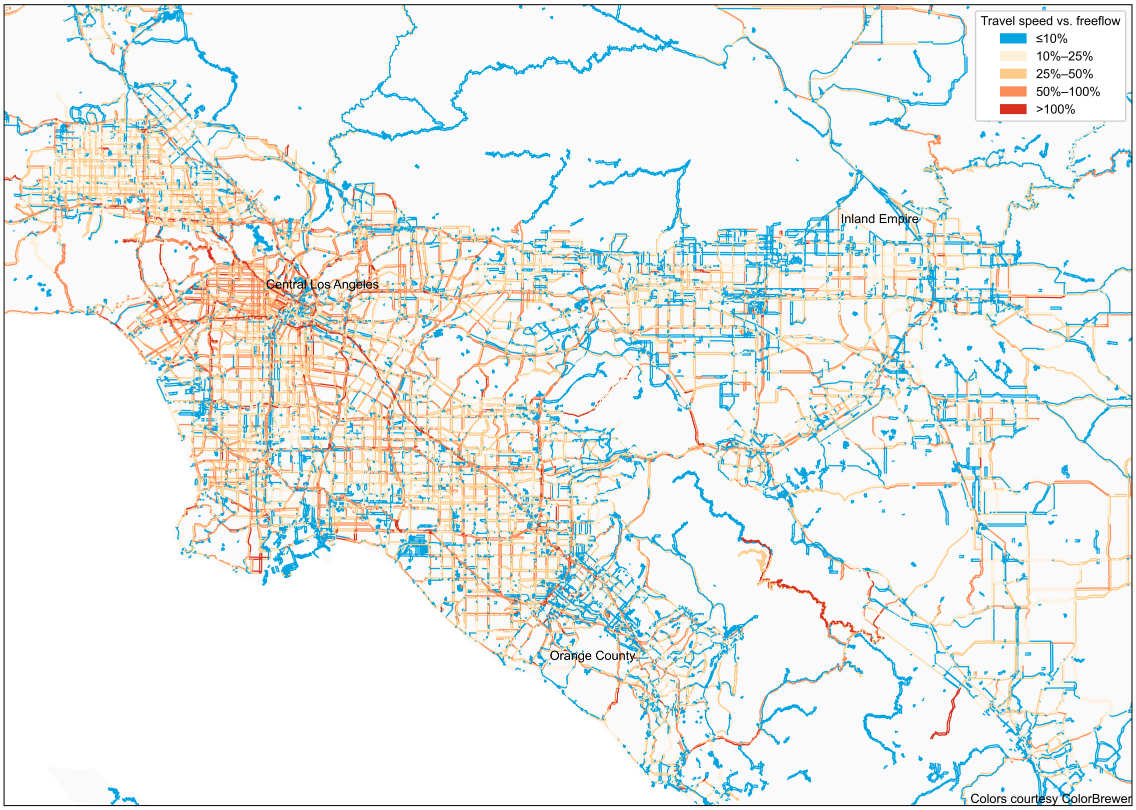


Figure 37. AM Peak Congestion, Baseline, Core of the SCAG Region

Congestion levels do not change much under the low operating cost land use scenario. To make results comparable, the travel models were both fit with populations scaled to be roughly the same size, as was done in the sorting model, so the only variable that changes between the two scenarios is the geographic distribution of residents, not the total population of the region.

Figure 38 shows overlaid histograms of congestion levels for street segments in Los Angeles, under the baseline and low operating cost scenarios. The histograms are remarkably close, almost identical, with a very small amount of road transitioning into the uncongested category. These slight changes in congestion mean that per-capita delay during one hour of

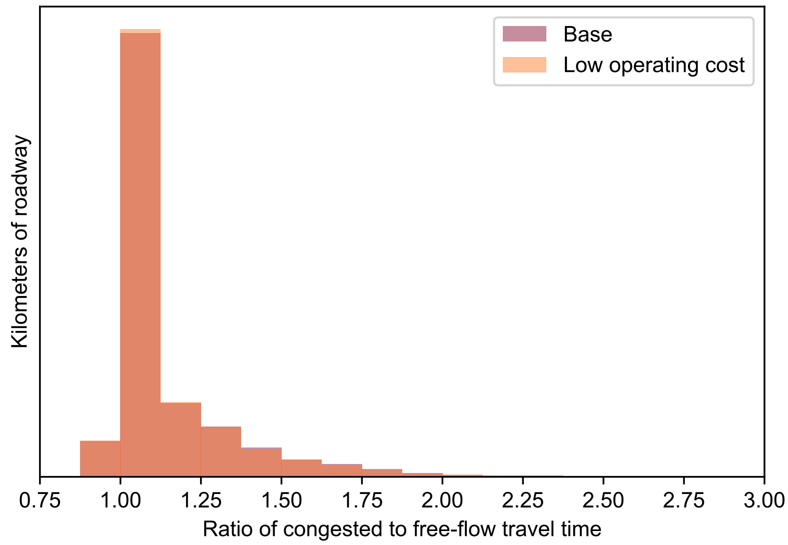


Figure 38. Kilometers of Roadway by Level of Congestion

the peak period decreases from 2.05 minutes to 1.97 minutes.[23] These small changes could be attributable to error in the model, or they could be consistent with development moving people away from the most congested areas.

Computing total time spent in congestion is misleading, however, since it implicitly assumes that free-flow traffic speeds are both achievable and desirable—neither of which is true in most urban areas (Downs 2004), and additionally masks spatial heterogeneity in changes in congestion. Figure 39 shows the change from the AM peak road segment congested travel

[23]These numbers seem very small compared to traffic congestion reports that one sees in the news. The most famous of these is the Texas Transportation Institute's Urban Mobility Report, which estimates that the average auto commuter in the Los Angeles-Long Beach-Anaheim metropolitan area spends 119 hours in congestion per year, and 70 hours in the Riverside-San Bernardino area. The estimate here is the total delay in a single peak hour on a single simulated weekday, divided by the total population of the region. Assuming 250 weekdays per year with 6 congested hours each, this equates to 51 annual hours of delay per capita, not accounting for weekend and off-peak congestion. Approximately 77% of residents of the Los Angeles-Long Beach Combined Statistical Area were in the labor force, and 86% of workers commuted by car, in the 2019 ACS (US Census Bureau 2021). Dividing 51 hours by (0.77×0.86) yields an estimate of 77 annual hours of delay per auto commuter, in the same order of magnitude as the TTI estimates, without accounting for off-peak and weekend congestion.

time in the baseline to the congested travel time in the low operating cost scenario. There are some increased travel times on the west side of Los Angeles and in Orange County, where much of the development from the scenario is occurring. The east side of Los Angeles and the Inland Empire further east show decreased travel times, not unexpected because population has been redistributed further west. However, this decreased congestion is found primarily on freeway entrance ramps. This could be a result of fewer residents accessing the freeway in these neighborhoods, since the scenario redistributes population west (recall that the total population is held constant, so growth in one part of the region implies fewer residents in other parts). Delay on freeway entrance ramps is estimated by a separate equation in the SCAG roadway congestion cost estimation methodology (Southern California Association of Governments 2012), which could also explain why these links are more sensitive to traffic flows.

6.5 Conclusion

These results suggest that a blanket upzoning across Southern California will not, in the aggregate, lead to large scale shifts in transport outcomes. While this may be disappointing to some advocates who hope to use zoning to influence transport outcomes, it should give hope to those who favor increased housing construction to improve housing affordability. These results suggest that such increased construction will not dramatically worsen regional transport outcomes.

These results are only applicable for a blanket upzoning, not a more targeted upzoning aimed to improve transport and environmental outcomes. I only applied the travel demand model to this blanket redevelopment scenario, rather than more of the scenarios presented earlier in this dissertation, due to a lack of funding for computation (each run of the activity-

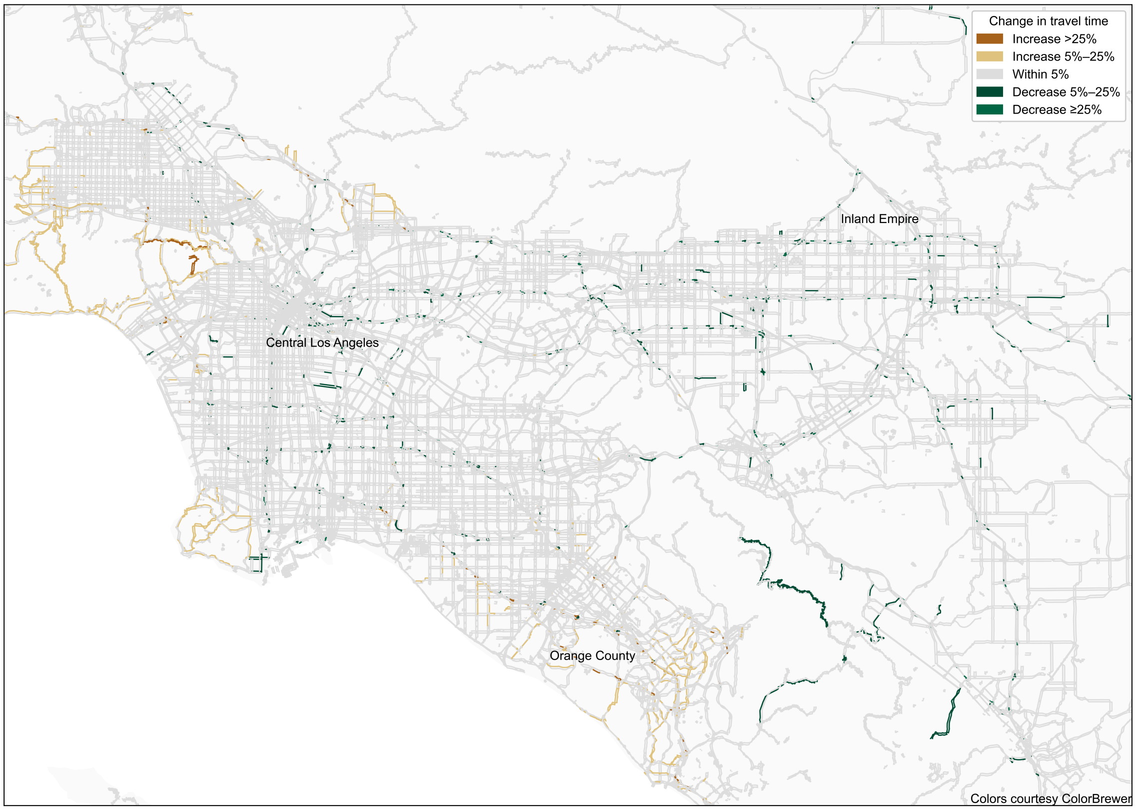


Figure 39. Change in Segment Travel Time, Low Operating Cost Scenario vs. Baseline, AM Peak

based travel model costs about $55 for computation, and each scenario generally requires multiple runs to debug issues with input data). The scenario I ran it for demonstrates the most aggressive levels of development, but the new housing is not in places that have systematically higher access to the region's opportunities than the existing housing stock. It is possible that running the model for one of the other scenarios, particularly one of the ones that targets upzoning to areas near transit, would produce different outcomes; this is left for future research.

While the model reproduces baseline travel in the region fairly well, future work could calibrate the model more closely to observed travel patterns in Southern California. In particular, the model significantly overestimates the use of for-hire vehicles. Additionally, the outputs from the congested network assignment have not been validated against observed vehicle counts and travel times, largely due to a lack of data. Calibrating this model could lead to more accurate results regarding the congestion effects of land use changes.

Chapter 7

DISCUSSION, POLICY IMPLICATIONS, AND CONCLUSION

This dissertation has developed a system of microsimulation models to evaluate the effects of a significant change to housing policy on transport outcomes. From a methodological standpoint, this dissertation has made several key contributions. First, my prototype building approach to construction simulation allows additional detail in the forecasts of future development, especially on small parcels. Data are readily available from building-industry data suppliers to enable this approach to be replicated across the country. Secondly, explicitly accounting for the probability that a property will be sold, and that it will be redeveloped given that it is sold, leads to more realistic estimates of construction than evaluating profitability without accounting for the possibility that a profitable project might not be developed. Thirdly, the use of equilibrium sorting model borrowed from microeconomics is relatively unique in this literature. I have extended this model to jointly model car ownership and to be robust to exogenous shocks to housing supply.

7.1 Implications for Housing Outcomes

This model suggests that there is not sufficient market potential to meet Southern California's housing shortfall purely by converting existing single family homes into multifamily homes, even under the most optimistic scenario. However, converting single family homes to small multifamily homes is only one method of housing provision, and indeed most new multifamily construction in Southern California is in much larger developments. There is also significant potential for accessory dwelling units to contribute to housing supply, especially with new streamlined permitting processes and manufactured units provided by one-stop firms (e.g. Abodu). Combined with other options, easing single-family zoning will positively contribute to solving housing supply shortages.

The model forecasts that development will largely occur in relatively wealthy areas, easing some concerns about gentrification. That said, there could also be short-term negative effects, since the homes most likely to be redeveloped in a particular neighborhood are likely the least valuable ones—a detail too nuanced to be captured in the data used for this model.

The people moving into the new units come largely from high-income households, suggesting that market development cannot directly meet the needs of lower-income households now (though, through housing filtering, it may in a decade or two). Some have argued for an expanded low-income housing tax credit to increase production of homes at the low-income end of the market (Phillips 2020). Furthermore, units that become financially feasible due to these subsidies would likely not displace units that would be built due to market forces, as they occupy different market segments and do not directly compete.

The number of units the model forecasts will be produced varies by two orders of magnitude given different plausible assumptions for key variables. This uncertainty in levels of development is not surprising given the volatile and cyclical nature of the housing market, but it stymies policy analysis. The SCAG region needs to plan for 1.3 million new housing units to meet its Regional Housing Needs Assessment obligations (McCauley 2019), and recently the California Department of Housing and Community Development has taken more interest in ensuring that the units regions plan for are actually feasible (Monkkonen, Carlton, and Macfarlane 2020). However, given the volatility of the development climate, planning for a particular number of market-feasible units is extremely difficult. While interviewing developers to better understand their decision variables could reduce the uncertainty, I suspect much uncertainty is related to the uncertain nature of the underlying housing market.

Due to the cyclical nature of housing supply, Phillips (2020) argues that goverment should build housing and/or fund housing construction when the economy is poor. It is less expensive to build at these times, because labor and material costs are lower. Such a policy would also spread the peaks of housing, reducing the volatility in the market and making housing planning more straightforward.

While market feasibility analysis is increasingly used in housing policy evaluation, it is much less common to evaluate the probability of redevelopment given market feasibility. This dissertation demonstrates that incorporating this effect is crucial, as estimated levels of development drop significantly once this effect is accounted for. While the approach used herein to account for the probability that a developer buys a property is crude, it is preferable to implicitly assuming that the probability is 1. Including this effect can drastically change policy outcomes, and it should not be overlooked.

7.2 Implications for Transport Outcomes

By and large, across several scenarios, I have found minimal influence of this change to housing policy on travel behavior outcomes. There are several possible explanations for this outcome. First, it could be that such changes to housing policy simply do not have large transport effects, at least when implemented at the scale of the entire Los Angeles megalopolis. This certainly seems possible, particularly since the Los Angeles region is relatively car-dependent even in central areas. It could simply be that constructing buildings in places where the market demands them simply does not change how residents choose to get around.

It is also possible that the magnitude of changes I simulate are simply not large enough to show an effect. The central "current appreciation" scenario only results in approximately one additional year's worth of additional construction, and only a small fraction of the total units in Southern California. This theory is lent credence by the fact that there was much more between-scenario variation in vehicle ownership among residents of new buildings than among the population overall—although this could also result from self-selection of low-vehicle households into urban areas in the more heavily urban scenario. That said, the low operating cost scenario results in a forecasted 644,000 new units across the Los Angeles area, which is a very significant change, and this scenario too does not affect aggregate travel behavior (either vehicle ownership, mode choice, or vehicle kilometers traveled) appreciably. That such a large scenario does not cause a change in the the forecasted outcome suggests that it is more likely that travel behavior is largely not affected by blanket upzonings, rather than that the scenarios are insufficiently ambitious.

Finally, it could be that my models of vehicle ownership and travel demand are insufficiently sensitive to effects of the built environment. This seems unlikely, as the sorting model includes a fairly complete set of covariates, and ActivitySim/Travel Model 1 is an industry-standard model designed to be sensitive to important characteristics like the built environment to produce useful forecasts. That said, the model assumes that no land-use changes occur other than new housing. In reality, if more housing is constructed, stores and offices are likely to follow. This model only evaluates the effects of people moving to more accessible locations, not of people locations becoming more accessible due to commercial development.

That the changes I simulate did not result in changed transport outcomes (positive or negative) does not necessarily imply that no zoning changes would, only that the broad-brush zoning changes I explored likely will not. Zoning changes that are targeted to specific locations around transit may be more successful at changing travel behavior. While I examined several such scenarios, the development they induced was rather modest at a regional scale.

A specific zoning change that I did not investigate due to lack of data is parking. All of my prototype buildings assumed at least parking space per apartment, but having or lacking parking has a strong effect on vehicle ownership, even in randomized trials (Millard-Ball et al. 2020). If zoning codes were to remove parking minimums, perhaps even implement parking maximums, and/or require unbundling of parking from apartment rents, travel behavior changes well beyond what I find here could transpire. A number of cities have already reduced or eliminated parking requirements, including Buffalo, Hartford, San Francisco (Shill 2020, 69), Los Angeles (Manville 2013), Minneapolis (Magrino 2018), and San Diego (NBC7 San Diego 2019). Lehe (2018) and Shoup (2011) argue that minimum parking requirements make housing more expensive and reduce the number of units built. Thigpen (2018) found that residential parking was underutilized and housing density could be increased on the same amount of land. In addition to increasing housing costs, parking requirements represent a subsidy to driving (Shill 2020).

7.3 Future Research

This dissertation is based on microsimulation models. Since significant zoning changes have taken place in a number of locations in the last few years, there will soon be an opportunity to supplement these results with survey-based research. Cities and regions which are relaxing their zoning codes should track the structures that are built under the new rules, to understand their effects. Ideally, residents of the new buildings could be surveyed about their travel behavior, perhaps as an oversample in a regional household travel survey. This would allow researchers to directly compare the travel behavior of residents of the new buildings to the region as a whole. Self-selection of residents into these buildings will still be an issue, so statistical models will have to include carefully-selected controls to try to reduce this issue. In particular, including attitudinal variables in the survey may help reduce bias due to self-selection (Cao, Mokhtarian, and Handy 2009).

The zoning changes examined in this research were limited; I only evaluated a complete elimination of single-family zoning, and elimination of single-family zoning in areas close to transit. Future research should consider more nuanced scenarios. For instance, one scenario that may ease gentrification concerns is to limit development to high-income and/or high-resource areas; since construction is concentrated in these areas anyhow, such a policy may not materially reduce redevelopment potential. Reducing parking requirements could result in significant transport changes, another policy option that should be considered in future research. Additionally, I did not propagate all of the zoning and development scenarios I developed through to the final travel demand model, due to cost and time constraints; future research should bring all scenarios through to the travel demand model.

If the geography of residential development changes significantly, commercial development is likely to change as well. As an area densifies, it is able to support more businesses, meaning residents do not have to travel as far to access the services they require. This could have positive effects on travel behavior that are not modeled herein. Future simulation work should investigate how commercial geography might change in response to a change in residential zoning, to better understand the full effects of a zoning change.

This dissertation focused on a single type of redevelopment—the replacement of individual single-family homes with larger multifamily homes. However, most multifamily construction in recent years has occurred on a much larger scale, and zoning changes may enable larger multifamily construction in existing single-family areas if there are larger parcels, or if developers are able to assemble several parcels. Future research could investigate the potential for larger developments. In addition, even on single lots, evaluating more prototype buildings would be a valuable addition to this dissertation.

This dissertation consists of a string of models—of development, residential location, and travel behavior—strung together. Each model has associated uncertainty. While I have propagated some uncertainty through the model system through the use of multiple scenarios, additional uncertainty is not propagated. In future work it would be valuable to better understand the uncertainty in the final results by accounting for the uncertainty in earlier models. The most promising method is to use a combination of scenario modeling and bootstrapping, although this requires many model runs and would require significant and expensive computer time.

7.4 Recommendations for Policy

The major policy implication of this project is that the transport impacts of blanket upzoning are relatively small. Transport concerns are often a key objection to development and to changes in zoning policy. This research suggests that under a blanket upzoning, these concerns may be unfounded.

This research does not, however, undermine policies that attempt to change transport outcomes through land-use policy (e.g., California's SB 375 Barbour and Deakin 2012). These policies do not propose blanket upzoning, but rather targeted upzoning to focus development in more-accessible places. More research is needed to understand the effects of these policies.

REFERENCES

ActivitySim team. 2020. *ActivitySim: Example.* Accessed February 28. https://activitysim.github.io/activitysim/abmexample.html.

Alonso, William. 1964. *Location and Land Use.* Cambridge, MA: Harvard University Press, January 1.

Anas, Alex. 1982. *Residential Location Markets and Urban Transportation: Economic Theory, Econometrics, and Policy Analysis with Discrete Choice Models.* New York: Academic Press.

Angrist, Joshua D, and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics.* An Empiricist's Companion. Princeton University Press.

Association of Metropolitan Planning Organizations. 2019. *ActivitySim.* Software package. Accessed June 5, 2019. http://activitysim.org.

Aston, Laura, Graham Currie, Alexa Delbosc, Md. Kamruzzaman, and David Teller. 2020. "Exploring Built Environment Impacts on Transit Use – an Updated Meta-Analysis." *Transport Reviews:* 1–24. doi:10.1080/01441647.2020.1806941.

Atlanta Regional Commission. 2019. *ARC Model Documentation.* Accessed March 28, 2021. https://atlregional.github.io/ARC_Model/index.html.

Barbour, Elisa, and Elizabeth A Deakin. 2012. "Smart Growth Planning for Climate Protection: Evaluating California's Senate Bill 375." *Journal of the American Planning Association* 78 (1): 70–86. doi:10.1080/01944363.2011.645272.

Barron, Leonard. 2012. *Investing 101: Estimating Rental Property Expenses,* August. Accessed March 28, 2021. https://www.zillow.com/blog/investing-101-estimating-rental-property-expenses-94824/.

Bayer, Patrick, Robert McMillan, and Kim Rueben. 2004. *An Equilibrium Model of Sorting in an Urban Housing Market* 10865. National Bureau of Economic Research.

Been, Vicki. 2005. "Impact Fees and Housing Affordability." *Cityscape* 8 (1): 139–185.

Been, Vicki, Ingrid Gould Ellen, and Katherine O'Regan. 2019. "Supply Skepticism: Housing Supply and Affordability." *Housing Policy Debate* 29 (1): 25–40. doi:10 . 1080 / 10511482.2018.1476899.

Ben-Akiva, Moshe, and Stephen R Lerman. 1985. *Discrete Choice Analysis: Theory and Application to Travel Demand.* MIT Press.

Bezanson, Jeff, Alan Edelman, Stefan Karpinski, and Viral B. Shah. 2017. "Julia: A Fresh Approach to Numerical Computing." *SIAM Review* 59 (1): 65–98. doi:10.1137/141000671.

Bhat, Chandra R., Rajesh Paleti, Ram M. Pendyala, Karen Lorenzini, and Karthik C. Konduri. 2013. "Accommodating Immigration Status and Self-Selection Effects in a Joint Model of Household Auto Ownership and Residential Location Choice." *Transportation Research Record* 2382 (1): 142–150. doi:10.3141/2382-16.

Biggs, Jeremy. 2019. *factor_analyzer.* Version 0.3.2. https://github.com/EducationalTestingService/factor_analyzer.

Bliss, Laura. 2019. "Where Oregon's Single-Family Zoning Ban Came From." *CityLab* (New York), July 2. Accessed April 8, 2020. https://www.citylab.com/equity/2019/07/oregon-single-family-zoning-reform-yimby-affordable-housing/593137/.

Boyce, David, Biljana Ralevic-Dekic, and Hillel Bar-Gera. 2004. "Convergence of Traffic Assignments: How Much Is Enough?" *Journal of Transportation Engineering* 130, no. 1 (January): 49–55. doi:10.1061/(ASCE)0733-947X(2004)130:1(49).

Boyles, Stephen. 2019. "Beckmann's Formulation, MSA, and Frank-Wolfe." https://sboyles.github.io/teaching/ce392c/5-beckmannmsafw.pdf.

Bromberger, Seth, James Fairbanks, and other contributors. 2017. *JuliaGraphs/LightGraphs.jl: an optimized graphs package for the Julia programming language.* doi:10.5281/zenodo.889971.

Brooks, Leah, and Byron Lutz. 2016. "From Today's City to Tomorrow's City: An Empirical Investigation of Urban Land Assembly." *American Economic Journal: Economic Policy* 8 (3): 69–105. doi:10.1257/pol.20130399.

Bureau of Labor Statistics. 2020. *Consumer Price Index, All Items in West—Size Class A, All Urban Consumers, Not Seasonally Adjusted.* Series CUURS400SA0,CUUSS400SA0.

———. 2021. *Consumer Price Index, All Items in West—Size Class A, All Urban Consumers, Not Seasonally Adjusted.* Series CUUR0000SA0.

California Department of Education. 2012. *STAR Test Results.* Accessed November 23, 2019. https://star.cde.ca.gov/star2012/.

———. 2017. *Enrollment by School: 2016–2017.* Accessed March 6, 2021. https://www.cde.ca.gov/ds/sd/sd/filesenr.asp.

California Department of Housing and Community Development. 2019. *HCD Regional Housing Needs Allocation and Housing Elements.* Accessed December 31. https://www.hcd.ca.gov/community-development/housing-element/index.shtml.

California School Campus Database. 2021. *California School Campus Database.* Both 2018 and 2021 versions were combined. https://californiaschoolcampusdatabase.org/.

Cao, X., P.L. Mokhtarian, and S.L. Handy. 2009. "Examining the Impacts of Residential Self-Selection on Travel Behaviour: A Focus on Empirical Findings." *Transport Reviews* 29 (3): 359–395. doi:10.1080/01441640802539195.

Chazelle, Bernard. 1983. "The Polygon Containment Problem." *Advances in Computing Research* 1:1–33.

Chazelle, Damien, dir. 2016. *La La Land.* Film.

Clarke, Keith C, S Hoppen, and L Gaydos. 1997. "A Self-Modifying Cellular Automaton Model of Historical Urbanization in the San Francisco Bay Area." *Environment and Planning B: Planning and Design* 24:247–261.

Clift, Theresa. 2021. "Sacramento Moves Forward with Controversial Zoning Change Designed to Address Housing Crisis." *The Sacramento Bee* (Sacramento, CA), January 19. Accessed March 27, 2021. https://www.sacbee.com/news/local/article248544635.html.

Committee of the Transport Access Manual. 2020. *Transport Access Manual: A Guide for Measuring Connection between People and Places.* Accessed March 27, 2021. https://hdl.handle.net/2123/23733.

Consumer Financial Protection Bureau. 2019. *Home Mortgage Disclosure Act data, Los Angeles County.* Accessed November 4, 2020. https://ffiec.cfpb.gov/data-browser/data/2019?category=counties&items=06037.

Conway, Matthew Wigginton. 2020. *SQMake.* https://github.com/mattwigway/sqmake.

Conway, Matthew Wigginton, Hunter Owens, and Ryan Gomba. 2018. *Zoning.Space dataset.* Electronic database. https://zoning.space.

Conway, Matthew Wigginton, Deborah Salon, and David A King. 2018. "Trends in Taxi Use and the Advent of Ridehailing, 1995–2017: Evidence from the US National Household Travel Survey." *Urban Science* 2 (3): 79. doi:10.3390/urbansci2030079. http://www.mdpi.com/2413-8851/2/3/79.

Conway, Matthew Wigginton, and Anson F. Stewart. 2019. "Getting Charlie off the MTA: A Multiobjective Optimization Method to Account for Cost Constraints in Public Transit Accessibility Metrics." *International Journal of Geographical Information Science* 33 (9): 1759–1787. doi:10.1080/13658816.2019.1605075.

Delling, Daniel, Thomas Pajor, and Renato F Werneck. 2015. "Round-Based Public Transit Routing." *Transportation Science* 49 (3): 591–604. doi:10.1287/trsc.2014.0534.

De Palma, Andre, Matthew de Lapparent, and Nathalie Picard. 2015. "Modeling Real Estate Investment Decisions in Households." In *Integrated Transport and Land Use Modeling for Sustainable Cities,* edited by Michel Bierlaire, Andre de Palma, Richard Hurtubia, and Paul Waddell, 137–159. Lausanne: EPFL Press.

Dezember, Ryan, and Marco Quiros-Guiterrez. 2021. "New Houses Are Costing More as Prices Jump for Wood, Bricks." *Wall Street Journal: Markets,* March 17. Accessed March 26, 2021. https://www.wsj.com/articles/commodities-boom-hits-home-11615973404.

Dillon, Liam. 2019. "Southern California Must Plan for 1.3 Million New Homes in the next Decade, Newsom Says." *Los Angeles Times* (Los Angeles), August 22. Accessed December 31, 2019. https://www.latimes.com/california/story/2019-08-22/southern-california-housing-growth.

———. 2020. "California Bill to Dramatically Increase Home Building Fails for the Third Year in a Row." *Los Angeles Times: California* (Los Angeles), January 30. Accessed April 8, 2020. https://www.latimes.com/california/story/2020-01-29/high-profile-california-housing-bill-to-allow-mid-rise-apartments-near-transit-falls-short.

Downs, Anthony. 2004. *Still Stuck in Traffic: Coping with Peak-Hour Traffic Congestion.* James A. Johnson Metro Series. Washington, D.C: Brookings Institution Press.

Dumpsters.com. 2021. *Concrete Weight Calculator.* https://www.dumpsters.com/resources/concrete-weight-calculator. Accessed January 10.

Erhardt, Gregory, David Ory, Amar Sarvepalli, Joel Freedman, Jeff Hood, and Ben Stabler. 2012. "MTC's Travel Model One: Applications of an Activity-Based Model in Its First Year." In *Innovations in Travel Modeling.* Accessed March 20, 2021. https://www.researchgate.net/publication/262261101_MTC's_Travel_Model_One_Applications_of_an_Activity-Based_Model_in_its_First_Year.

Ewing, Reid, and Robert Cervero. 2010. "Travel and the Built Environment: A Meta-Analysis." *Journal of the American Planning Association* 76 (3): 265–294. doi:10.1080/01944361003766766.

Federal Emergency Management Agency. 2007. *Public Assistance Debris Management Guide* 325. Washington, D.C.: Federal Emergency Management Agency.

Fischel, William A. 2001. *The Homevoter Hypothesis: How Home Values Influence Local Government Taxation, School Finance, and Land-Use Policies.* Cambridge, Mass: Harvard University Press.

Freddie Mac. 2021. *30-Year Fixed-Rate Mortgage Average in the United States [MORTGAGE30US].* https://fred.stlouisfed.org/series/MORTGAGE30US.

Fulton, William B, and Paul Shigley. 2005. *Guide to California Planning.* Point Arena, Calif.: Solano Press Books.

Galster, George C. 2004. "The Effects of Affordable and Multifamily Housing on Market Values of Nearby Homes." In *Growth Management and Affordable Housing,* 176–211. Washington, DC: Brookings Institution Press.

Gammage, Grady, Jr. 2016. *The Future of the Suburban City: Lessons from Sustaining Phoenix.* Washington, DC: Island Press.

Garcia, David. 2019. "Making It Pencil: The Math Behind Housing Development." Accessed March 28, 2021. https://ternercenter.berkeley.edu/research-and-policy/making-it-pencil/.

GDAL/OGR contributors. 2021. *GDAL/OGR Geospatial Data Abstraction software Library.* Open Source Geospatial Foundation. https://gdal.org.

Giles, Dave. 2013. *Forecasting From Log-Linear Regressions.* R-bloggers, August 22. Accessed March 10, 2021. https://www.r-bloggers.com/2013/08/forecasting-from-log-linear-regressions/.

Gillies, Sean. 2020. *The Shapely User Manual: Shapely 1.7.1 Documentation.* https://shapely.readthedocs.io/en/stable/manual.html.

Glaeser, Edward. 2011. *Triumph of the City.* New York: Penguin.

Guan, Xiaodong, Donggen Wang, and Xinyu (Jason) Cao. 2020. "The Role of Residential Self-Selection in Land Use-Travel Research: A Review of Recent Findings." *Transport Reviews,* no. 3: 267–287. doi:10.1080/01441647.2019.1692965.

Guerra, Erick. 2015. "The Geography of Car Ownership in Mexico City: A Joint Model of Households' Residential Location and Car Ownership Decisions." *Journal of Transport Geography* 43:171–180. doi:10.1016/j.jtrangeo.2015.01.014.

Hanchett, Tom. 2000. "The Other 'Subsidized Housing' Federal Aid To Suburbanization, 1940s-1960s." In *From Tenements to Taylor Homes: In Search of Urban Housing Policy in Twentieth Century America,* edited by John Bauman, Christian Biles, and Kristin Szylvian, 163–179. University Park, PA: Pennsylvania State University Press.

Harris, Charles R., K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, et al. 2020. "Array Programming with NumPy." *Nature* 585, no. 7825 (September): 357–362. doi:10.1038/s41586-020-2649-2.

Hirt, Sonia A. 2014. *Zoned in the USA: The Origins and Implications of American Land-Use Regulation.* Ithaca, NY, USA: Cornell University Press.

HomeAdvisor. 2020. *Learn How Much it Costs to Hire an Architect.* Accessed October 22. https://www.homeadvisor.com/cost/architects-and-engineers/hire-an-architect/.

Increase Housing Opportunities. 2021. In collaboration with Chuck Edwards, Paul Newton, and Milton F. "Toby" Jr Fitch, March 24. Accessed March 27, 2021. https://www.ncleg.gov/Sessions/2021/Bills/Senate/PDF/S349v0.pdf.

Infranca, John. 2019. "The New State Zoning: Land Use Preemption Amid a Housing Crisis." *Boston College Law Review* 60, no. 3 (March 29): 825–887.

Jackson, Mick, dir. 1991. *LA Story.* Film.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning, with Applications in R.* New York: Springer.

Johnson, Jerry, Jeff Frkonja, Maribeth Todd, and Dennis Yee. 2018. "Additional Detail in Aggregate Integrated Land-Use Models via Simulating Developer Pro Forma Thinking." *Journal of Transport and Land Use* 11 (1): 1–14. doi:10.5198/jtlu.2018.1189.

King, David A, Michael J Smart, and Michael Manville. 2019. "The Poverty of the Carless: Toward Universal Auto Access." *Journal of Planning Education and Research* 7 (4). doi:10.1177/0739456X18823252.

Klaiber, H. Allen, and Nicolai V. Kuminoff. 2014. "Equilibrium Sorting Models of Land Use and Residential Choice." In *The Oxford Handbook of Land Economics.* Oxford University Press, July 22. doi:10.1093/oxfordhb/9780199763740.013.011.

Klaiber, H. Allen, and Daniel J. Phaneuf. 2010. "Valuing Open Space in a Residential Sorting Model of the Twin Cities." *Journal of Environmental Economics and Management* 60 (2): 57–77. doi:10.1016/j.jeem.2010.05.002.

Konduri, Karthik C, Daehyun You, Venu M Garikapati, and Ram M Pendyala. 2016. "Enhanced Synthetic Population Generator that Accommodates Control Variables at Multiple Geographic Resolutions." *Transportation Research Record* 2563 (1): 40–50. doi:10.3141/2563-08.

Kuminoff, Nicolai V, V Kerry Smith, and Christopher Timmins. 2013. "The New Economics of Equilibrium Sorting and Policy Evaluation Using Housing Markets." *Journal of Economic Literature* 51, no. 4 (December 1): 1007–1062. doi:10.1257/jel.51.4.1007.

Lam, Siu Kwan, Antoine Pitrou, and Stanley Seibert. 2015. "Numba: A LLVM-Based Python JIT Compiler." In *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC,* 1–6. LLVM '15. New York, NY, USA: Association for Computing Machinery, November. doi:10.1145/2833157.2833162.

Lehe, Lewis. 2018. "How Minimum Parking Requirements Make Housing More Expensive." *Journal of Transport and Land Use* 11, no. 1 (January 5): 1–13. doi:10.5198/jtlu.2018.1340.

Levine, Jonathan. 2006. *Zoned Out: Regulation, Markets, and Choices in Transportation and Metropolitan Land Use.* Washington, DC: Resources for the Future.

Logan, Tim. 2021. "Zoning Measure Would Require Municipalities to Allow Higher-Density Housing near MBTA Stations." *The Boston Globe* (Boston, MA) (January 11). https://www.bostonglobe.com/2021/01/11/business/new-law-would-require-high-density-housing-near-t-stations/.

Luxen, Dennis, and Christian Vetter. 2011. "Real-Time Routing with OpenStreetMap Data." In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS '11,* 513. Chicago, Illinois: ACM Press. doi:10.1145/2093973.2094062.

Magrino, Nick. 2018. "What Happens When You Ease Parking Requirements for New Housing." *nickmagrino.com.* Accessed March 28, 2021. https://www.nickmagrino.com/blog/2018/1/30/when-you-dont-have-to-build-so-much-parking.

Manville, Michael. 2013. "Parking Requirements and Housing Development." *Journal of the American Planning Association* 79 (1): 49–66. doi:10.1080/01944363.2013.785346.

Mapbox. 2018. *Rasterio: Access to Geospatial Raster Data — Rasterio Documentation.* https://rasterio.readthedocs.io/en/latest/.

Mawhorter, Sarah L, David Garcia, and Hayley Raetz. 2018. *It All Adds Up: The Cost of Housing Development Fees in Seven California Cities.* Berkeley, CA: Terner Center for

Housing Innovation. http://ternercenter.berkeley.edu/uploads/Development_Fees_Report_Final_2.pdf.

McCauley, Douglas R. 2019. *Final Regional Housing Need Assessment.* Letter. Accessed March 27, 2021. https://www.hcd.ca.gov/community-development/housing-element/docs/southern_california_association_of_governments_regional_housing_need_determination_for_the_sixth_housing_element_update_1.pdf.

McFadden, Daniel. 1978. "Modeling the Choice of Residential Location." *Transportation Research Record,* no. 673: 72–77.

McKerns, Michael M., Leif Strand, Tim Sullivan, Alta Fang, and Michael A. G. Aivazis. 2012. "Building a Framework for Predictive Science." Accessed November 4, 2019. arXiv: 1202.1056 [cs]. http://arxiv.org/abs/1202.1056.

Mervosh, Sarah. 2018. "Minneapolis, Tackling Housing Crisis and Inequity, Ends Single-Family Zoning." *The New York Times* (December 13).

Miceli, Thomas J, and C F Sirmans. 2007. "The Holdout Problem, Urban Sprawl, and Eminent Domain." *Journal of Housing Economics* 16 (3-4): 309–319. doi:10.1016/j.jhe.2007.06.004.

Microsoft. 2018. *US Building Footprints, version 1.1.* https://github.com/microsoft/USBuildingFootprints.

Miles, Mike E, Laurence M Netherton, and Adrienne Schmitz. 2015. *Real Estate Development: Principles and Process.* Washington, DC: Urban Land Institute.

Millard-Ball, Adam, Jeremy West, Nazanin Rezaei, and Garima Desai. 2020. "How the Built Environment Affects Car Ownership and Travel: Evidence from San Francisco Housing Lotteries." Accessed March 28, 2021. doi:10.7922/G2319T55. https://escholarship.org/uc/item/7tw5x9p7.

Missouri Census Data Center. 2014. *Geocorr 2014.* Accessed December 31, 2019. http://mcdc.missouri.edu/applications/geocorr2014.html.

Molloy, Raven. 2018. "The Effect of Housing Supply Regulation on Housing Affordability: A Review." *Regional Science and Urban Economics.* doi:10.1016/j.regsciurbeco.2018.03.007.

Monkkonen, Paavo, Ian Carlton, and Kate Macfarlane. 2020. "One to Four: The Market Potential of Fourplexes in California's Single-Family Neighborhoods" (July).

Monkkonen, Paavo, Michael Lens, and Michael Manville. 2020. *Built-out Cities? How California Restricts Housing Production through Prohibition and Process.* Berkeley, CA: Terner Center for Housing Innovation. Accessed January 16, 2021. http://californialanduse.org/download/MLM%20Built-Out%20Cities%202020.pdf.

Murphy, Alvin. 2018. "A Dynamic Model of Housing Supply." *American Economic Journal: Economic Policy* 10 (4): 243–267. doi:10.1257/pol.20150297.

National Center for Education Statistics. 2017. *Integrated Postsecondary Education Data System.* https://nces.ed.gov/ipeds/use-the-data.

NBC7 San Diego. 2019. *Some Future Housing Developments Won't Need Parking Spaces.* NBC 7 San Diego. Accessed May 28, 2019. http://www.nbcsandiego.com/news/local/City-Council-to-Consider-Eliminating-Parking-Requirements-for-Some-Housing-Developments-506664061.html.

Nocedal, Jorge, and Stephen J. Wright. 2006. *Numerical Optimization.* 2nd ed. Springer series in operations research. New York: Springer.

O'Flaherty, Brendan. 2005. *City Economics.* Cambridge, MA: Harvard University Press.

OpenStreetMap contributors. 2020. *OpenStreetMap.* https://openstreetmap.org.

Ortúzar, Juan de Dios, and Luis G Willumsen. 2011. *Modelling Transport.* 4th ed. Chichester, UK: Wiley.

Ory, David. 2016. *MasterNetworkLookupTables.* MTC/ABAG Analytical Modeling Wiki. Accessed March 20, 2021. http://analytics.mtc.ca.gov/foswiki/bin/view/Main/MasterNetworkLookupTables.

Ory, David, Flavia Tsang, and Lisa Zorn. 2019. *HwySkims.Job.* GitHub. Accessed March 20, 2021. https://github.com/BayAreaMetro/travel-model-one.

Owen, Andrew, and David M Levinson. 2015. "Modeling the Commute Mode Share of Transit Using Continuous Accessibility to Jobs." *Transportation Research Part A* 74 (C): 110–122. doi:10.1016/j.tra.2015.02.002.

Owen, Andrew, and Brendan Murphy. 2018. *Access Across America: Transit 2017 Data.* doi:10.13020/D6G68V.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al. 2011. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research* 12:2825–2830.

Pendall, Rolf. 1999. "Opposition to Housing: NIMBY and Beyond." *Urban Affairs Review* 35 (1): 112–136. doi:10.1177/10780879922184310.

Phillips, Shane. 2020. *The Affordable City: Strategies for Putting Housing within Reach (and Keeping It There).* Washington: Island Press.

Pollack, Stephanie, Barry Bluestone, and Chase Billingham. 2010. *Maintaining Diversity In America's Transit-Rich Neighborhoods:* Boston: Dukakis Center for Urban and Regional Policy.

Pucher, John, and John L Renne. 2003. "Socioeconomics of Urban Travel: Evidence from the 2001 NHTS." *Transportation Quarterly* 57 (3): 49–77.

Pyle, Lizbeth A. 1985. "The Land Market beyond the Urban Fringe." *Geographical Review* 75 (1): 32–43. doi:10.2307/214576.

Python Wiki. 2020. *Global Interpreter Lock.* Accessed January 12, 2021. https://wiki.python.org/moin/GlobalInterpreterLock.

QGIS Project. 2021. *QGIS Desktop 3.16 User Guide.* Version 3.16.2. https://docs.qgis.org/3.16/pdf/en/QGIS-3.16-DesktopUserGuide-en.pdf.

Rambha, Tarun. 2020. *CE 272 Traffic Network Equilibrium.* Bengaluru: Indian Institute of Science and Technology. http://civil.iisc.ernet.in/~tarunr/CE_272/Lecture_9.pdf.

Revels, J., M. Lubin, and T. Papamarkou. 2016. "Forward-Mode Automatic Differentiation in Julia." *arXiv:1607.07892 [cs.MS].* https://arxiv.org/abs/1607.07892.

Rey, Sergio J., and Luc Anselin. 2007. "PySAL: A Python Library of Spatial Analytical Methods." *The Review of Regional Studies* 37 (1): 5–27.

Rodríguez-Pose, Andrés, and Michael Storper. 2019. "Housing, Urban Growth and Inequalities: The Limits to Deregulation and Upzoning in Reducing Economic and Spatial Inequality." *Urban Studies:* 44. http://eprints.lse.ac.uk/100738/.

Rossi, Peter H. 1980. *Why Families Move.* 2nd ed. Beverly Hills: Sage.

Rothstein, Richard. 2017. *The Color of Law: A Forgotten History of How Our Government Segregated America.* New York: Liveright, January 1.

RSMeans. 2020. *Square Foot Estimator.* https://rsmeansonline.com.

Ruggles, Steven, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas, and Matthew Sobek. 2019. *IPUMS USA: Version 9.0 [Dataset].* Minneapolis: IPUMS. https://doi.org/10.18128/D010.V9.0.

Salon, Deborah. 2009. "Neighborhoods, Cars, and Commuting in New York City: A Discrete Choice Approach." *Transportation Research Part A: Policy and Practice* 43, no. 2 (February 1): 180–196. doi:10.1016/j.tra.2008.10.002.

SciPy Developers. 2019. *Scipy.* https://scipy.org/scipylib/index.html.

Seabold, Skipper, and Josef Perktold. 2010. "Statsmodels: Econometric and Statistical Modeling with Python." Austin, TX. http://conference.scipy.org/proceedings/scipy2010/pdfs/seabold.pdf.

Shannon, Megan Elizabeth. 2015. "Quantifying the Impacts of Regulatory Delay on Housing Affordability and Quality in Austin, Texas." Master's thesis. University of Texas at Austin. http://hdl.handle.net/2152/32194.

Shaw, Randy. 2018. *Generation Priced Out: Who Gets to Live in the New Urban America.* Oakland, CA, USA: University of California Press.

Shill, Gregory H. 2020. "Should Law Subsidize Driving?" *New York University Law Review* 95:498. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3345366.

Shoup, Donald. 2011. *The High Cost of Free Parking.* Chicago: American Planning Association Planners Press.

Solon, Gary, Steven J Haider, and Jeffrey M Wooldridge. 2015. "What Are We Weighting For?" *Journal of Human Resources* 50, no. 2 (March 31): 301–316. doi:10.3368/jhr.50.2.301.

Southern California Association of Goverments. 2019. *High Quality Transit Areas (HQTA) 2045 – SCAG Region.* https://gisdata-scag.opendata.arcgis.com/datasets/43e6fef395d041c09deaeb369a513ca1_1.

Southern California Association of Governments. 2012. *SCAG Regional Travel Demand Model and 2012 Model Validation.* Accessed March 25, 2021. https://scag.ca.gov/sites/main/files/file-attachments/scag_rtdm_2012modelvalidation.pdf?1605571641.

———. 2018. *Los Angeles Combined Land Use.* Accessed February 10, 2020. http://gisdata-scag.opendata.arcgis.com/datasets/389cd8eadde94fe9b1cc20bb8f3a6d1a_0.

———. 2020. *6th Cycle Draft RHNA Allocations Based on Final RHNA Methodology and Final Connect SoCal,* September 3. Accessed January 29, 2021. https://scag.ca.

gov/sites/main/files/file-attachments/rhna-draft-allocations-090320-updated.pdf?1602188695.

Stiphany, Kristine, and Jake Wegmann. 2020. "From Owner-Occupant to Senhorio: The Micreconomics of Rentrification in São Paulo." Paper presented at the 2020 Association of Collegiate Schools of Planning conference, online.

The Toll Roads of Orange County. 2021. *Rate Card: Two-Axle Vehicles and Motorcycles.* https://thetollroads.com/sites/default/files/FY21_RateCard.pdf.

Thigpen, Calvin G. 2018. "Giving Parking the Time of Day: A Case Study of a Novel Parking Occupancy Measure and an Evaluation of Infill Development and Carsharing as Solutions to Parking Oversupply." *Research in Transportation Business & Management* 29:108–115. doi:10.1016/j.rtbm.2018.03.003.

Tra, Constant I. 2007. "Evaluating the Equilibrium Welfare Impacts of the 1990 Clean Air Act Amendments in the Los Angeles Area." Doctoral dissertation, University of Maryland, College Park. http://hdl.handle.net/1903/7236.

———. 2010. "A Discrete Choice Equilibrium Approach to Valuing Large Environmental Changes." *Journal of Public Economics* 94 (1): 183–196. doi:10.1016/j.jpubeco.2009.10.006.

———. 2013. "Measuring the General Equilibrium Benefits of Air Quality Regulation in Small Urban Areas." *Land Economics* 89 (2): 291–307. Accessed October 10, 2019. http://muse.jhu.edu/article/502176.

Transportation Secure Data Center. 2019. *2017 National Household Travel Survey: California add-on: cleansed data.* National Renewable Energy Laboratory. https://www.nrel.gov/transportation/secure-transportation-data/tsdc-nhts-california.html.

Uber. 2020. *Uber Movement: Los Angeles 2019 Q4.* Accessed February 17, 2021.

US Census Bureau. 2017. *Longitudinal Origin-Destination Employment Statistics.* http://lehd.ces.census.gov/lodes/.

———. 2018. *US Census Bureau Building Permits Survey.* Accessed December 31, 2019. https://www.census.gov/construction/bps/.

———. 2019. *Tenure by Units in Structure, Los Angeles-Long Beach CSA, 2019 American Community Survey 1-year estimates.* Accessed January 12, 2020. https://censusreporter.org/data/table/?table=B25032&geo_ids=33000US348&primary_geo_id=33000US348#valueType%7Cestimate.

US Census Bureau. 2021. *2019 American Community Survey 1-year Estimates: Tables B08006 and B18120.* Accessed March 27. http://censusreporter.org/profiles/33000US348-los-angeles-long-beach-ca-csa/.

US EPA. 2013. *Smart Location Database, Version 2.0.* Data and Tools. Accessed July 12, 2019. https://www.epa.gov/smartgrowth/smart-location-mapping.

US Geological Survey. 2021. *3D Elevation Program 1/3 Arc-Second DEMs.* Accessed March 20. https://www.usgs.gov/core-science-systems/ngp/3dep.

ValuePenguin. 2020. *Average Commercial Real Estate Loan Rates for 2020.* Accessed October 20, 2020. https://www.valuepenguin.com/average-commercial-real-estate-loan-rates.

Van Duijn, M., and J. Rouwendal. 2013. "Cultural Heritage and the Location Choice of Dutch Households in a Residential Sorting Model." *Journal of Economic Geography* 13 (3): 473–500. doi:10.1093/jeg/lbs028.

Waddell, Paul. 2000. "A Behavioral Simulation Model for Metropolitan Policy Analysis and Planning: Residential Location and Housing Market Components of Urbansim." *Environment and Planning B: Planning and Design* 27, no. 2 (April 1): 247–263. doi:10.1068/b2627.

———. 2010. "Modeling Residential Location in UrbanSim." In *Residential Location Choice,* edited by Francesca Pagliara, John Preston, and David Simmonds, 165–180. Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-12788-5_8.

———. 2013. *Draft Technical Documentation: San Francisco Bay Area UrbanSim Application.* Accessed May 28, 2019. http://analytics.mtc.ca.gov/foswiki/pub/Main/Documents/2013_04_01_DRAFT_RELEASE_UrbanSim_Documentation.pdf.

Waddell, Paul, Geoff Boeing, Max Gardner, and Emily Porter. 2018. "An Integrated Pipeline Architecture for Modeling Urban Land Use, Travel Demand, and Traffic Assignment." *arXiv:1802.09335 [cs]* (February 7). Accessed May 28, 2019. arXiv: 1802.09335. http://arxiv.org/abs/1802.09335.

Waddell, Paul, A Borning, M Noth, N Freier, M Becke, and G Ulfarsson. 2003. "Microsimulation of Urban Development and Location Choices: Design and Implementation of UrbanSim." *Networks and Spatial Economics* 3 (1): 43–67. doi:10.1023/A:1022049000877.

Waddell, Paul, Liming Wang, Billy Charlton, and Aksel Olsen. 2010. "Microsimulating Parcel-Level Land Use and Activity-Based Travel: Development of a Prototype Application in San Francisco." *Journal of Transport and Land Use* 3 (2): 1–20. doi:10.5198/jtlu.v3i2.124.

Waste Management. 2020. *Simi Valley Landfill and Recycling Center Service and Fee Schedule.* Accessed January 10, 2021. https://web.archive.org/web/20201020040440/https://www.wm.com/location/california/ventura-county/_documents/PriceSheet.pdf.

White, Roger, Guy Engelen, and Inge Uljee. 2015. *Modeling Cities and Regions as Complex Systems: From Theory to Planning Applications.* Cambridge, MA: MIT Press.

Wiener, Scott. 2017. *SB-35 Planning and Zoning: Affordable Housing: Streamlined Approval Process.* Accessed October 22, 2020. https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180SB35.

Wiener, Scott, Anna Caballero, Ben Hueso, John Moorlach, Nancy Skinner, Jeff Stone, Autumn Burke, et al. 2019. *California SB 50: Planning and Zoning: Housing Development: Incentives.,* January 5. Accessed June 1, 2019. https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=201920200SB50.

Winter, Stephan. 2002. "Modeling Costs of Turns in Route Planning." *GeoInformatica* 6 (4): 345–361. doi:10.1023/A:1020853410145.

Zillow. 2019a. *How Much Does It Cost to Sell a House?* Accessed March 28, 2021. https://www.zillow.com/sellers-guide/costs-to-sell-a-house/.

———. 2019b. *Zillow Transacton and Assessment Database (ZTRAX).* Electronic database. Accessed May 29. https://www.zillow.com/research/ztrax/.

———. 2020. *Zillow Observed Rental Index: Los Angeles County, September 2020.* Accessed November 4, 2020. https://www.zillow.com/research/methodology-zori-repeat-rent-27092/.

———. 2021. *Zillow Home Value Index, Single Family Homes Time Series.* Accessed February 5, 2021. https://www.zillow.com/research/data/.

APPENDIX A

MAPS OF SENSITIVITY TESTS FOR PROFITABILITY MODEL

These maps show where it is profitable to redevelop single-family homes in all scenarios tested in Chapter 2.

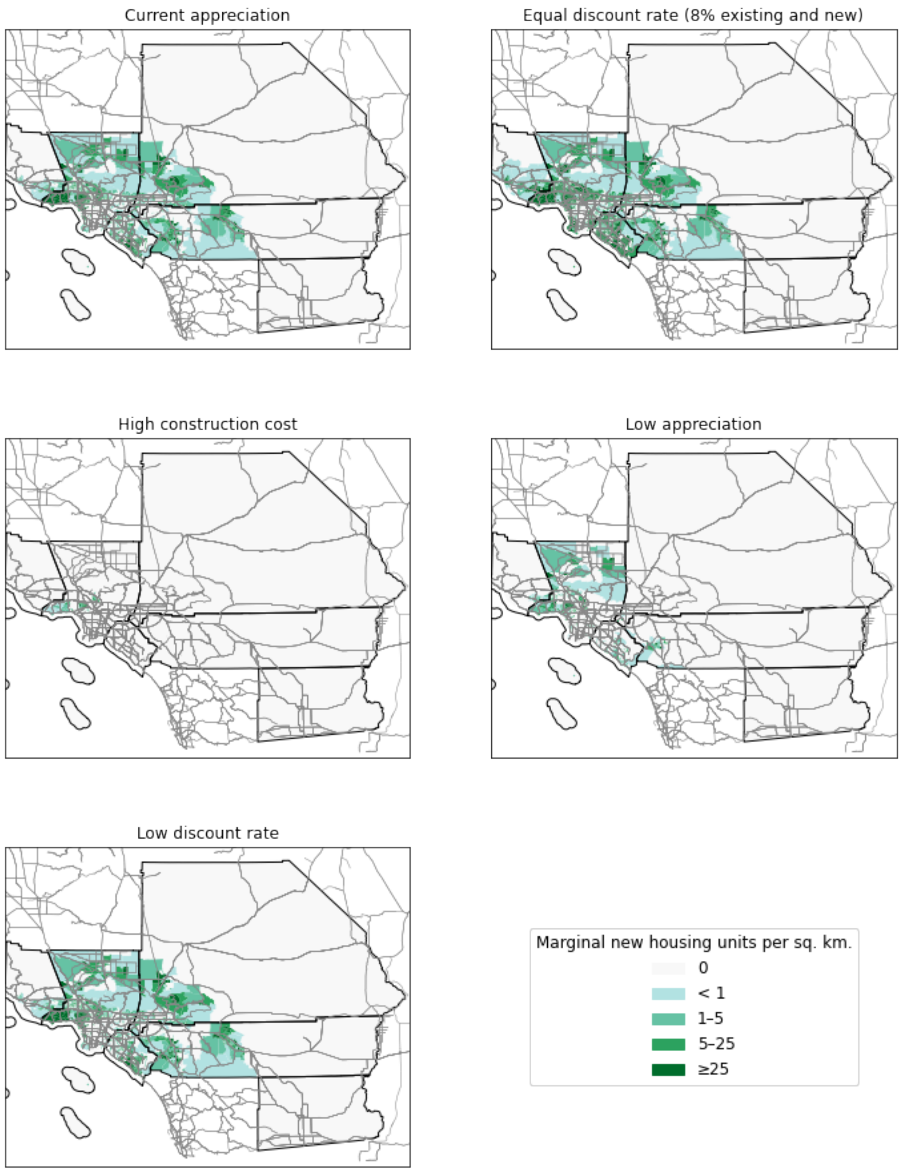


Figure 40. Geographic Distribution of Growth for Each Scenario

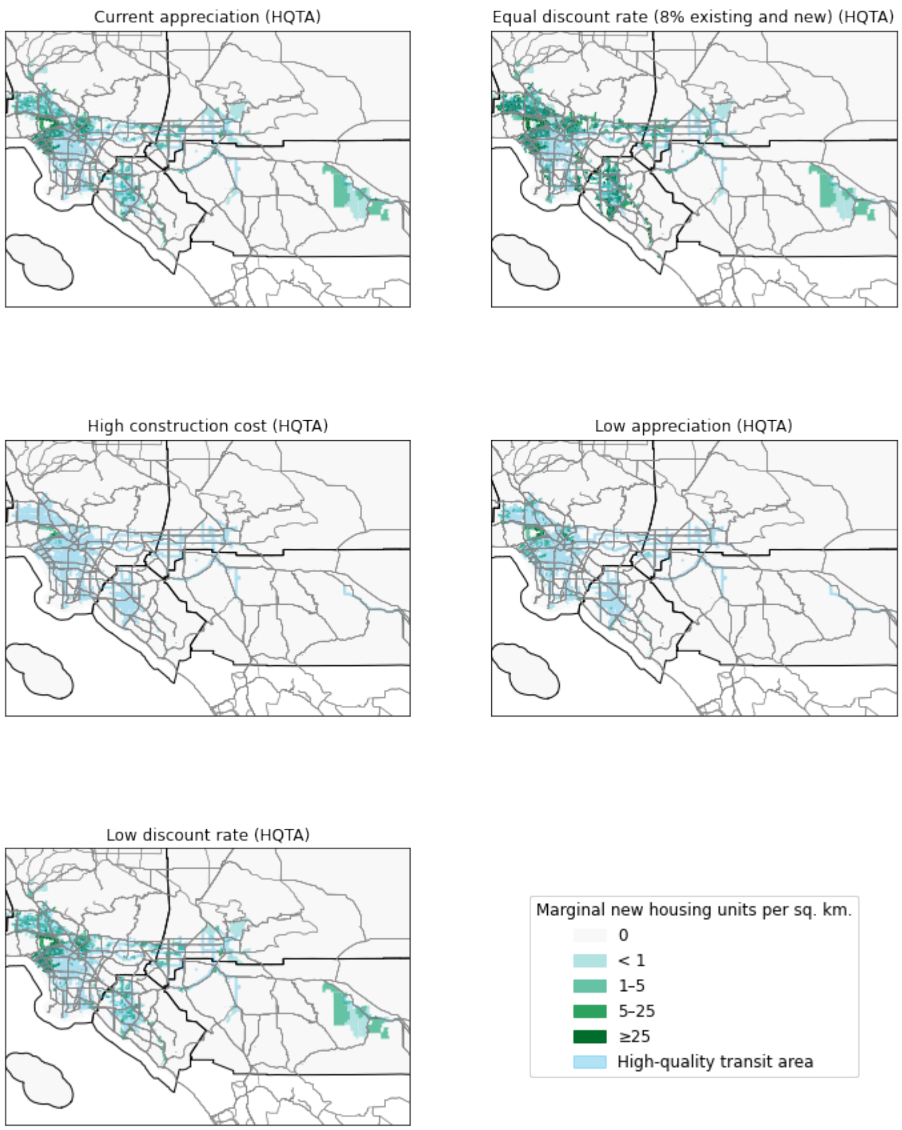


Figure 41. Geographic Distribution of Growth for Each Scenario

APPENDIX B

COMPUTATIONAL APPROACHES USED IN SORTING MODEL

The computational requirements of estimating and equilibrating the sorting model described in Chapter 4 are significant, and a number of techniques were needed in order to successfully run the model. Those techniques are detailed in the sections that follow. For context, the model has 260,933 households, 974 housing types, 4 levels of vehicle ownership, and 105 estimated parameters exclusive of alternative specific constants. Both estimation and simulation take place on an Amazon Web Services `r5.4xlarge` instance with 16 cores, 128 GB of RAM, and a 300 GB solid-state disk, running Amazon Linux, or on Intel x86 or AMD Epyc nodes with 100GB of allocated RAM in the ASU high-performance compute cluster

B.1 Computing Utility Piecewise

The utility of all alternatives for all households is needed to compute the full alternative specific constants to guarantee that the observed prices clear the existing market, and in market clearing during sorting. Conceptually, this is simply the product of the matrix of all variables for all household-alternative pairs multiplied by the vector of coefficients. However, with 105 variables (excluding ASCs), 974 housing types, 4 levels of vehicle ownership, and 260,933 households, the matrix of variables contains $105 \times 974 \times 4 \times 260,933 \approxeq 107$ billion floating point numbers. Since each double-precision floating point number requires 8 bytes of memory, this array would require 795 GB of memory—far more than the 128 GB available on the machine used to fit the model. However, the final vector of utilities is has only $974 \times 4 \times 260,933$ elements, and thus requires only 8 GB of memory. Since the utility of each row is independent, I compute the utility in chunks, allocating intermediate matrices of 2 GB each, multiplying by coefficients, and then combining the results into a single final array. With multithreading (see Section B.3), this computation takes about 30 minutes.

B.2 Taking Advantage of the Linear Utility Function

There are several places where derivatives of the utility function need to be numerically estimated with respect to some variable. Most notably, this is needed when computing the market-clearing prices in Equation (4.8). It is also needed when computing the Hessian matrix to produce standard errors for the coefficient estimates in Appendix C. When the derivatives are used in sorting, this is especially computationally intensive, as derivatives need to be calculated based on all alternatives, rather than the sampled alternatives used in model fitting—indeed, the vast majority of the compute time to find market-clearing prices is used in finding derivatives. To make this process faster, rather than calculating the full utility function with one parameter slightly modified, I first calculate the full utility function. Then, for each parameter or housing price I need to calculate the derivative with respect to, I add only the difference in the utility function implied by that change, rather than recalculating the full utility function. Similarly, in some cases, I calculate the full utility function except for a particular variable, and then add different versions of that variable to calculate the derivative. This drastically speeds the process of computing derivatives.

B.3 Multithreading

To speed computation, multithreading is used in several locations throughout the model fitting process. Piecewise utility computation and Hessian computation are both implemented using as many threads as the machine has available CPUs, while derivative computations in the sorting phase use 2 threads to avoid overloading available memory. These are implemented using Python's `threading` library. The *global interpreter lock* in Python prevents Python code from executing in more than one thread simultaneously, usually making multithreading impractical in Python programs. However, most of the computation in these processes is done with `numpy`, which is an extension to Python written in C which releases this global interpreter lock (Python Wiki 2020), allowing the heavy compuations to happen in parallel, even though the lightweight Python code that starts them and stores their results does not.

Multithreading is also heavily used in traffic assignment; Julia does not have a global interpreter lock, making implementation simpler.

B.4 Memory-Mapped Arrays

Memory-mapping allows a file on disk to be addressed as if it were in memory, and the operating system kernel handles moving pages of the file in and out of memory as needed, and caches the file in memory whenever available memory allows. I memory-map several of the large but infrequently-used arrays, such as indices into the vector of utilities. These indices are generally accessed sequentially, so once an access starts the CPU is able to easily predict which parts of the file will be read next and have that part of the file ready in memory when the program needs it, making accessing these memory-mapped files stored on disk almost as fast as if they were stored directly in memory, but without the memory hit.

Memory-mapped arrays can also be opened in *copy-on-write* mode, in which modifications to the array are stored in memory but not persisted back to disk. This is used when calculating derivatives of demand with respect to price. A small part of the utility array must be changed in each thread that is computing derivatives. To prevent each thread from having to have its own copy of this large array to modify, I write the array to a file and then memory-map it in copy-on-write mode into each thread, so the threads can modify their own private copies of the array, while still sharing the unmodified portions to save memory.

APPENDIX C

SORTING MODEL COEFFICIENTS

Table 17. Equilibrium Sorting Model Results

	Coefficient	Std. Err.	p-value
First-stage tenure choice model†			
Binary logit for renting vs. owning			
University-educated in HH	-0.75***	0.08	0.00
Number of workers	0.29***	0.01	0.00
Child in HH	-0.11	0.09	0.22
Senior in HH	-0.68***	0.07	0.00
Household size	0.25***	0.01	0.00
Income \$50,000–\$100,000	-2.25***	0.08	0.00
Income ≥ \$100,000	-3.32***	0.10	0.00
Income < \$50,000 × immigrant in HH	-0.14***	0.03	0.00
Income \$50,000–\$100,000 × immigrant in HH	-0.25***	0.03	0.00
Income ≥ \$100,000 × immigrant in HH	-0.36***	0.03	0.00
First-stage residential choice model†			
	Coefficient	Std. Err.	p-value
Budget (ln(income - housing cost))	0.46***	0.01	0.00
Renters			
College-educated individual in household ×			
Median fifth-grade math proficiency	3.42***	0.09	0.00
PUMA access	0.15***	0.02	0.00
Regional access	0.07***	0.01	0.00
PUMA access > 90th pctile	0.38***	0.04	0.00
Regional access > 90th pctile	-0.02	0.03	0.57
Single-family home	-0.06***	0.02	0.00
Number of workers in household ×			
PUMA access	0.08***	0.01	0.00
Regional access	0.14***	0.01	0.00
PUMA access > 90th pctile	-0.08***	0.02	0.00
Regional access > 90th pctile	-0.0	0.02	0.78
Immigrant in household ×			
Single-family home	-0.29***	0.02	0.00
PUMA access	0.26***	0.01	0.00
Regional access	0.51***	0.01	0.00
PUMA access > 90th pctile	-0.31***	0.04	0.00
Regional access > 90th pctile	-0.43***	0.03	0.00
Child in household ×			
Single-family home	0.02	0.02	0.32
PUMA access	-0.18***	0.02	0.00
Regional access	-0.06***	0.01	0.00

Table 17. Equilibrium Sorting Model Results, Continued

	Coefficient	Std. Err.	p-value
PUMA access > 90th pctile	-0.13**	0.04	0.00
Regional access > 90th pctile	-0.11***	0.03	0.00
Senior in household ×			
Median fifth-grade math proficiency	-0.18*	0.07	0.01
Single-family home	0.45***	0.02	0.00
PUMA access	-0.12***	0.01	0.00
Regional access	0.02	0.01	0.12
PUMA access > 90th pctile	-0.06	0.04	0.14
Regional access > 90th pctile	-0.11***	0.03	0.00
Household size ×			
Single-family home	0.38***	0.01	0.00
Income \$50,000–\$100,000 ×			
Median fifth-grade math proficiency	0.91***	0.09	0.00
Single-family home	0.18***	0.02	0.00
PUMA access	-0.01	0.02	0.44
Regional access	0.07***	0.01	0.00
PUMA access > 90th pctile	0.1*	0.04	0.02
Regional access > 90th pctile	-0.12***	0.03	0.00
Income ≥ \$100,000 ×			
Median fifth-grade math proficiency	3.07***	0.12	0.00
Single-family home	0.63***	0.02	0.00
PUMA access	0.11***	0.02	0.00
Regional access	0.08***	0.02	0.00
PUMA access > 90th pctile	0.19***	0.06	0.00
Regional access > 90th pctile	-0.0	0.04	1.00
Child in household ×			
Median fifth-grade math proficiency ×			
Income < \$50,000 × No college-educated individual in household	-2.43***	0.10	0.00
Income \$50,000–\$100,000 × No college-educated individual in household	-2.11***	0.12	0.00
Income ≥ \$100,000 × No college-educated individual in household	-1.41***	0.16	0.00
Income < \$50,000 × College-educated individual in household	-1.14***	0.21	0.00
Income \$50,000–\$100,000 × College-educated individual in household	-0.71***	0.15	0.00
Income ≥ \$100,000 × College-educated individual in household	0.55***	0.13	0.00

Table 17. Equilibrium Sorting Model Results, Continued

	Coefficient	Std. Err.	p-value
Owners			
College-educated individual in household ×			
Median fifth-grade math proficiency	2.99***	0.08	0.00
PUMA access	0.17***	0.02	0.00
Regional access	0.09***	0.01	0.00
PUMA access > 90th pctile	0.19***	0.04	0.00
Regional access > 90th pctile	-0.09**	0.03	0.00
Single-family home	-0.29***	0.02	0.00
Number of workers in household ×			
PUMA access	0.05***	0.01	0.00
Regional access	0.11***	0.01	0.00
PUMA access > 90th pctile	-0.1***	0.02	0.00
Regional access > 90th pctile	0.07***	0.01	0.00
Immigrant in household ×			
Single-family home	-0.3***	0.02	0.00
PUMA access	0.24***	0.02	0.00
Regional access	0.43***	0.01	0.00
PUMA access > 90th pctile	-0.42***	0.04	0.00
Regional access > 90th pctile	-0.5***	0.03	0.00
Child in household ×			
Single-family home	-0.03	0.04	0.49
PUMA access	-0.03	0.02	0.08
Regional access	-0.06***	0.01	0.00
PUMA access > 90th pctile	-0.08	0.05	0.07
Regional access > 90th pctile	0.13***	0.03	0.00
Senior in household ×			
Median fifth-grade math proficiency	0.24***	0.06	0.00
Single-family home	0.57***	0.02	0.00
PUMA access	-0.0	0.02	0.85
Regional access	0.11***	0.01	0.00
PUMA access > 90th pctile	-0.07	0.04	0.06
Regional access > 90th pctile	-0.1***	0.03	0.00
Household size ×			
Single-family home	0.6***	0.01	0.00
Income \$50,000–\$100,000 ×			
Median fifth-grade math proficiency	-1.33***	0.09	0.00
Single-family home	-0.11***	0.03	0.00
PUMA access	-0.01	0.02	0.70

Table 17. Equilibrium Sorting Model Results, Continued

	Coefficient	Std. Err.	p-value
Regional access	-0.06***	0.01	0.00
PUMA access > 90th pctile	-0.12*	0.05	0.02
Regional access > 90th pctile	0.02	0.04	0.55
Income ≥ $100,000 ×			
Median fifth-grade math proficiency	0.38***	0.09	0.00
Single-family home	0.23***	0.03	0.00
PUMA access	0.06*	0.02	0.01
Regional access	-0.04**	0.01	0.00
PUMA access > 90th pctile	0.17**	0.06	0.00
Regional access > 90th pctile	-0.07	0.04	0.10
Child in household ×			
Median fifth-grade math proficiency ×			
Income < $50,000 × No college-educated individual in household	-3.58***	0.13	0.00
Income $50,000–$100,000 × No college-educated individual in household	-2.61***	0.12	0.00
Income ≥ $100,000 × No college-educated individual in household	-1.56***	0.15	0.00
Income < $50,000 × College-educated individual in household	-1.49***	0.22	0.00
Income $50,000–$100,000 × College-educated individual in household	-0.81***	0.15	0.00
Income ≥ $100,000 × College-educated individual in household	1.08***	0.11	0.00

First-stage vehicle ownership model†

	One car (Std. Err.)	Two cars (Std. Err.)	Three or more cars (Std. Err.)
ASC‡	2.25	-0.06	-2.91
Child in household	0.96*** (0.03)	0.14*** (0.03)	-0.9*** (0.03)
PUMA access	-0.22*** (0.01)	-0.43*** (0.01)	-0.68*** (0.02)
Regional access	-0.1*** (0.01)	-0.24*** (0.01)	-0.34*** (0.01)
PUMA access > 90th pctile	-0.01 (0.03)	-0.09* (0.04)	-0.13** (0.04)

Table 17. Equilibrium Sorting Model Results, Continued

	Coefficient	Std. Err.	p-value
Regional access > 90th pctile	-0.19***	-0.38***	-0.49***
	(0.02)	(0.03)	(0.03)
Income ≥ $100,000	0.38***	1.69***	2.45***
	(0.03)	(0.03)	(0.03)
Income $50,000–$100,000	0.63***	1.26***	1.62***
	(0.02)	(0.02)	(0.03)
Household size	-0.46***	0.3***	0.8***
	(0.01)	(0.01)	(0.01)
College-educated individual in household	0.59***	0.81***	0.67***
	(0.02)	(0.02)	(0.02)
Number of workers in household	0.07***	0.4***	0.83***
	(0.01)	(0.01)	(0.02)
First-stage model diagnostics			
McFadden's Pseudo-R^2	0.14		
Number of households	262,924		

Second-stage housing choice model (OLS)			
	Coefficient	Std. Err.§	p-value
Intercept (ν)	0.75***	0.06	0.0
Rent	0.57***	0.09	0.0
Rent ×			
Single-family home	-1.27***	0.02	0.0
PUMA access	0.06***	0.02	0.0
Regional access	-0.21***	0.01	0.0
PUMA access > 90th pctile	0.15**	0.05	0.0
Regional access > 90th pctile	0.48***	0.03	0.0
Median fifth-grade math proficiency	-1.9***	0.09	0.0
Own ×			
Single-family home	-1.58***	0.02	0.0
PUMA access	0.09***	0.02	0.0
Regional access	-0.11***	0.01	0.0
PUMA access > 90th pctile	0.27***	0.05	0.0
Regional access > 90th pctile	0.46***	0.03	0.0
Median fifth-grade math proficiency	-1.71***	0.09	0.0
R^2	0.94		
Adj. R^2	0.94		
Sample size	974		

.: $p < 0.1$; *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$

† Jointly estimated.

Table 17. Equilibrium Sorting Model Results, Continued

	Coefficient	Std. Err.	p-value

‡ ASCs are found through a contraction mapping process, standard errors are not estimated.
§ These standard errors do not account for any error in the dependent variable from the first-stage estimation.

APPENDIX D

OPEN-SOURCE SOFTWARE DEVELOPED FOR THIS DISSERTATION

In order to support my work on this project, I developed several pieces of software that may be generally useful to the research community. I have released them as open-source under the Apache license to promote their re-use.

D.1 `eqsormo`

https://github.com/mattwigway/eqsormo

`eqsormo` is a Python package for fitting equilibrium sorting models. It currently supports the Tra-style sorting models used in my dissertation. Its modular design means it could be adapted to support other types of sorting models.

D.2 `TransitRouter.jl`

https://github.com/TransitRouter.jl

`TransitRouter.jl` is a Julia package to find shortest paths in public transport networks, using the RAPTOR algorithm (Delling, Pajor, and Werneck 2015) and machine-readable schedule data in GTFS format. It also performs routing on the street network for access to transit through an integration with the OSRM route-finding software (Luxen and Vetter 2011). This package was used to calculate transit skims for the activity-based model

D.3 `OSMPBF.jl`

https://github.com/mattwigway/OSMPBF.jl

`OSMPBF.jl` is a Julia package for reading OpenStreetMap from the binary OpenStreetMap Protocol Buffers format, including files that are larger than available RAM. This format can be processed far more efficiently than the original XML-based representation of OpenStreetMap data, but no package for reading these files was available for Julia. This package was used to build the road network for the Frank-Wolfe traffic assignment process.

D.4 `sqmake`

https://github.com/mattwigway/sqmake

`sqmake` is GNU Make-like tool for managing data analysis that takes place in SQL databases. Much of the construction model in this dissertation is implemented as a series of long-runing queries in PostGIS database. `sqmake` allows me to write all of these scripts, and then simply type `sqm <task>` to run `<task>` as well as all the other tasks that must be run in preparation for that task. `sqmake` figures out which tasks have already been run and do not need to be re-run based on the state of the database.