

ABSTRACT

AMINDARBRI, REZA. Spatially Disaggregated Simulation of Interconnections between Land Use Policy, Housing Markets, and the Affordability Crisis. (Under the direction of Dr. Perver Baran and Dr. Ross Meentemeyer).

Access to affordable housing is a major determinant of one's quality of life, yet many households in the United States continue to be burdened by the cost of housing. Land use regulations play a key role on both the supply and demand sides of the real estate market by regulating the location and density of employment, as well as housing. Given their impact on the real estate market, land use regulations must be an integral consideration when forming policies that seek to mitigate the housing affordability crisis. Developing geospatial models for this interaction between land use and home price on a spatially disaggregated level enables decisionmakers to evaluate the impact of their land use policies and plans from the perspective of housing affordability. However, existing standalone residential real estate pricing models are insensitive to changes in land use. Moreover, the data preparation, calibration, and training for integrated land use and transportation models are nontrivial and may be impractical for many municipalities and planning agencies. This dissertation presents a simple-to-implement framework, SimP-R, for simulating changes in housing prices on a spatially disaggregated level in response to land use change. It is composed of a residential real estate pricing model and an algorithm for computing a novel measure of supply-to-demand ratio. SimP-R's pricing model predicts home prices based on a series of building and location-related attributes, including the proposed disaggregated gravity-based measure of supply-to-demand ratio. This dissertation also demonstrates the implementation of SimP-R in the city of San Francisco, with the entire Bay Area serving as the influence geography. The results demonstrate this framework's ability to simulate the effect of land use changes in one location on housing prices in another within the same metropolitan area.

This framework was then used to evaluate the impact of the California Transit Housing Bill (SB 827) on housing prices and affordability in San Francisco. The findings show that SB 827 has the potential to substantially decrease San Francisco's housing prices by overriding local zoning restrictions on medium density residential development within the half-mile buffer around major train stations in the Bay Area. However, according to the findings of this research, despite the likely substantial impact on housing prices, SB 827 alone doesn't appear to be capable of resolving the housing affordability crisis in San Francisco. Finally, this dissertation offers a data and software model for implementing the proposed simulation framework in other metropolitan areas in the US.

© Copyright 2020 by Reza Amindarbari

All Rights Reserved

Spatially Disaggregated Simulation of Interconnections between Land Use Policy, Housing
Markets, and the Affordability Crisis

by
Reza Amindarbari

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Design
Forestry and Environmental Resources

Raleigh, North Carolina
2020

APPROVED BY:

Dr. Perver Baran
Committee Co-Chair

Dr. Ross Meentemeyer
Committee Co-Chair

Dr. Robin Abrams

Dr. Justin Post

Dr. Andres Sevtsuk

DEDICATION

To my wonderful mom, Atefe Fazeli, and loving wife, Shani Sharif

BIOGRAPHY

Reza is a geospatial and data scientist, with a background in architecture and urban design. His current work seeks to bring the power of geospatial analytics, machine learning, and statistics into Architecture, Engineering, and Construction Software. Prior to his doctoral studies, Reza was a researcher at City Form Lab at Singapore University of Technology and Design, and at MIT Department of Urban Studies and Planning, working on studies focused on combining city design with analytics. Reza holds a Master of Statistics from NC State University, a Master of Urban Design from the University of Michigan, and a Master of Architectural Technology and a Bachelor of Architecture from the University of Tehran. He also practiced as an architect for about 6 years between 2003 and 2009 before moving to the United States.

ACKNOWLEDGMENTS

This work would not have been possible without the support of my professors, fellow doctoral candidates, family, and friends. First, I would like to thank my advisors, Drs. Perver Baran and Ross Meentemeyer, for their continuous guidance, care, and encouragement throughout the years. I would also like to express my gratitude to the other members of my doctoral committee, Drs. Robin Abrams, Justin Post, and Andres Sevtsuk, for their guidance and support throughout this process. I am grateful for the teaching opportunities Dr. Abrams made possible for me. Dr. Post's courses on the fundamentals of statistics introduced me to new avenues of study, exposing me to the world of data analytics and science. My interest in data analytics for urban planning and the very early seeds of my doctoral research were formed after Dr. Sevtsuk's course entitled Measuring Urban Form, and later when working under his supervision at the City Form Lab.

I would also like to thank the faculty and staff of the College of Design and Center for Geospatial Analytics, especially Dr. Soolyeon Cho, head of the PhD in Design program, for his exceptional care, passion, and support, and Dr. Laura Tateosian for trusting me, giving me the opportunity to collaborate on the Geo-visualization of Conflict Economies project, and allowing me to assist her in teaching Geospatial Programming Fundamentals. I am also thankful for Sarah Slover, who has been a tireless resource for making sure everything administrative was in order.

I am also so grateful to my amazing friends and colleagues in the PhD in Design Program and Center for Geospatial Analytics, with whom I shared many treasured experiences and unforgettable conversations. Special thanks to Dr. Mohsen Ghiasi, Dr. Payam Tabrizian, Dr. George Hallowell, Makiko Shukunobe, Ghazal Kamyabjou, Sahand Azarby, Saeed Ahmadi,

Ezgi Balkanay, Dr. Vaclav Petras, Dr. Anna Petrasova, and Dr. Ece Altinbaşak for their moral support and friendship over the years.

Finally, I am so very thankful for the unconditional love, provision, and constant encouragement of my mother, Atefe Fazeli. Without her continuous support, I would not be where I am today. And words cannot express how indebted I am to my beloved wife, Shani. None of this would have been possible without her love and encouragement.

TABLE OF CONTENTS

LIST OF TABLES	vi
LIST OF FIGURES	vii
Chapter 1: Introduction	1
1.1. The Current State of Housing Affordability Crisis in the United States	1
1.2. Zoning and Housing Affordability.....	3
1.3. Modeling the Interaction between Land Use and Home Prices.....	7
Chapter 2: Spatially Disaggregated Simulation of Interactions between Home Prices and Land Use Change	9
2.1. Introduction.....	9
2.2. Simulation Framework.....	14
2.2.1. Overview	14
2.2.2. Residential Real Estate Pricing Model	15
2.2.3. Supply-to-Demand Ratio Measurement	17
2.3. Simulation Application	21
2.3.1. Study and Influence Areas	21
2.3.2. Data and Model Implementation	22
2.3.3. Simulation Experimentation Scenarios.....	25
2.4. Results.....	26
2.4.1. Model Validation and Parameter Estimates.....	26
2.4.2. Simulation Experimentation Results.....	30
2.5. Conclusion	36
Chapter 3: Impact of the California Transit Housing Bill on Housing Affordability in San Francisco	38
3.1. Introduction.....	38
3.2. Method	41
3.2.1. Study area and influence geography	41
3.2.2. SimP-R: A Real Estate Pricing Simulation Framework	44
3.2.3. Housing and Job Scenarios	45
3.2.4. Measuring Housing Affordability.....	48
3.3. Simulation Results and Discussion.....	49
3.4. Conclusion	56
Chapter 4: A Conceptual Data and Software Model	58
4.1. Challenges and Opportunities in Implementing SimP-R.....	58
4.2. A Data and Software Model for SimP-R.....	62

LIST OF TABLES

Table 2.1	Coefficient estimates (absolute and elasticity) for the pricing model's linear regression	29
Table 2.2	Summary of scenarios and simulation results	30
Table 3.1	Summary of scenarios and their predicted price effect	53
Table 3.2	Housing affordability under different scenarios.....	56

LIST OF FIGURES

Figure 2.1	SimP-R’s conceptual framework.....	15
Figure 2.2	Locations of hypothetical developments scenarios for simulation experimentations	26
Figure 2.3	Optimization of α	27
Figure 2.4	Negative correlation between Census block group-level supply-to-demand ratio and average price per square foot.....	27
Figure 2.5	The importance of predictors in the random forest implementation of the real estate pricing model measured as total decrease in the sum of squares of errors from splitting the predictors space on each predictor.....	28
Figure 2.6	Census block group-level supply-to-demand ratio values and simulated price changes	31
Figure 2.7	Predicted average prices vs. supply-to-demand ratio	35
Figure 3.1	Population of the Bay Area’s top 4 most populated counties from 1860 to 2010.....	43
Figure 3.2	Percent price changes in response to scenarios that build upon ABAG’s 2025 and 2040 housing and job projections	54
Figure 4.1	A low-fidelity data and software model for SimP-R.....	63

CHAPTER 1: INTRODUCTION

1.1. The Current State of Housing Affordability Crisis in the United States

Access to affordable housing is a major determinant of one's quality of life.

Unfortunately, many households' in the United States are burdened by the cost of housing. While real wages have been stagnant for most Americans, home prices (after adjusting for inflation) have increased about 60% on average since the 1980s. According to the latest American Community Survey (US Census Bureau, 2018), nearly half of renter households (20.8 million) across the United States spend more than 30% of their income on housing, and for more than half of those households (11 million) the monthly rent accounts for more than 50% of their take-home pay. The situation is just slightly better for homeowners. Approximately 17 million homeowner households (about 23%) spend more than 30% of their take-home pay on housing, with 7.3 million (about 43%) spending more than 50% of their income on housing.

Spending a sizable share of income on housing leaves many low-income households with insufficient resources for other essentials, including food, savings, and even healthcare, making their finances particularly fragile (California Legislative Analysts' Office, 2015). The high cost of housing forces households to make serious tradeoffs between housing cost, commute time, and housing size. To reduce the housing cost burden, many households choose less expensive housing options further away from their workplaces, and commute further to work each day, which aside from having adverse effects on the environment and personal health (Costa, Pickup, & Di Martino, 1988; Jackson, 2003), can also be a substantial cost burden itself. Alternatively, some households choose smaller and thus more crowded housing situations, which often adversely impact their wellbeing (Myers et al., 1996; Solari & Mare, 2012).

The housing affordability crisis is not limited to major metropolitan areas of the US, where housing prices are higher than average. About 47% of renter households in rural areas of the United States are also burdened by the cost of housing (US Census Bureau, 2018). In fact, nowhere in the United States (either urban or rural areas), a renter household would be able to afford a median-priced two-bedroom apartment by working 40 hours a week on minimum wage (National Low-Income Housing Coalition, 2018).

While the housing affordability crisis is a nationwide issue, its root causes are not the same in all cities. Comparing housing prices to the cost of new construction, Glaeser and Gyourko (2003) found that while in many areas of the United States housing affordability is more due to poverty and low wages, in many major metropolitan areas, it is primarily a supply issue. These researchers found that in most of the US, housing prices are generally very close to and sometimes even lower than the cost of new construction (e.g., in Detroit and Philadelphia). In such areas, housing affordability cannot be achieved by increasing supply without sufficient growth in income or technological breakthroughs that substantially reduce construction cost. However, the authors showed that in many areas where home prices are substantially higher than construction and land costs, it is the insufficient supply of housing due to zoning restrictions that drives housing unaffordability. Matlack and Vigdor's (2003) findings also suggest that in tight housing markets, increasing income inequality since the 1980s has exacerbated low-income households' housing cost burden and accelerated gentrification because high-income households enter low-income households' housing market sector when there is a housing shortage. Many other works (Calder, 2017; Dowall, Landis, & Frieden, 1979, 1982; Glaeser, Gyourko, & Saks, 2005; Quigley & Raphael, 2005) have also blamed zoning (or land use regulations in general) for housing shortage and affordability crisis in major metropolitan areas across the US.

In addition to impacting the supply side of the housing market by limiting the supply of housing and determining the locations and densities of residential developments, zoning (and other land use regulations) plays a key role on the demand side. An extensive body of literature (e.g., Alonso, 1964; Kockelman, 1997; Franklin & Waddell, 2003; Hwang & Thill, 2010; Mills, 1967; Muth, 1969) has shown both empirically and theoretically that proximity to employment locations is a major determinant of housing demand and, as a result, housing cost. Land use ordinances can impact the demand side of the housing market by controlling the locations and densities of workplaces (e.g., offices, retail establishments, industrial sites).

1.2. Zoning and Housing Affordability

In the United States, more than 90% of local governments regulate land use by zoning. Zoning is the practice of dividing the geographic extent of a city into zones and specifying certain land use ordinances for each (Fischel, 2015). These ordinances vary vastly, but often include restrictions on the type and intensity of activity and physical properties of buildings.

Many communities in the United States had basic forms of land use ordinances as early as the seventeenth century, typically addressing issues such as fire safety, light and air rights, street frontage, and building height and style (Fischel, 2004; Hirt, 2014). In 1916, New York City became the first American metropolis to enact a comprehensive zoning code that covered the entire urban area (Hirt, 2014; Fischel, 2015). Many other cities and suburban communities created their own zoning ordinances in the 1920s and 1930s (Fischel, 2004, Fischler, 1998; Weiss, 1987). Eight cities had enacted comprehensive zoning ordinances by the end of 1916, 68 more by the end of 1926, and an additional 1,246 by the end of 1936. Today, zoning is still an integral part (and in many cases, even the main driver) of comprehensive urban planning in America (Babcock, 1966; Haar, 1955; Nelson, 1977; Karkkainen, 1994).

Land rights controls are economically justified as a means of protecting property values from the negative externalities of non-conforming uses and densities (Ellickson, 1973; Fischel, 2004, 2015; Lai, 1994; McMillen & McDonald, 1993, 1999). They are also justified as a tool for preserving natural and environmental resources (Fischel, 2004) and historic landmarks and urban fabrics (Costonis, 1972). However, a historical review of American zoning suggests that zoning has first and foremost served to protect exclusively single-family residential areas from mixed-use and denser forms of residential development. Zoning has also often been used to deliberately exclude certain racial and sociodemographic groups from certain areas, often exclusively single-family residential areas with dominantly white residents.

Fischel (2004) explained that zoning ordinances were first introduced in response to major urban transformations caused by shifts in transportation and construction technologies in the post-Industrial Revolution era. As Fischel (2004) suggested, these transformations began with the emergence of streetcars in rapidly growing industrial cities, which for the first time allowed urban workers to live away from urban centers. These streetcars could not carry large amounts of freight, so only limited small-scale commercial activities such as retail stores moved away from the central city, while noxious heavy industries remained in central urban areas. In fact, streetcars shaped the first transit-oriented development in the United States, with multifamily housing and limited commercial establishments emerging close to streetcar lines while single-family houses were primarily located a few blocks away.

With the introduction of Ford's Model-T in 1908, more and more middle-class families began to live in exclusively single-family districts, further away not only from the central city but also from streetcar lines (Fischel, 2004). However, the introduction of motorized buses and trucks a few years later changed this dynamic. Motorized trucks allowed heavy industry to also

move away from the central city, taking advantage of the abundance of cheap land. Their workers could use passenger buses to commute to these new job locations. Passenger buses also liberated the construction of multifamily apartments from proximity to streetcar lines, and thus multifamily rental units began to emerge in existing single-family districts. Heavy industry and multifamily apartments became a major source of concern for single-family homeowners and developers. Homeowners were concerned about property values, and developers needed to deal with new uncertainties affecting their investment decisions; vacant lots in exclusively single-family areas could now be used for industrial purposes or apartments. It didn't take long for cities to begin enacting zoning ordinances to protect single-family districts from other forms of development that weren't desirable to single-family homeowners (Fischel, 2004; Hirt, 2014).

In many places in the US, zoning ordinances weren't only used to restrict the development of multifamily and rental housing in exclusively single-family districts, but also to exclude certain racial and socioeconomic demographics from white neighborhoods (Donovan & Neiman, 1995; Rothwell & Massey, 2009; Pendall, 2000; Pogodzinski & Sass, 1990; Resseger, 2013). Many cities across the United States explicitly attempted exclusionary zoning ordinances to keep and/or move nonwhite residents out of white neighborhoods (Prakash, 2013). While some of these provisions were blocked by local and federal courts, it wasn't until 1917 that the United States Supreme Court ruled explicit racial zoning to be unconstitutional (*Buchanan v. Warley*, 1917).

For instance, in 1890, San Francisco passed a provision known as the Bingham Ordinance that mandated the residential segregation of Chinese Americans; however, it was quickly invalidated by the Circuit Court for the Northern District of California (*In re Lee Sing*, 1890). Given the legal challenges to explicit racial zoning provisions, cities often resorted to

implicit ordinances that made certain districts unaffordable for low-income communities of color and immigrants (e.g., strict minimum unit size requirements), particularly by creating exclusively single-family zones that excluded rental apartments. San Francisco's 1870 Cubic Air Ordinance, which required 500 cubic feet of space per person living in lodging houses, is an example of such a regulation. The Cubic Air Ordinance, which targeted Chinese American residents and Chinese immigrants, was initially blocked by a local court, but went into effect after a similar provision was introduced at the state level (Yang, 2009).

Most cities continued to introduce implicit exclusionary zoning codes even after the Supreme Court's ruling against racial zoning provisions, and such provisions have been shaping America's racially segregated cities and metropolitan areas to the present day. For instance, Pendall (2000) studied 1,510 cities, towns, counties, and townships in 25 metropolitan areas, finding that communities with more restrictive zoning tended to have smaller percentages of Hispanic and African-American residents. The work also showed that the share of African-American and Hispanic residents dropped between 1980 and 1990 in communities with tight restrictions on density and multifamily housing, after controlling for the overall change in the share of African-American and Hispanic populations in the metropolitan area. Using block-level zoning and census data for Massachusetts, Resseger (2013) also showed a significant and strong positive association between the share of African-American and Hispanic residents and the number of units permitted per acre, lot size, and quantity of multifamily housing permits.

In general, these implicit exclusionary zoning ordinances excluded the targeted racial and socioeconomic demographics by setting restrictions and requirements that blocked housing options that were affordable for those demographics.

1.3. Modeling the Interaction between Land Use and Home Prices

Given the key role that land use regulation plays on both the supply and demand sides of the real estate market, they must be an integral part of policies and plans for mitigating the housing affordability crisis. The geospatial nature of the interaction between land use control and the real estate market is also important in such considerations. While the authority behind land use regulations is often split among the municipalities that comprise a metropolitan area, the real estate market is naturally continuous across the metropolitan area due to the households' mobility. A household's choice to live in one jurisdiction and commute to their workplace in another area suggests that every housing unit can serve as a substitute for the housing supply in other parts of the metropolitan area. On the demand side, every household can generate demand for housing options in other areas of the metropolitan area. Given the mobility element of households' housing decisions, home prices in a municipality are expected to be affected by land use decisions across a metropolitan area; equally, a municipality's land use decisions are expected to impact home prices across the metropolitan area.

Developing geospatial models for this interaction between land use and home price on a spatially disaggregated level enables decisionmakers to evaluate the impact of their land use policies and plans from a housing affordability perspective. Using such models allows them to explore how different land use scenarios may affect housing prices in different locations in a metropolitan area. However, existing standalone residential real estate pricing models are insensitive to changes in land use. In addition, while the output of existing pricing models can be adjusted to reflect land use change price effects by connecting them to supply/demand models in an integrated land use and transportation model (ILUTM), the data preparation, calibration, and

training for ILUTMs are nontrivial and impractical for many municipalities and planning agencies.

Thus, this dissertation focuses on geospatial modeling of the interaction between land use and the real estate market and its application in informing housing affordability policies. Chapter 2 presents a paper titled “Spatially disaggregated simulation of interactions between home prices and land use change.” In this chapter, I first review the current state of residential real estate pricing models both standalone and in ILUTMs and discuss their respective limitations. I then propose a framework for simulating changes in housing prices in response to land use change in metropolitan areas on a spatially disaggregated level. I introduce a novel gravity-based measure for the housing supply-to-demand ratio, which makes the real estate pricing model of the proposed framework sensitive to land use change. I then demonstrate the application of the proposed simulation framework in San Francisco through a series of simulation experiments.

Chapter 3 presents a paper titled “Impact of the California transit housing bill on housing affordability in San Francisco.” In this chapter, I use the proposed simulation framework proposed in Chapter 2 to evaluate the effects of California Senate Bill 827 (2018), *Planning and zoning: Transit-rich housing bonus*, on housing prices and housing affordability in San Francisco. In this Chapter, I first review the current state of housing affordability in San Francisco and the Bay Area. I then present a series of simulation experiments designed to demonstrate the effect of SB 827 on housing prices and discuss the simulation results, with a focus on housing affordability. Chapter 4 offers a data and software model for implementing the proposed simulation framework in other metropolitan areas in the US. I conclude and offer directions for future work in Chapter 5.

CHAPTER 2: SPATIALLY DISAGGREGATED SIMULATION OF INTERACTIONS BETWEEN HOME PRICES AND LAND USE CHANGE

2.1. Introduction

Land use regulations, despite being essential for preventing land-use conflicts and protecting environmental resources and public health and safety, tend to limit the supply of housing by restraining density (e.g., floor area ratio, unit density, land coverage) and building height or by imposing other forms of requirements. Strict land use regulations have been blamed for insufficient housing supply and housing affordability challenges in many growing metropolitan areas of the United States (Frieden, 1979; Dowall & Landis, 1982; Glaeser & Gyourko, 2003; Glaeser, Gyourko, & Saks, 2005; Quigley & Raphael, 2005; Calder, 2017). In addition to the supply side, land use regulations can play a key role on the demand side of the housing market. They regulate the location of places of work (such as offices, retail establishments, and industrial facilities), the proximity to which determines spatial variation in housing demand and prices across a metropolitan area (Kockelman, 1997; Franklin & Waddell, 2003; Hwang & Thill, 2010). Also, the impact of a local jurisdiction's land use regulations on housing prices is not necessarily contained within its boundaries. Household housing location choice is a trade-off between commute length and housing cost, and many opt to live and work in different jurisdictions. Housing options in different jurisdictions of a metropolitan area can serve as substitutes for one another, and job opportunities in one location can generate housing demand across a metropolitan area.

A residential real estate pricing model that represents the interaction between land use change and housing market on a spatially disaggregated level could serve a critical role in the evaluation of land use policies and plans, particularly in metropolitan areas facing a housing

affordability crisis. Such a model would enable decision makers to explore the effects of different land use scenarios on housing prices and their spatial variability across the study area. Existing real estate pricing models have, however, focused primarily on automated valuation and mass appraisal, often under existing (observed) equilibrium. The interaction between land use change and home price has not been a primary concern in prior efforts in real estate price modeling.

Except for rare examples that use geostatistical methods like kriging (e.g., Basu & Thibodeau, 1998; Case, Clapp, Dubin, & Rodriguez, 2004; Bourassa, Cantoni, & Hoesli, 2010; Kuntz & Helbich, 2014) or location choice models with price formation (Wang & Waddell, 2013), real estate prices are often statistically estimated simply as a function of a series of explanatory variables. Most commonly, using an ordinary least squares linear regression formulation, this function is expressed as a hedonic model (e.g., Rosen, 1974; Goodman, 1978) in which housing is considered a composite good, with its overall price being the sum of the equilibrium prices of its constituent qualities. More recently, non-parametric (and thus, non-hedonic) regression methods such as random forest (Antipov & Pokryshevskaya, 2012) and artificial neural networks (e.g., Tay & Ho, 1992; Din, Worzala, Lenk, & Silva, 1995; Hoesli & Bender, 2001; Selim, 2009; Curry, Morgan, & Silver, 2002; Kauko, 2003; Liu, Zhang, & Wu, 2006) have also been used in real estate pricing models, mainly due to their flexibility and often better performance.

Regardless of the learning method, these models are not sensitive to shifts in the observed supply-demand equilibrium, and if the models are temporally explicit, they only implicitly capture supply and demand changes as a function of time and temporal dependencies (e.g., Case, Clapp, Dubin, & Rodriguez, 2004). One main challenge to explicitly expressing home prices as a

function of housing supply and demand is to specify a disaggregated measure of supply to demand. Coppola, Ibeas, del l'Olio, and Cordera (2013) appears to be the only study that has tried to include a disaggregated measure of the supply-to-demand ratio as a predictor in their real estate pricing model. The authors specified the supply-to-demand ratio as the ratio of the number of housing units to the number of households within small aggregated zones (very similar to the inverse of the vacancy rate). However, this specification fails to account for household mobility. That is, housing units in one zone can serve as substitute choices for units in other zones, and people living in one zone can generate demand for units in others because they can move and substitute their current units with units in other locations.

While standalone real estate pricing models are insensitive to land use change, if placed in an integrated land use and transportation model (ILUTM) (e.g., de la Barra, 1989; Echenique et al., 1990; Martínez, 1996; Waddell, 2002; Hunt & Abraham, 2003; Salvini & Miller, 2005), their outputs can be adjusted with feedback input from the ILUTM's supply and demand module to reflect the price effect of land use change. However, the process for such a setup, data preparation, calibration, and output validation across all modules and for all interactions is nontrivial, and full implementation in a specific metropolitan region can take years and be very expensive. The significant computational resources that an ILUTM needs for operation and their long run times are major barriers to their calibration and scenario exploration (Wegener, 2004; Wagner & Wegener, 2007; Moeckel, Garcia, Chou, & Okrah, 2018). While ILUTMs can be powerful tools for answering complex questions about the interactions of land use, the real estate market, and transportation, particularly in metropolitan areas that have access to quality data and technical resources, they are still far from being a practical option for many municipalities and

decisionmaking authorities, due to their complexity and implementation, operation, and maintenance costs.

Aside from practical complications and data needs, the hyper-comprehensiveness of ILUTMs (Lee, 1973; Klosterman, 1994) in modeling complex interactions can add to the uncertainty of model outputs. Along with information, feedback loops pass uncertainties (including errors and random and unobserved processes) from one module to another. While integrated models can provide a more realistic representation of urban processes and interactions in comparison to single-purpose standalone models, the uncertainties of one module (because of data quality, model assumptions, or arbitrary choices in hyperparameters) can affect the outputs of all interacting modules due to this integration. Linking price modeling (for which more and more high-quality detailed data have been becoming available) to models of location choice or transportation would be suitable only if the detailed and reliable behavioral data that the models need are available, which has rarely been the case. Making inferences about ILUTMs' outputs (e.g., creating prediction intervals) can also be complicated. Neither formal nor numerical methods of inference seem practical for their outputs. Formal methods of inference are impractical because outputs from each ILUTM module relies on the outputs of other modules, each implemented in different ways using a variety of data sources on various scales and units of analysis. Numerical methods (e.g., bootstrapping) can be extremely computationally expensive, as that requires many iterations of resampling, training all models, estimating all parameters, and conducting simulations runs, which can be practically impossible.

Given the complexity of ILUTMs both in their implementation and operation and estimating the uncertainty of their output, there is a strong case for the development of simple single-purpose real estate pricing models that are sensitive to land use change. Building such

models requires the introduction of novel spatially disaggregated specifications of supply-to-demand ratio.

In this research, we present SimP-R, a simple-to-implement framework for simulating disaggregated residential real estate prices that is sensitive to residential and commercial development (i.e., land use) scenarios. We introduce a novel gravity-based measure of the supply-to-demand ratio as a predictor in SimP-R's residential real estate pricing model. This measure represents the effect of disaggregated residential and commercial development scenarios on disaggregated residential real estate prices. We also explore the application of a non-parametric machine learning technique – random forest – on residential real estate pricing in a sample city (i.e., San Francisco) and compare its performance (i.e., prediction accuracy) and output patterns to commonly used linear hedonic pricing models. While linear models are more interpretable due to their explicit representation of effects, random forest models tend to achieve a lower prediction bias given their flexibility and lack of functional form. Unlike other nonparametric methods, random forest models are also not prone to overfitting (i.e., they vary little from one sample set to another) because their predictions are the average of many trees. Also, random forest models don't require observations to be independently distributed. Thus, they can be used with spatially or temporally correlated data without any explicit treatment of the spatial and temporal autocorrelations present in the data.

We organized this research as follows. In Section 2, we describe the SimP-R simulation framework and its core residential real estate pricing model, including its supply-to-demand ratio algorithm. In Section 3, we present the implementation of the simulation framework for the city of San Francisco's housing prices, with the San Jose–San Francisco–Oakland Combined Statistical Area (CSA) as the influence geography. In this section, we also present data and

features specific to the study area and discuss model performance and validation. In Section 4, we report results from the simulation experiments that we designed to demonstrate the capabilities of SimP-R in exploring the effects of alternative development scenarios on housing prices. We conclude and offer direction for future work in Section 5.

2.2. Simulation Framework

2.2.1. Overview

SimP-R is a framework for simulating changes in housing prices in a study area in response to shifts in the quantity and spatial distribution of jobs and housing within the study area's influence geography (i.e. the geographical extend beyond the study area, whose housing supply and employment opportunities can still impact the study area's home prices). The core of SimP-R is a residential real estate pricing model with a novel gravity-based measure of the housing supply-to-demand ratio as a primary predictor of housing prices. This measure is a function of the spatial distribution of jobs and housing, and thus can make the pricing model sensitive to residential development and employment change scenarios.

The other integral part of SimP-R is the Supply-to-Demand Ratio Measurement Algorithm (SDRMA). For a given input development scenario (i.e., a new spatial distribution of housing and jobs within the influence geography), the SDRMA first computes the supply-to-demand ratio values at all locations across the study area at the desired level of analysis. Then, using the computed supply-to-demand ratio values, the pricing model predicts new prices for housing units or a random sample of them under the given scenario (see Figure 2.1).

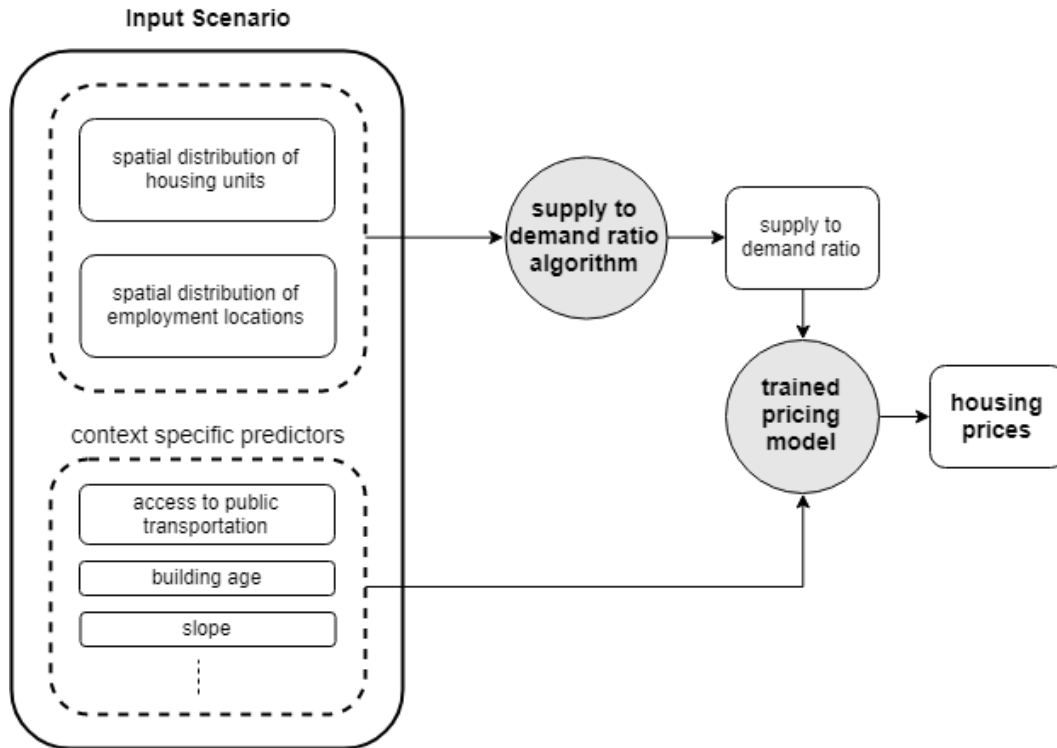


Figure 2.1. SimP-R's conceptual framework. SimP-R is composed of two modules: the pricing model and the SDRMA. For a new spatial distribution of housing and jobs, the SDRMA computes updated supply-to-demand ratio values at every location in the study area and passes them to the pricing model. The pricing model predicts housing prices based on new supply-to-demand ratio values.

2.2.2. Residential Real Estate Pricing Model

The residential pricing model is a regression function that formulates the price of a unit area of housing based on a series of predictors, each describing a quality of housing units or their location. Regardless of the learning technique used, the pricing model is trained based on the price (per unit of area) and value of predictors observed in a randomly sampled set of housing units in the study area. We used the price at which the sample unit was last sold as its observed price.

The set of predictors can include variables such as building age, plot size, slope, or access to public transit, parks, or retail establishments. While the set of predictor variables can vary

from one metropolitan area to another and be determined with the help of any feature selection technique, the supply-to-demand ratio is a required predictor for SimP-R’s real estate pricing model. It is the explanatory variable that makes the pricing model sensitive to residential development and changes in the spatial distribution of jobs. In the next section, we show that the supply-to-demand ratio is in fact the most important predictor of home prices in our study area of San Francisco. By using spatial predictors such as the supply-to-demand ratio, the pricing model can implicitly explain variability in prices attributable to location.

Part of variability in observed home prices can also be due to variations in sales transaction times. We accounted for the temporal variability of the observed home prices by adjusting them to a single base time using zip code-level monthly price indices. We scaled the sales prices observed to the ratio of the base year-month index to the observed year-month index (Equation 2.1):

$$P_{adj} = P_{obs} * \frac{I_b}{I_o} \quad (2.1)$$

where P_{adj} is the price adjusted to the base year-month, P_{obs} is the observed sales price, I_b is the price index of the base year (for the zip code into which the unit falls), and I_o is the price index of the month-year in which the transaction occurred. The values of some of the predictors can also be time dependent. For such variables (e.g., the supply-to-demand ratio) we used the value observed in the base year.

We took this implicit approach to the spatial and temporal variability of home prices rather than explicitly modeling their spatial and temporal autocorrelation structure because using a spatially and temporally explicit model can result in a circular dependency paradox in the simulation process. In a spatially and temporally explicit model, the price of each unit is dependent upon the price of its spatial and temporal neighbors. Changing an input variable under

a given scenario (e.g., changing the supply-to-demand ratio) is expected to impact the price of all neighboring units, and thus predicting the new price of each unit requires the updated price of its neighbors under the new scenario. However, the updated prices of the initial unit's neighbors are unknown as well, since their prices are dependent upon their neighbors' prices, including the initial unit. Using a spatially and temporally explicit model for simulation would also require a strong assumption that the estimated covariances would remain valid as prices change in response to changes in input variables.

Model validation and feature selection

We used k-fold cross validation to measure the pricing model's performance in terms of mean average percentage error (MAPE). In this method, the sample data are randomly split into k equal subsamples. Each subsample is then used once as the test set for validation, while the model is trained using the other k-1 subsets. The average of average errors across all k subsamples provides an estimate of the model's prediction error.

To identify and exclude less important predictor variables from the initial set of candidate variables, we applied a forward selection approach to the k-fold cross-validation process. We sequentially added variables that most improved prediction accuracy in the k-fold test sets. We first added the variable that resulted in the largest reduction in MAPE across all k-fold test sets. Since accuracy is measured in the test sets, adding a new variable can sometimes increase the prediction error. Therefore, forward selection continued until all features were included in the model. Then, from the sequence of fitted models we chose the one with lowest average MAPE.

2.2.3. Supply-to-Demand Ratio Measurement

The monocentric models introduced by Alonso (1964), Mills (1967), and Muth (1969) were the earliest to link housing demand and prices to the accessibility of employment locations.

In their models, all individuals were assumed to commute to the central business district (CBD) for work, and thus proximity to the CBD reduced individuals' cost of daily trips to work. Given their lower cost of travel to the CBD, locations closer to the CBD were considered more in demand and thus had higher housing prices. That is, individuals maintained the same level of utility at different locations in a city by trading off between housing and travel costs.

Using gravity-based measures, a series of studies (Kockelman, 1997; Franklin & Waddell, 2003; Hwang & Thill, 2010) demonstrated the impact of job accessibility on housing prices in settings where employment locations were dispersed or agglomerated in multiple centers. A gravity-based measure assesses any given location for the total quantity of a type of activity (e.g., number of jobs) accessible from that location, weighting destinations with an exponential decay function of travel cost or distance. Unlike Kockelman (1997) and Hwang and Thill (2010) who used aggregated housing data by census tract, Franklin and Waddell (2003) presented a disaggregated model of housing prices by individual housing units based on historical sales transactions. All of these studies showed a positive correlation between job accessibility and housing price. That is, everything else being equal, housing units offering greater accessibility to employment locations tended to have higher prices.

While access to employment locations can represent the demand for housing and prior studies have confirm that it is a strong predictor of housing prices, job accessibility is not sufficient when there is a spatial mismatch between housing supply and demand. As with any other type of goods, the price of housing does not merely depend upon demand, but also on supply. When the supply of housing follows the same spatial pattern as its demand (as it would naturally do in the absence of land use regulations), adding housing supply to the pricing model as an explanatory variable does not offer any additional information. When the supply of housing

follows the demand, supply and demand are highly correlated and provide the same or very similar information. However, strict land use regulations can distort the market and limit the supply of housing where it is being demanded. As the spatial patterns of housing supply and demand diverge, the ratio between housing supply and demand becomes a more important feature for predicting housing prices. In the presence of a spatial mismatch between supply and demand in a market with polycentric or dispersed employment locations, locations with the same intensity of demand can have different housing prices due to differences in the levels of housing supply.

To simultaneously capture housing supply and demand, we propose a novel measure, the ratio of the access to available housing units to the access to employment locations, as the main driver of housing demand, both measured with a gravity-based function:

$$SDR_i = \frac{\sum_{j \in G} \frac{Q_{Hj}}{e^{\alpha * C(ij)}}}{\sum_{j' \in G'} \frac{Q_{Jj'}}{e^{\beta * C(ij')}}} \quad (2.2)$$

where SDR_i is the measure of the supply-to-demand ratio for location i ; Q_{Hj} and $Q_{Jj'}$ are the quantities of housing units at location j and jobs at location j' , respectively; α and β are the decay rates for access to housing and jobs, respectively; $C(ij)$ and $C(ij')$ are the travel cost from i to j and j' , respectively; G is the set of all housing locations; and G' is the set of all job locations. Origins (i 's) and destinations (G and G') can be fully disaggregated. In other words, they can represent the exact locations of housing units and employment. They can also be aggregated to the same or different levels of aggregation if fully disaggregated data are not available or to reduce computation time.

The choice of decay rate hyperparameters for access to housing α and jobs β not only affects the model's accuracy, but also determines how the sensitivity of home prices to new

developments changes with distance. A numerical approach for finding proper decay rates would be to run k-fold cross-validations for different combinations of hyperparameters α and β and choose the combination that minimizes the average cross-validation MAPE. Alternatively, the decay rate for the housing demand generated by employment locations can be estimated based on survey data of commute times (or distances) for employment locations:

$$Q_{Jjd} = \frac{Q_{Jj}}{e^{d*\beta_j}} \quad (3)$$

where Q_{Jj} is the total number of jobs at location j and Q_{Jjd} is the number of workers at location j who commute the distance/time d . With the log transformation of both sides in Equation 3, β_j , the decay rate for location j , can be estimated by linearly regressing $\log(Q_{Jj})$ on time (d) and $\log(Q_{Jjd})$ as the intercept, using Q_{Jj} and Q_{Jjd} values observed from the survey data (e.g. from the American Community Survey for work geographies):

$$\log(Q_{Jj}) = \log(Q_{Jjd}) - \beta_j * d \quad (4)$$

The average of the β_j values can be used as an estimate for β for the entire influence geography. With β estimated, α can be derived numerically based on cross-validation MAPEs as described above. Searching for only one minimizing hyperparameter substantially reduces the number of training runs, thus allowing for finer search intervals and a wider search span for the single hyperparameter α .

This measure of the supply-to-demand ratio accounts only for the housing demand that is generated due to the presence of employment opportunities and doesn't capture the housing demand of households with no working members, such as non-working seniors or students. Thus, the simulated price changes demonstrate only the isolated effect of changes in employment-related demands. That is, the simulation is not sensitive to changes in non-employment-related housing demands and treats them as if they are constant. This, however,

doesn't create a major practical limitation because those segments of the demand for housing rarely fluctuate substantially (as compared to employment-related housing demands), and are rarely the primary driver of housing price in most major metropolitan areas.

SIMP-R's SDRMA computes the proposed measure of the supply-to-demand ratio (see Equation 2) for input scenarios. A component of the SDRMA is a precomputed origin-destination matrix of commute times/cost (the $c(ij)$ and $c(ij')$ values in Equation 2). In this matrix, origins are locations for which prices are predicted. Destinations are employment locations (i.e., demand generators) in denominator and housing units (supply) in the numerator. The matrix can represent a fully disaggregated set of origins and destinations (i.e., individual housing units and employment locations). Instead of computing commute times for a fully disaggregated set of origins and destination (which is computationally expensive), supply-to-demand ratio values can be computed for the desired level of aggregation, such as by census block or block group for destinations, origins, or both. For each destination, the SDRMA takes the quantity of housing units Q_H and jobs Q_J from input scenarios.

2.3. Simulation Application

2.3.1. Study and Influence Areas

This section demonstrates application of the proposed simulation framework, SimP-R, to the city of San Francisco. We used the entire San Jose-San Francisco-Oakland CSA as the influence area for computing the gravity-based measure of the supply-to-demand ratio, since housing prices in the city of San Francisco are expected to be connected to housing supply and employment opportunities in other locations across the greater metropolitan area. Housing location choice is a tradeoff between travel cost and the cost of housing, and thus housing units

in nearby cities in a metropolitan area can serve as substitutes for one another. Employment locations also generate housing demand across the entire metropolitan area's housing market.

In the 1960s, local municipalities in the San Francisco Bay Area began to enact strict land use regulations that significantly constrained the supply of housing (Elliott, 1982; Beitel, 2007). Around the same time, economic growth in the high-tech industry in the Bay Area – particularly San Francisco and nearby Silicon Valley – created tens of thousands of new jobs every year. Unsurprisingly, this was followed by significant and continuous population growth. The combination of a growing demand for housing and severely restricted housing supply has resulted in an extended and ever-increasing housing shortage and extremely high housing prices (Beitel, 2007). The high cost of housing is a major contributing factor to poverty in the Bay Area. San Francisco Bay Area households (particularly low-income households) have to spend a larger share of their income on housing, and thus can spend less on essentials (including saving for emergencies). The high cost of housing has also forced many households to move to relatively more affordable areas further away from employment locations, and thus spend much more time on super-commutes (Taylor, 2015).

2.3.2. Data and model implementation

Learning methods for the real estate pricing model

We used random forest and linear regression methods to implement two real estate pricing models for SimP-R's application to San Francisco. We employed the same sample of housing units and set of predictor variables in both models. RandomForest, an R library created by Liaw and Wiener (2002), was used to train the random forest implementation.

In the case of linear model implementation, given the presence of spatial and temporal autocorrelation in housing prices (Moran's $I = 0.24$, $z = 12.46$ and $p\text{-value} < 0.00001$), we

avoided using the estimated variance of regression coefficients to obtain p-values and make inferences about the significance of predictors. We estimated the predictors' p-values by non-parametric bootstrapping, as it doesn't rely on the independence of observations. We fit a null model for each predictor (a model with all predictors except the one for which the p-value is being estimated). Then, using the null model's coefficient estimates and by resampling its residuals 10,000 times, we generated many datasets that represented the null hypothesis assumption (i.e., that the coefficient for the predictor of interest would be zero). We then fit full models with each simulated null hypothesis dataset, and calculated the p-values as the frequency of cases where the absolute value of the coefficient estimate from the full model with the generated null hypothesis dataset was greater than the absolute value of the coefficient estimates from the full model with the real data.

Housing data

We used public records of home sales from 1999 to 2018 from the Assessor's Office of the City and County of San Francisco to train the pricing model. We randomly selected five units in each census block from the collection of units with at least one sales transaction since 1999 and information related to area and year of construction. In rare cases, when blocks had fewer than five units with the above-mentioned criteria, we used all units in the block that met the criteria. In total, we sampled 10,592 units that met these initial requirements. Since these sales transactions encompassed an extended period of time, we linearly adjusted prices to December 2016 (Equation 1), using Zillow Group Research's monthly zip code-level price per square foot index (n.d.). This index tracks monthly changes in each zip code's average listed price. We chose 2016 as the benchmark for housing prices as it corresponded with our regional job and housing count data.

Finally, we performed single and multivariate outlier analyses based on adjusted per square foot prices and areas of units to identify potentially erroneous records. Using sampling within different ranges of outliers and manual exploration combined with other sources (including Google Maps Street View and satellite imagery), we purged erroneous records. The final purged sample dataset contained 8,204 housing units.

Supply-to-demand ratio

Computing the proposed measure of the supply-to-demand ratio required the travel cost (i.e., time) between all origins (i.e., the locations for which prices were predicted) and destinations (i.e., job and housing locations). Using a routing service to solve the travel cost between all origin-destination pairs would have been extremely computationally expensive for these large sets of origins and destinations. We used census blocks as the level of aggregation for origins within the study area (the city of San Francisco) for which we computed the supply-to-demand ratio SDR_i . For destinations, we used job counts obtained from Esri's 2016 Business Locations and housing counts taken from the 2016 five-year American Community Survey, aggregated at the census block group (CBG) level across the San Jose-San Francisco-Oakland CSA, the designated influence geography. Even at this level of aggregation, the origin-destination matrix for the city of San Francisco with the entire CSA as the influence area had about three million elements for which travel times needed to be computed. We computed commute times between all origin-destination pairs by locally running the Open Source Routing Machine API (Luxen & Vetter, 2011) on OpenStreetMap's roads and traffic data (OpenStreetMap contributors, 2018). The matrix of the origin-destination travel times was then used by the SDRMA to recompute the supply-to-demand ratios used to predict new prices under new input scenarios.

We used the 2016 five-year American Community Survey for the work geographies of workers' commute patterns to estimate the destination-specific decay rate hyperparameter for access to jobs (β_{js}) with the regression formulation in Equation 4.

Study area-specific predictors

In addition to the supply-to-demand ratio, we included the following predictors in the pricing model for San Francisco:

Access to public transportation: For each housing unit, we measured access to public transportation as walking distance to the nearest station. We included one variable for access to San Francisco's local public transit system (Muni), and one for access to the Bay Area's regional transportation systems, including Caltrain and Bay Area Rapid Transit.

Slope: We measured the average slope around each unit based on a slope grid generated from 20-foot interval elevation contours. We assigned to each unit an average of nine surrounding cells, an area of approximately 550 x 550 feet.

Density of street trees: Using San Francisco Department of Public Works' street tree dataset, we measured the linear density of trees along street segments around the block that contained the unit and street segments immediately connected to those segments.

Distance from shoreline: Finally, we included the Euclidean distance from the unit to the shoreline around San Francisco as a proxy for views of the ocean or bay, as well as walking access to the waterfront.

2.3.3. Simulation Experimentation Scenarios

We used Simp-R to simulate the impacts of four hypothesized housing development scenarios on home prices in the city of San Francisco. Three of these scenarios increased residential density (i.e., housing supply) in (a) the low-density residential neighborhood of

Sunset in the city of San Francisco, (b) across the city of Oakland, and (c) around major transit stations (Caltrain) on the San Francisco Peninsula from Millbrae to San Carlos. The Oakland and Millbrae-San Carlos development scenarios show how housing prices in a city can be affected by development decisions in other cities in the same metropolitan area. In the fourth scenario, along with increasing housing density in the city (in the Sunset neighborhood), we increased employment numbers around Caltrain stations from Millbrae to San Carlos. Figure 2.2 shows the locations of the hypothesized developments, and Table 2.2 in Section 4 reports a summary of the scenarios and results.



Figure 2.2. Locations of hypothetical developments scenarios for simulation experimentations

2.4. Results

2.4.1. Model Validation and Parameter Estimates

Figure 2.3 shows how the average five-fold cross-validation MAPE value varies when changing the decay rate hyperparameter for access to housing α while β is set at 0.06. As shown in Figure 2.3 the average five-fold cross-validation MAPE is minimized at $\alpha = 0.045$. We estimated β (0.06) using the regression formulation in Equation 4 with the 2016 five-year American Community Survey for the work geographies of workers' commute patterns.

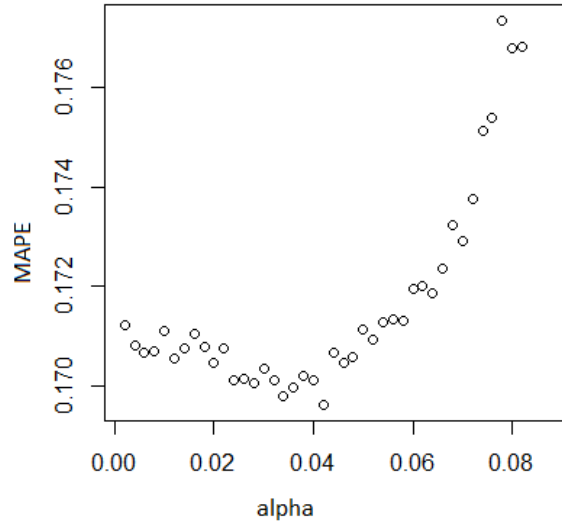


Figure 2.3. Optimization of α . $\alpha = 0.045$ minimizes the average cross validation MAPEs.

Using the decay rate estimates, computed matrix of travel times, and observed housing and job counts at destination locations in the influence geography, we computed the supply-to-demand ratio (Equation 3) for all origins in the study area. Figure 2.4 demonstrates San Francisco’s CBG-level supply-to-demand ratios and average per square foot housing prices, and the inverse relationship between the two ($\rho = -0.43$).



Figure 2.4. Negative correlation between census block group-level supply-to-demand ratio (left) and average price per square foot (right).

The final model selected by forward selection based on the average cross-validation MAPEs included all initial candidate predictors except for the distance to Muni. Results from the random forest model implementation confirmed that the proposed measure of the supply-to-demand ratio was the most important predictor of housing prices in San Francisco (see Figure 2.5). Figure 2.5 shows the total decrease in the sum of squares of errors from splitting the predictors space on each predictor. The supply-to-demand ratio contributed the most to the reduction in prediction error among the other predictors in the pricing model.



Figure 2.5. The importance of predictors in the random forest implementation of the real estate pricing model measured as total decrease in the sum of squares of errors from splitting the predictors space on each predictor

The supply-to-demand ratio was also selected first in the forward selection process, as it resulted in the lowest prediction error (17.8%) among all univariate models made with other predictors, while adding all other features only reduced the average k-fold cross-validation MAPE from 17.8% to 16.9%.

Since we were particularly interested in the effect of land use decisions – here presented by the measure of the supply-to-demand ratio – on the overall spatial pattern of housing price, we aggregated the predicted prices on the CBG level. Averaging naturally reduces error by reducing the variance. Here too, aggregating the predicted home prices by CBG (CBGs’ average home prices) reduced the cross-validation MAPE from 16.9% to 5.5%.

We included the same set of predictors in the linear regression implementation of the pricing model. We only applied log transformation to the supply-to-demand ratio, which resulted in a lower mean squared error. Table 2.1 shows the regression coefficient estimates and their bootstrapped p-values, as well as the elasticity of each predictor (i.e., coefficients from the log-log regression). All predictors turned out to be significant (i.e., they had very small bootstrapped p-values). Elasticity estimates from the log-log regression also confirmed that the supply-to-demand ratio was by far the most important predictor of housing prices. However, all predictors together explained only 27.9% of the variability in housing prices ($R^2 = 0.279$).

Table 2.1. Coefficient estimates (absolute and elasticity) for the pricing model’s linear regression

Coefficient	Estimate	p-value (bootstrapped)	Elasticity	Standard Error	p-value
Intercept	392.20	< 0.0001	5.572	18.14	< 0.0001
Building age	0.26	0.012	0.001	0.10	0.01
Slope	17.01	< 0.0001	0.031	1.05	< 0.0001
Tree density	377.70	< 0.0001	0.549	19.71	< 0.0001
Distance to shore	-0.01	< 0.0001	-0.046	0.00	0.01
Distance to regional transit	0.05	< 0.0001	0.162	0.00	< 0.0001
Supply to demand ratio-log	-3179.00	< 0.0001	-2.352	90.94	< 0.0001

Notes: Standard errors and p-values (highlighted columns) are computed based on the assumption that observations are independent. Since this assumption doesn’t hold due to spatial autocorrelation, standard errors and p-values shouldn’t be used for statistical inference. However, in this case, p-values calculated based on the assumption of independence and normality are very similar to bootstrapped p-values.

In comparison to the random forest model, the linear model resulted in more substantial errors. The linear hedonic pricing model predicted home prices with an average of 22.2% error at the unit level and 14.6% error at the CBG level. The average of the five-fold cross validation MAPE values from the random forest model were, however, 16.9% and 5.5% at the unit and CBG levels, respectively.

2.4.2. Simulation Experimentation Results

We tested both the linear and random forest models in simulation experiments for each scenario and compared the outputs by reflecting upon fundamental differences in the mechanics of the two models. Table 2.2 and Figure 2.6 summarize the simulation experimentation results.

Table 2.2. Summary of scenarios and simulation results

Scenarios	Added units (% change in scenario area)	Added jobs (% change)	Average supply to demand ratio change	Average price change - random forest (Variance)	Average housing price change - linear (Variance)
(a) Increased residential density - Oakland	20563 (10%)	-	0.99%	-1.9% (0.0013)	-3.1% (0.0001>)
(b) Increased residential density - Millbrae to San Carlos stations	22641 (50%)	-	1.15%	-2.4% (0.0017)	-3.7% (0.0001>)
(c) Increased residential density - Sunset	19720 (60%)	-	2.10%	-4.2% (0.0035)	-6.7% (0.0007)
(d) Mixed scenario: Increased residential density & employment	19720 (60%)	50396 (50%)	0.43%	-0.93% (0.0013)	-1.3% (0.0006)

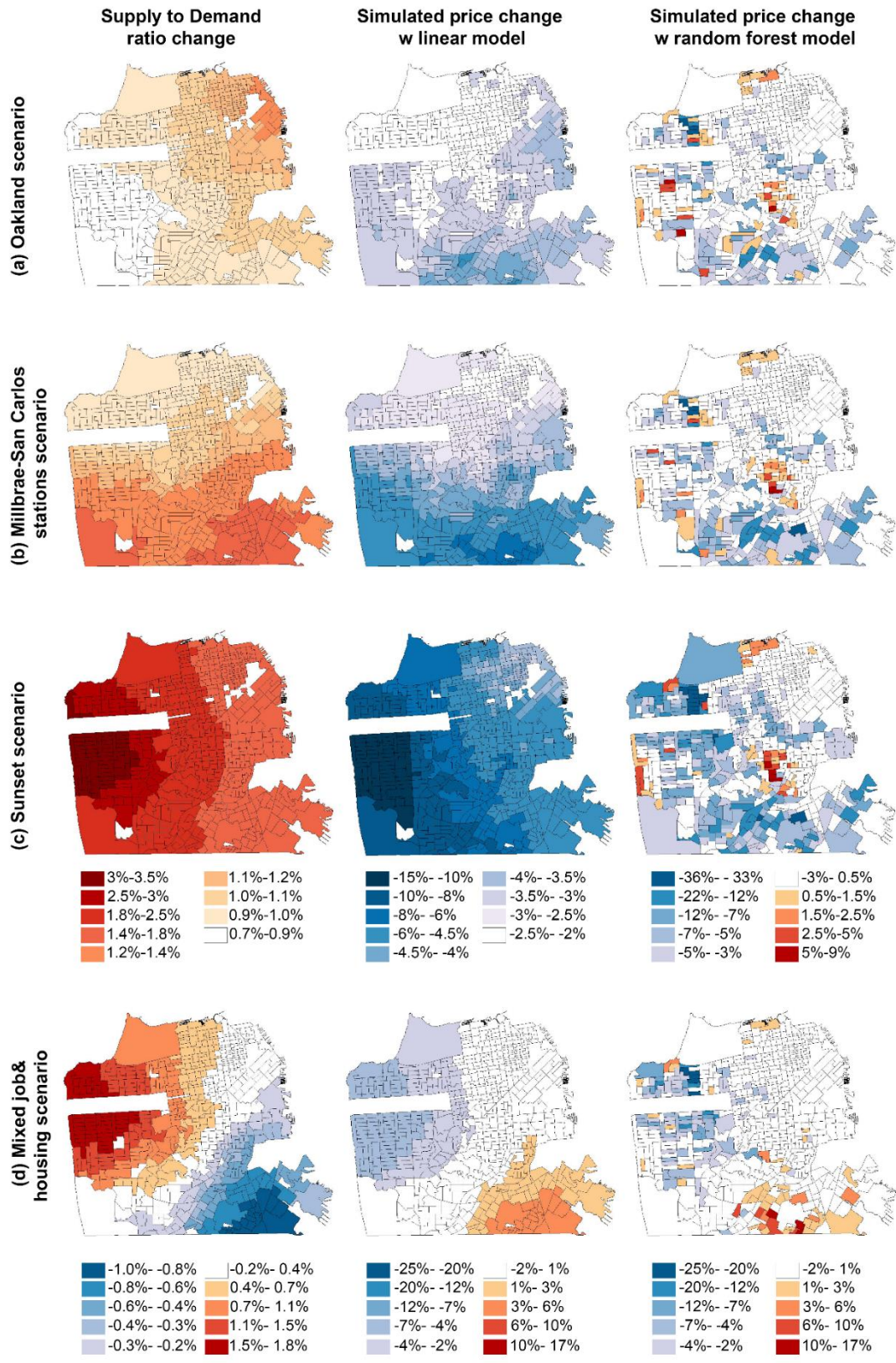


Figure 2.6. Census block group-level supply-to-demand ratio values and simulated price changes

With the linear model, given the negative sign of the regression coefficient for the supply-to-demand ratio effect, the simulated changes in housing prices linearly followed the changes in the supply-to-demand ratio, but in the opposite direction. Thus, for the three scenarios that only increased housing supply (Scenarios a, b, and c), housing prices were predicted to drop in all block groups, on average around 2.4% for each 1% increase in the housing supply-to-demand ratio. Similarly, under the combined housing-commercial development scenario (Scenario d), housing prices were expected to increase where the supply-to-demand ratio dropped, and vice versa. The overall pattern and direction of simulated housing price changes obtained from the random forest model were similar to the outputs of the linear model. That is, on average, housing prices decreased with an increase in the supply-to-demand ratio. In the (a) Oakland and (b) transit area scenarios, both models also demonstrated that even residential developments beyond the city of San Francisco's jurisdiction had a measurable impact on San Francisco's housing prices.

In a small percentage of block groups, however, the random forest model outputs showed an increase in housing price where the supply-to-demand ratio increased, which seems inconsistent with the supply, demand, and price interaction rationale. This inconsistency and the differences between the outputs from the two models can be explained by looking into the internal mechanics of learning and prediction in both models. The linear model breaks down the dependent variable (i.e., housing prices) as a linear combination of predictors and random error. Because of this formulation, the linear model naturally showed that home prices change linearly at a constant rate in response to changes in the supply-to-demand ratio if all other variables are held constant. However, this perfect breakdown of effects and their estimates is invalid in the presence of multicollinearity (i.e., correlations between predictors), endogeneity (i.e.,

correlations between predictors and the error term), and heteroskedasticity (i.e., nonconstant error variance), or if the relationship between the response and predictors is not linear. The random forest model, however, doesn't impose any functional form on the data and does not depend on any of the above-mentioned assumptions of the linear model. This model partitions the feature space into sub-spaces using decision trees such that the variance of observed responses in the sub-spaces is minimized on average to create homogeneous sub-spaces. It then models the dependent variable as the average of observed response values in each sub-space.

In the absence of a perfect linear relationship between the dependent variable and predictors or when one predictor is related to unobserved processes (i.e., the error term) or other predictors, the effect size estimates in the linear regression function (i.e., the regression coefficients) do not represent the unbiased pure and separate effect of the predictors. Our proposed measure of the supply-to-demand ratio only captures job-related demands, and we control for some of the other qualities that influence demand (such as access to transit) by adding them to the model as predictors. If the increase in the supply-to-demand ratio in an area is followed by changes in these other attributes (even in omitted processes) the linear model's outputs will only demonstrate the effect of the supply-to-demand ratio, which would be biased given multicollinearity, endogeneity, or both. Conversely, the random forest model can implicitly capture the combined effect of related processes because of its flexibility and that it doesn't explicitly model the isolated effect of the predictors. Thus, where the net effect of all processes that vary by increasing the supply-to-demand ratio is positive, the random forest model can demonstrate the increase in housing prices.

For example, in areas where the supply of housing increases, cities may tend to allocate more resources for public transportation and more retail establishments may follow the housing

development, which tends to have a positive impact on housing prices. While the isolated price effect of increasing the supply in this scenario is expected to be negative, the combined price effect of all related processes can be positive. The linear model shows the isolated effect of increasing supply in this case, which might not be perfectly isolated given the presence of multicollinearity, endogeneity, or both. The random forest model, however, demonstrates the effect of increasing the supply combined with other changes that may follow it, without explicitly explaining the individual isolated effects and their sizes or directions.

The differences between the two models became more apparent in scenarios that more drastically changed the supply-to-demand ratio. We compared the two models' behaviors in response to changes in the supply-to-demand ratio by gradually adding housing units to Scenario b and c's areas while keeping job numbers constant (and thus increasing the supply-to-demand ratio). To decrease the supply-to-demand ratio, we gradually increased job numbers while we kept housing numbers at the current state. Figure 2.7 demonstrates that the average predicted housing prices from the two models diverged more and more as we increased or decreased the average supply-to-demand ratio away from the observed state. This was because the linear model extrapolates housing prices without any bound, but the random forest model's predictions are always constrained within the observed range of the response value (i.e., price). As a result, with the random forest model, the rate of change in the average simulated housing prices slowed down and eventually flattened out as the average supply-to-demand ratio deviated further away from the observed state.

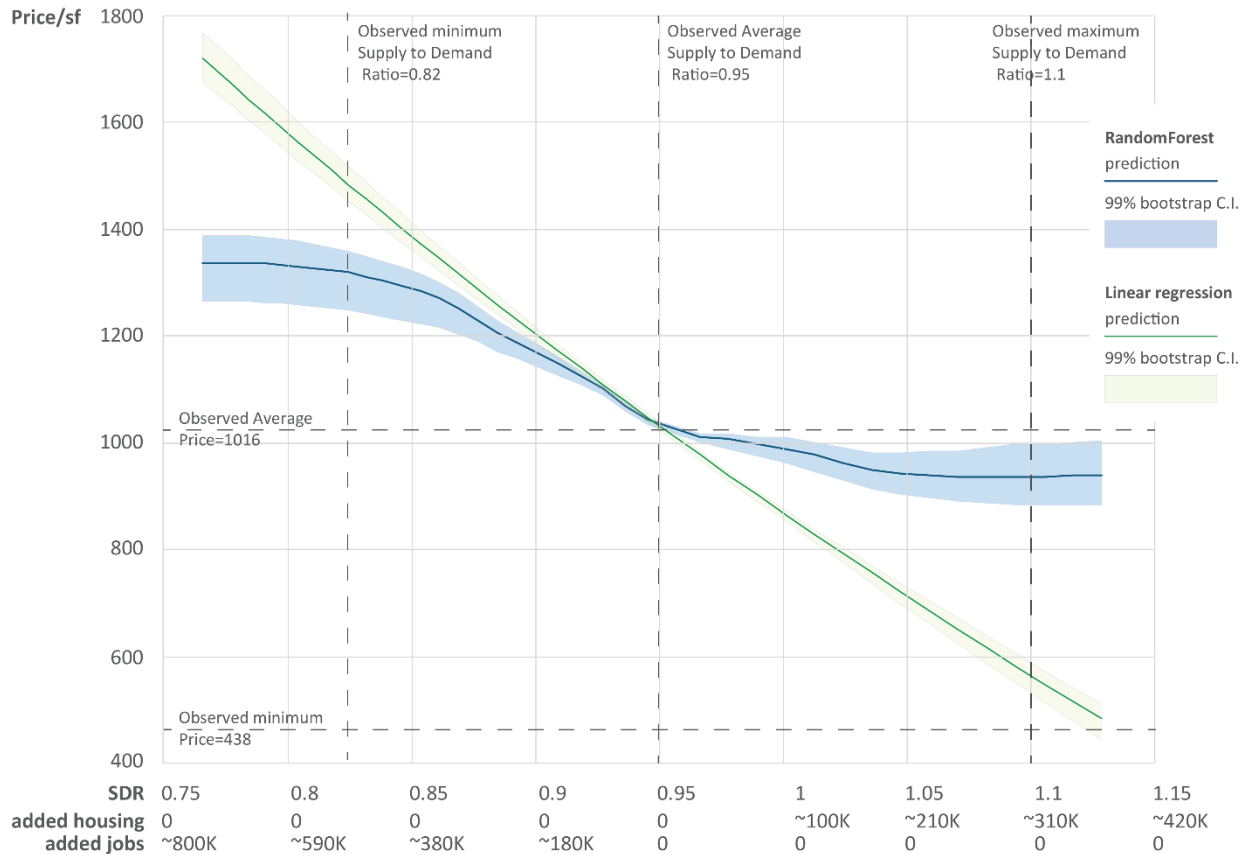


Figure 2.7. Predicted average prices vs. supply-to-demand ratio. As the supply-to-demand ratio values in the input scenarios moved further away from the observed values, the difference between the outputs from the two models increased.

In addition, regardless of the model used, the level of uncertainty increases as we extrapolate farther from the observed state. Given the higher level of uncertainty in more extreme scenarios, using the outputs of the two models can offer a better insight into the possible range of changes in housing prices, particularly considering that the random forest model predicts more conservatively while the linear model predicts without bounds.

2.5. Conclusion

In this research we presented SimP-R, an easy-to-implement simulation framework for exploring the effect of development scenarios on housing prices. The framework couples a real estate pricing model with an algorithm for computing a novel gravity-based measure of the

supply-to-demand ratio as a predictor of housing prices. We implemented SimP-R for the city of San Francisco with the San Jose–San Francisco–Oakland CSA as the influence geography. The results confirmed our proposed measure of the supply-to-demand ratio to be a strong predictor of and inversely related to housing prices in San Francisco. In fact, it was the most important predictor among the variables we examined in this work.

We explored the use of linear and random forest regression learning techniques for implementing the pricing model. The random forest model demonstrated a substantially higher performance (i.e., a lower cross-validated mean absolute percentage error in prediction), particularly at the CBG level (22.2% vs. 16.9% error at the unit level and 14.6% vs. 5.5% error at the CBG level). The outputs of the two models tended to diverge more under scenarios where the supply-to-demand ratio deviated more drastically from the observed state. The differences between the two models' outputs suggest that it is critical to consider each model's internal mechanics of learning and prediction and their limitations and assumptions when interpreting the simulation outputs.

Results from the SimP-R scenario exploration with both the linear and random forest models also confirmed the inverse relationship between housing price and the supply-to-demand ratio. As expected, increasing the housing supply and holding the demand for housing (i.e., job number) constant caused housing prices to drop in both models. The random forest model's results also suggest that while the isolated effect of increasing supply on prices is in general negative, in some locations the combined effect of processes that follow an increase in the supply could result in an increase in housing prices. However, since the random forest model doesn't explicitly model the effects of different processes and their interactions, the results do not offer further insight into the exact mechanics of price increases in these areas.

The simulation experimentation results demonstrated SimP-R's ability to simulate the effect of land-use changes in one location on housing prices in another location within the same metropolitan area. This spatial dependency of local housing prices on development and land-use in other locations highlights the importance of metropolitan area-level coordination among local municipalities to achieve housing affordability. Our work also showcases the capabilities of geospatial models like SimP-R in such intermunicipal collaborations.

CHAPTER 3: IMPACT OF THE CALIFORNIA TRANSIT HOUSING BILL ON HOUSING AFFORDABILITY IN SAN FRANCISCO

3.1. Introduction

In the last 50 years, home prices in metropolitan areas in California, and particularly in San Francisco, have grown much faster than anywhere else in America. While the United States home price index has increased 4.7% every year on average, it has increased 6.95% in California and 8% in the San Francisco-San Mateo-Redwood City metropolitan area (Federal Housing Finance Agency, 2019). The high cost of housing has forced California households to make significant tradeoffs among location, cost, quantity, and quality when making housing choices. Compared to other locations, households in California metropolitan areas, and particularly those that are low income, tend to spend a larger share of their take-home pay on housing and less on other essentials; this situation has become a primary contributor to poverty in California (Alamo & Uhler, 2015). California households also tend to commute further to work and live in more crowded housing in order to reduce housing-related spending (Alamo & Uhler, 2015).

The high cost of housing in California has mainly been attributed to the insufficient and poorly distributed housing supply available to the rapidly growing population in coastal metropolitan areas. For instance, Alamo and Uhler's study for the California Legislative Analysts' Office (2015) suggested that to keep its median housing price from growing faster than the nation's, California should have built 70,000 to 110,000 additional units a year between 1980 and 2010. The most significant shortage has been in the San Francisco Bay Area, with an annual deficit of 50,000 units. This study also suggested that additional units were mainly needed in coastal areas and a more effective strategy would have been to focus on building denser housing in central urban areas. While in locations such as the San Francisco Bay Area the geography and

natural landscape have served as an important barrier to housing development, zoning and local land-use decisions have been primary drivers of the current housing shortage and affordability crisis (Frieden, 1979; Dowall & Landis, 1982; Glaeser & Gyourko, 2003; Glaeser, Gyourko & Saks, 2005; Quigley & Raphael, 2005; Calder, 2017).

California Senate Bill 827 (2018), *Planning and zoning: Transit-rich housing bonus*, seeks to mitigate the housing shortage in California metropolitan areas by overriding local zoning to permit mid-rise housing units in transit-rich areas. The proposed law would require the local government to grant density bonuses to housing developers in transit-rich areas (defined as a half-mile radius around major transit stops or quarter-mile radius around stops on high-quality bus corridors). The law would also exempt developers of transit-rich housing projects from requirements such as minimum parking and limitations on maximum height and floor area ratio if the limitations specified in the local ordinance are less than a specified amount (SB 827, 2018).

Not surprisingly, SB 827 has faced strong opposition from local governments, as it would reduce their control over local planning. Advocates for low-income housing and tenants' rights groups such as the LA Tenants Union and Housing is a Human Right (2018) were other major opponents of the bill, arguing that it would exacerbate the gentrification of low-income communities. Later amendments to the bill required developers receiving a transit-rich bonus to provide a specified percentage of units as affordable housing and increase protections for existing residents (SB 827, 2018), but these didn't garner support from tenants' right groups. SB 827 has also been criticized for its one-size-fits-all approach (the uniform treatment of all transit-rich areas across the state) and disregard for local peculiarities (Dovey, 2018; Schneider, 2018). Surprisingly, some advocates for public transportation and transit-oriented development such as Sierra Club also opposed the California transit housing bill out of concern for further pushback

on public transportation expansion that SB 827 might generate in wealthy low-density neighborhoods (Sierra Club, 2018). Finally, after a tense hearing, SB 827 failed to advance from the Senate Transportation and Housing Committee. The discussion around this bill continues, however, and its sponsors have announced their intention to introduce a new version of SB 827 in the near future (sd11.senate.ca.gov, April 18, 2018).

In this ongoing debate, discussion of the magnitude and spatial variability of SB 827's effect on home prices has been absent. Thus far, no effort has been made to examine on a spatially disaggregated level how home prices might change in different locations in a metropolitan area in response to SB 827's increasing the supply of housing in transit-rich locations and with respect to projected housing and employment changes.

We used SimP-R (see Chapter 2) to explore how average block group-level home prices in the city of San Francisco might change in response to a series of SB 827-compliant development scenarios across the Bay Area. Our work provides insights into how home prices in one location interact with the supply of housing in other areas of a metropolitan region and showcases the capabilities of microscale geospatial modeling to support policy decisions by predicting the potential effects of alternative scenarios.

In Section 2, we describe SimP-R's framework and the input scenarios we developed for exploring SB 827's impact. In Section 3, we report the simulation results under different scenarios based on both the parametric (linear regression) and non-parametric (random forest) implementation of the framework and discuss housing affordability in response to new scenarios. We conclude in Section 4, we conclude with main findings regarding possible impacts of SB 827 on housing affordability in San Francisco and a discussion on the capabilities of microscale geospatial modeling in supporting land use policy decisions.

3.2. Method

We have taken a simulation approach to understanding the disaggregated price effect of SB 827. We used SimP-R to explore how home prices in the city of San Francisco (the study area) might be impacted by increasing housing density in transit-rich areas of the San Francisco Bay Area (the influence geography). We specified a series of scenarios likely to occur if SB 827 went into effect and explored their impact on San Francisco's housing prices. We first built two baseline scenarios for 2025 and 2040 only based on the Association of Bay Area Governments (ABAG) and Metropolitan Transportation Commission's (MTC) housing and employment projections without SB 827 added housing in transit-rich areas. We then built five scenarios based on each of the baseline scenarios, each applying a different ratio of a reference density that we defined based on minimums set by SB 827 for local maximum density and building height in transit-rich areas. We also created five similar scenarios that only apply different density levels to transit-rich areas based on SB 827 but keep other areas as observed in the base year (2016). We then simulated the price effect of each of these scenarios (17 in total) on the census block group (CBG) level and explored their impact on housing affordability by comparing predicted prices against the median household income.

3.2.1. Study Area and Influence Geography

Beginning with the California gold rush in the 1840s and continuing until the early 1900s, the initial influx of people into the San Francisco Bay Area was mainly concentrated in the city of San Francisco (see Figure 3.1). The second wave of people moving to the Bay Area was driven by shipyard jobs, particularly during World War II, and wasn't limited to San Francisco (Kamiya, 2018; Johnson, 1994). Nearly 45% of all cargo shipping tonnage and 20% of warship tonnage built in the United States during World War II was the work of San Francisco

Bay Area shipbuilders (Bonnett, 2000), many of whom were located in the East Bay, including Richmond, Oakland, and Berkeley. During this second influx (between 1900 and 1950), the population of the East Bay county of Alameda increased by nearly 500% (Bay Area Census, n.d.).

The Bay Area's economic boom continued after the war, building on the technological expertise of wartime with local initiatives such as the Stanford Research Park (Luger, 1991). Major companies in military technology and electronics such as Hewlett-Packard, Eastman Kodak, General Electric, and Lockheed Corporation began to grow rapidly in the Bay Area in the 1950s (Leslie & Kargon, 1996). The population continued to grow on average by 4.5% and 3% annually in the 1950s and 1960s, respectively (Bay Area Census, n.d.). High technology industries continued to be the primary driver of economic growth with the rise of the microprocessor industry between 1970 and 1990, followed by software and the Internet industries in the past three decades (Wolff, 2018).

However, while new employment opportunities either directly or indirectly related to high technology industries were attracting an influx of people, San Francisco and some of the other cities in the Bay Area, influenced by neighborhood groups, began to enact strict zoning ordinances in the 1960s (Smith, 1999). Mainly lowering the maximum height and unit density of new structures, these ordinances became a barrier blocking for an adequate supply of housing for the growing population (Dowall & Landis, 1982). As a result of the restrictions on medium and high-density residential developments, most developed areas with limited vacant land (like San Francisco) couldn't house the increase in population. Thus, other nearby areas with more land available for development, particularly Santa Clara county, experienced rapid population growth (see Figure 3.1).

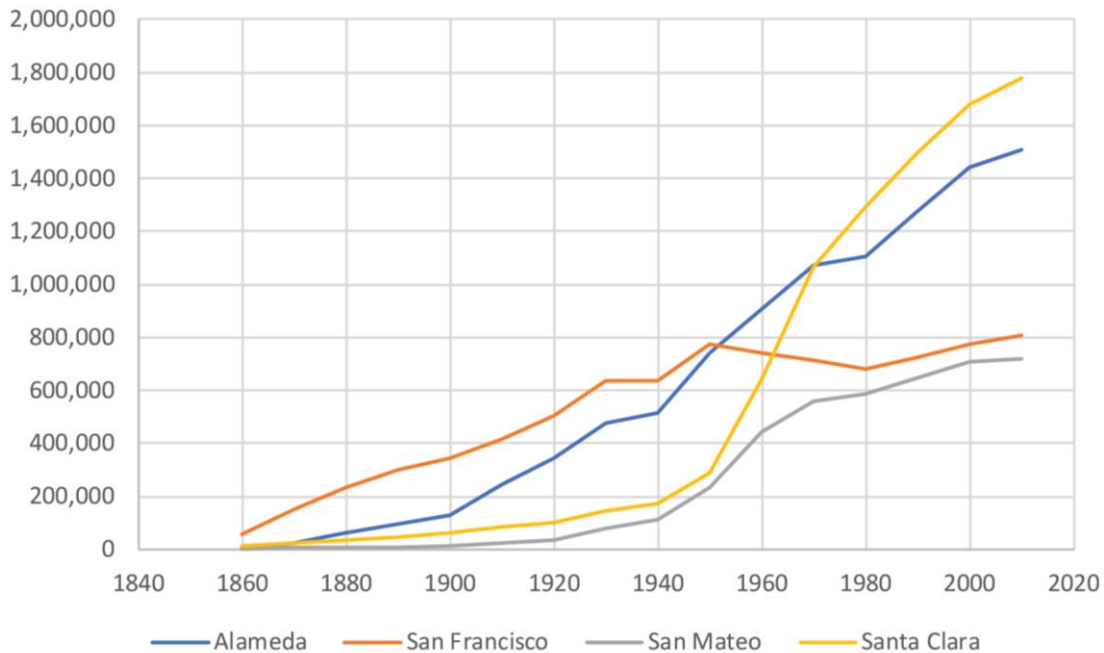


Figure 3.1. Population of the Bay Area’s top four most populated counties from 1860 to 2010

Because of zoning constraints on the supply of housing and nearly continuous employment growth (except for during the 2001 and 2008 recessions), the imbalance between housing supply and job growth has increased over the past seven decades in the Bay Area. For instance, between 1990 and 2020, cities in San Francisco-Oakland-Berkeley and San Jose–Sunnyvale–Santa Clara metropolitan statistical areas (MSAs) (the Bay Area’s two largest MSAs) on average issued permits for nearly 16,000 units annually (US Census Bureau, 2020). During this time, more than 26,000 new non-farm jobs were created on average every year (US Bureau of Labor Statistics, 2020). That’s 0.63 new units for every new job. Within the same timeframe at the national level, 0.83 units have been issued permits for every new job.

Because of job losses following the 2008 economic recession, the Bay Area’s housing-to-job ratio increased to above 0.8 but began dropping again in the past decade with the rapid growth of the technology industry. More than 725,000 new jobs were created between 2010 and 2020, while only about 166,000 new units were built – approximately 0.22 units per new job –

bringing down the housing-to-job ratio to less than 0.72 (ABAG & MTC, 2017a). Housing and employment projections suggest that the housing-to-job imbalance will continue to drop between 2020 and 2030, but at a slower pace (ABAG & MTC, 2017a). For the projected 270,000 new jobs that will appear during this time, about 175,000 units are expected to be built (nearly 0.65 units per job); this will slightly reduce the housing-to-job ratio from 0.72 to 0.71.

ABAG and MTC's longer-term projections (2017a), however, suggest that between 2030 and 2040, more housing units (1.16) will be built for every new job, but this will still be insufficient to remedy the housing shortage created over the past seven decades.

3.2.2. SimP-R: A Real Estate Pricing Simulation Framework

SimP-R is composed of a residential real estate pricing model and a supply-to-demand ratio measurement algorithm (SDRA). The pricing model predicts home prices based on a series of building and location-related attributes, including a disaggregated gravity-based measure of the supply-to-demand ratio. This measure is the ratio of access to housing (i.e., supply) to access to employment locations (as the main generator of the demand for housing). In Chapter 2, we showed that the gravity-based specification of the supply-to-demand ratio is a significant predictor of home prices in San Francisco, and in fact was the most important predictor among all the attributes they examined.

For any given spatial distribution of jobs and housing whether observed or based on hypothetical development scenarios, the SDRA module computes the supply-to-demand ratio for every location in the study area at the desired level of aggregation. That is, for an input development scenario (which changes the number of housing units, jobs, or both at different locations in the influence geography), the SDRA computes the new supply-to-demand ratio

values for every location in the study area. It then passes the new supply-to-demand ratio values to the pricing module, which predicts new home prices under the given scenario.

In addition to the supply-to-demand ratio, the implementation of SimP-R for San Francisco (see Chapter 2) used distance to regional transit, building age, slope, and street tree density as predictors of home prices. We used both linear regression and random forest to train SimP-R's pricing model, based on historical sales prices. The results showed that the random forest model had substantially smaller prediction errors (in terms of the mean absolute percentage error) than did the linear model: 22.2% vs. 16.9% at the housing unit level and 14.6% vs. 5.5% at the census block group level. Despite the substantial difference in accuracy, the overall pattern and direction of simulated price change from the two models was similar in their simulation experimentations. The results from both models confirmed that housing price and supply-to-demand ratio are generally negatively related. We used both model implementations across the study scenarios and in the next section discuss in detail the differences between the results obtained from the two.

3.2.3. Housing and Job Scenarios

We created three types of scenarios and simulated their price effects with SimP-R. We first created two baseline scenarios under which housing and job counts and their spatial distributions change only according to the 2025 and 2040 projections, without any additional increase in residential density around major transit stations that is directed by SB 827. We used ABAG and MTC's *Projection 2040* (2017a), which is the statistical companion to *Plan Bay Area 2040*, a state-mandated integrated Sustainable Communities Strategy for the Bay Area's transportation and land-use planning. The housing and jobs projections in this statistical compendium were produced by the Bay Area's implementation of Regional Economic Models

Inc. (REMI). They use a preliminary range of population projections from the Pitkin-Myers model for the Bay Area as the input for REMI (ABAG & MTC, 2017b).

The other two sets of scenarios we explored in this research increase the housing density within a half-mile radius of major transit stations (i.e., the Bay Area Rapid Transit [BART] and Caltrain stations) based on SB 827's provisions. In one set of these scenarios (five scenarios), we kept the employment and housing counts and distributions in areas beyond the half-mile buffer constant as were in 2016 (the pricing model's training base year), but in the other set (ten scenarios), we changed them according to the 2025 or 2040 projections.

We used a reference case with characteristics close to SB 827's requirements to specify the unit density that each scenario applied to the half-mile radius around major transit stations. We explored a range of different possibilities by applying a factor of the reference unit density (between 0.75 to 1.2) to stations' buffers. We took this reference-based approach instead of using the maximum unit density allowed under SB 827, because SB 827 is a state-level directive and lacks common details related to local zoning such as setbacks and minimum unit size, which directly affect the maximum number of units allowed on a site. While SB 827 overrides local zoning, it focuses mainly on minimizing the constraints on maximum density and building height and leaves land use allocation to local ordinances. Obviously, land use configuration and areas allocated for residential use also directly affect residential density and the number of housing units in the buffer around train stations. Moreover, the full zoned unit density can rarely be realized due to microscale factors such as individual owner and investor decisions, the size and form of parcels, architectural design requirements, and land-use adjacencies.

We used the residential unit density of the half-mile radius around the 16th and Mission Street BART station in San Francisco as the reference density since this area is reasonably

similar to SB 827's specified requirements. The area around this station is composed mostly of three-, four-, and five-story buildings with high land coverage and a mix of uses. SB 827 requires local zoning limits for building heights to be greater than 55 feet (five-story) and 45 feet (four-story) within the quarter- and half-mile radii of major transit stops. Except for the two baseline scenarios for 2025 and 2040, each scenario applies a ratio of reference density (75%, 90%, 100%, 110%, and 120% of reference density) to the half-mile radius around major transit stops.

In computing the reference residential unit density, we excluded parcels with uses unlikely to be converted to residential use, including those containing government buildings, schools, museums, hospitals, parks, theaters, major office buildings, factories, shopping malls, and other public and institutional uses. That is, we computed the reference residential unit density by dividing the number of residential units within the half-mile buffer around the reference station (the 16th and Mission Street BART station) across the area of all parcels within the buffer, except for those with the abovementioned uses. We used the block-level five-year (2011-2016) American Community Survey for housing counts in the reference area and included all blocks with 50% or more of their area within the half-mile buffer around the reference station.

For each scenario, the residential unit counts of CBGs that fully fell within the half-mile buffer around major stations were determined as described above by applying a factor of the reference density. The unit counts of the CBGs completely outside the stations' half-mile buffer were kept as in the base year (2016) for one group of scenarios (five scenarios) and were determined according to the city/township-level's projected rates of change from Projection 2040 (for 2025 and 2040) for the rest (12 scenarios). Finally, for the CBGs partially overlapping with

the half-mile buffer, we accounted for both sources of change in the number of residential units proportional to the ratio of the CBG's area that fell into vs. out of the buffer.

Table 3.1 lists all the scenarios we explored and offers the total and percentage change in the number of residential units and jobs under each scenario.

3.2.4. Measuring Housing Affordability

Affordability is generally measured as a relationship between income and housing cost. In the US, counseling and industry practitioners and local policy decisionmakers most commonly use housing affordability indices calculated and reported by the National Association of Realtors (NAR), Department of Housing and Urban Development (HUD), or National Low-Income Housing Coalition (NLIHC) (Jewkes & Delgadillo, 2010).

We used the NAR's Housing Affordability Index (HAI) specification (National Association of Realtors, n.d.) to measure housing affordability under different scenarios. The HAI measures the ratio of median household income to the qualifying income for median home prices. Qualifying income is calculated by requiring the monthly principal and interest mortgage payment for a median-priced home to be 25% of the qualifying income, assuming a 20% down payment and 30-year fixed rate mortgage, based on the effective mortgage rates reported monthly by the Federal Housing Finance Agency (FHFA). The NAR's approach can be implemented at any unit of geography or market definition so long as the median household income and median home price in that market are known. It also accounts for interest rates.

Unlike the NAR measure, HUD's gauge for housing affordability excludes homeowner households and compares renter households' median gross income to qualifying income for median-priced residences. Qualifying income requires that the total cost of housing (including insurance, utility bills, and HOA fees) be 30% of the median home price (HUD's New Rental

Affordability Index, n.d.). Computing HUD's affordability index requires detailed data on all housing costs on a local scale. The NLIHC's Housing Wage Index (Dolbeare, 1991, as cited in NLIHC, 2019) was not suitable for our analysis because it is rent-based, while SimP-R is price-based. It specifies housing affordability as the minimum hourly wage a household needs (called Housing Wage) to afford the Fair Market Rent (FMR) without spending more than 30% of their income. FMR is determined by HUD to be the standard payment amount for their Housing Choice Voucher program.

We calculated the NAR's housing affordability index based on a 4% fixed interest rate, as reported by the FHFA for December 2016 (the base year-month used in the SimP-R implementation we employed in this research) and a \$103,801 median household income estimate from the 2016 one-year American Community Survey. For each scenario, we estimated the median home price based on the sample housing data set used in training SimP-R, employing unit areas in the dataset and simulated per square foot home prices via SimP-R.

In addition to the housing affordability index, we also reported the percentage of households that could afford median price homes under each scenario. Again, we used the 2016 one-year American Community Survey for household counts in different income brackets.

3.3. Simulation Results and Discussion

Table 3.1 summarizes the simulation results for the three sets of scenarios we explored in this research. The results for the two baseline scenarios (which used ABAG and MTC's housing and jobs projections as the only input, without additional SB 827-related residential density increase in transit-rich areas) suggest that San Francisco's housing prices will continue to grow until 2025, but slowly. These results, however, show that housing prices will drop by 2040, eventually to become lower than 2016 prices (the base year). SimP-R's linear implementation

predicts that San Francisco's housing prices will increase a total of 5% between 2016 and 2025 (an average of 0.55% annually) in response to ABAG and MTC's 2025 projected housing and job counts changes. SimP-R's random forest implementation, however, forecasts a 2.9% increase in housing prices (on average, 0.28% annually) for the same period. This was expected, as for every new job only 0.63 units is projected to be added to the Bay Area's housing inventory between 2016 and 2025, which is below 0.76, the housing-to-job ratio in the base year of 2016. For the 2040 base scenario, the simulation results suggest that San Francisco's housing prices will drop by 10.8% based on the linear model and 7.6% based on the random forest model. According to ABAG and MTC's housing and job projections for 2040, 0.88 units will be added to the Bay Area's housing inventory for every new job, which is higher than the base year rate of 0.76, and on average will increase the supply-to-demand ratio in San Francisco.

Simulation results for the second group of scenarios (which increased residential density within the half-mile radius of the Bay Area's major train stations and kept everything in the other areas constant) suggest that SB 827 could substantially reduce housing prices even if it achieves only a fraction of its target housing supply. Applying only 75% of the reference density (i.e., 33 units per acre) to parcels with the potential for residential development brings the density in the half-mile buffer from 18 to 24 units per acre on average, and increases the Bay Area's housing inventory by about 6% by adding more than 200,000 units. If everything else is kept constant, this increase in housing inventory would result in San Francisco's housing prices dropping 12.9% on average based on the SimP-R random forest model and 31.9% based on the linear model.

As we increased residential density within the half-mile radius buffer around train stations, housing prices were found to drop even further. Simulation results from the SimP-R random forest model show that San Francisco's housing prices would drop 14.4%, 16%, 16.8%,

and 16.9% if we respectively applied 40, 45, 50, and 54 units per acre densities to parcels qualified for residential development within the half-mile radius buffer around train stations. However, for the same density levels, SimP-R's linear model predicts significantly more drastic – and seemingly less realistic – drops in San Francisco's housing prices: 41.6%, 47.9%, 54.2%, and 60.3%, respectively.

We observed a very similar pattern in the third group of scenarios, which combined ABAG and MTC's housing and job projections with SB 827-inspired scenarios for increasing residential density within the half-mile buffer around major train stations. That is, housing prices were found to decrease to a greater extent as greater increases were made to the residential density around train stations. Additionally, with increasing the density level in the input scenarios (i.e., extrapolating further from the observed density in the base year), the linear model predicted more drastic price changes and the difference between the two models' outputs grew. As we increased the residential density around major train stations, the linear model's simulated prices continued to drop at an approximately constant rate; however, with the random forest model, the rate at which housing prices dropped decreased.

As the density level that the scenarios applied to the areas around stations increased and further deviated from the current state (i.e., from what was observed in the base year), we treated the simulation outputs with more caution. In Chapter 2's simulation experimentations, we showed that uncertainty around model outputs increases as input scenarios deviate further from the observed state. The growing difference between the outputs of the two models in more extreme scenarios highlights the limitations of extrapolation and the systematic difference between the two models' behaviors in predicting home prices under scenarios that deviate substantially from the observed range of data. In general, the main problem with extrapolation is

that it builds upon the assumption that the observed relationship between predictors and the response variable will hold beyond the observed range of data, a notion that cannot be validated given the lack of data. In response to changes in the supply-to-demand ratio, the linear model extrapolates housing prices without bound at the same rate as in the observed range of data, which in extreme cases can result in unrealistically small or even negative price predictions. The random forest model's predictions are, however, constrained to the observed range of prices. That is why the average prices predicted by the random forest model stop dropping when the density is increased beyond a certain level.

Considering the mechanics of both models, the combination of the two models' outputs still offers an expected range of change in San Francisco's housing prices even in extreme scenarios, despite the higher level of uncertainty and substantial difference between the two models' outputs. Since random forest predictions are always constrained to the observed range of data, SimP-R's random forest model is likely to underpredict the magnitude of the price drop in scenarios that substantially increase the housing supply-to-demand ratio. Thus, output from SimP-R's random forest model can be interpreted as the lower end of the expected price change range. Conversely, SimP-R's linear model likely overpredicts the magnitude of price drop in extreme scenarios. While the linear model's simulated prices drop at a constant rate in response to increasing the supply-to-demand ratio, in reality it is expected that the decrease in home prices will slow and eventually stop, since housing prices cannot be negative or zero. Thus, outputs from SimP-R's linear model can be interpreted as the upper end of the expected price change range.

Table 3.1. Summary of scenarios and their predicted price effect

Year	Density level around train stations (% of reference density)	Total added jobs	Total added units	Average % price change (linear)	Average % price change (random forest)
2016	75%	0	208123	-31.9%	-12.9%
2016	90%	0	271804	-41.6%	-14.4%
2016	100%	0	314543	-47.9%	-16.0%
2016	110%	0	357530	-54.2%	-16.8%
2016	120%	0	400683	-60.3%	-16.9%
2025 ^a	0%	247851	156680	5.0%	2.9%
2025	75%	247851	358086	-23.9%	-11.7%
2025	90%	247851	421336	-33.0%	-12.9%
2025	100%	247851	463885	-39.0%	-13.8%
2025	110%	247851	506648	-44.9%	-15.1%
2025	120%	247851	549689	-50.8%	-16.3%
2040 ^a	0%	657933	580770	-10.8%	-7.6%
2040	75%	657933	769470	-34.6%	-13.6%
2040	90%	657933	831647	-42.7%	-15.4%
2040	100%	657933	873891	-48.0%	-16.4%
2040	110%	657933	916322	-53.3%	-16.9%
2040	120%	657933	959046	-58.6%	-17.0%

^a baseline scenarios.

Figures 3.2 shows the spatial pattern of price changes (in percentage) in response to 2025 and 2040 scenarios. Under less extreme scenarios (base projections and 75% of reference density), the overall spatial pattern of price changes derived from the two models are similar. Both models (particularly for 2025) show greater changes in southern and western areas, where mostly residential units with lower per square foot prices are located, than the northern and north-eastern part of the city. Spatial patterns derived from the two models began to deviate under the more extreme scenarios, and our results are not conclusive. According to the random forest model, as scenarios' density level increases, the magnitude of price changes increases in

the northern and central CBGs, but it remains mostly unchanged in the southern CBGs. However, the linear model shows the magnitude of change in home prices continues to grow at the same rate in all CBGs in response to increasing scenarios' density level.

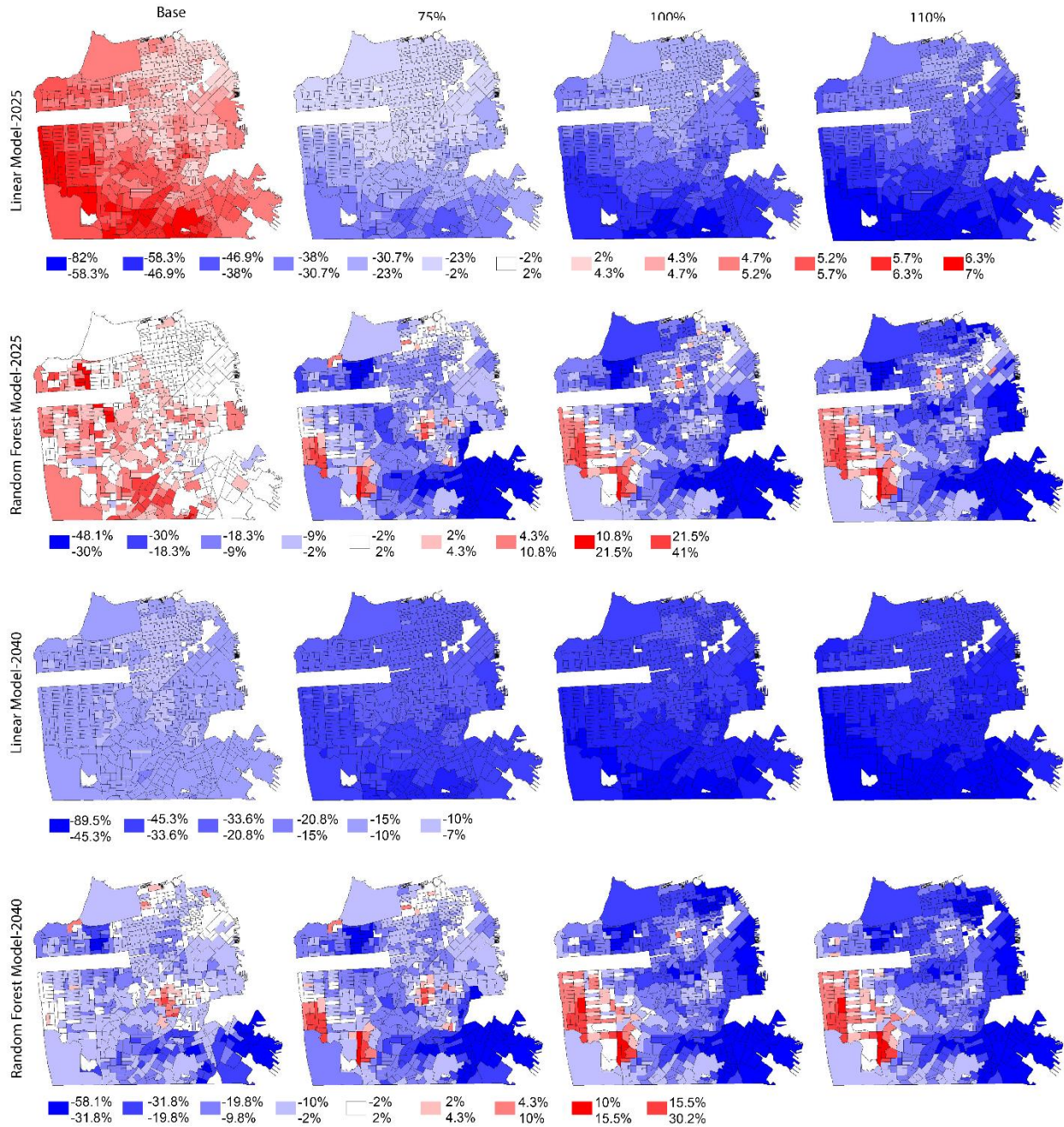


Figure 3.2. Percent price changes in response to scenarios that build upon the Association of Bay Area Governments' 2025 and 2040 housing and job projections

Finally, Table 3.2 reports housing affordability indices calculated for each scenario based on prices simulated by the linear and random forest models. It also reports the percentage of households that could afford median-priced homes in each scenario, based on the two implementations of the pricing model. Despite a substantial drop in home prices and improving housing affordability under the SB 827 scenarios (see Tables 3.1 and 3.2), it appears that San Francisco's housing affordability problem would persist even after increasing residential densities around major train stations to the level required by SB 827. According to the random forest model, under none of the scenarios will a median-priced home become affordable for more than 24% of households. Even based on the linear model (which is expected to overpredict price drops in extreme scenarios), a median-priced home will become affordable for around 50% of households (between 42% to 53% of households) only if we increase the residential density in the half-mile buffer around train stations to a level above SB 827's reference density.

Table 3.2. Housing affordability under different scenarios

Year	Density level around train stations (% of reference density)	Random forest model			Linear model		
		Qualifying income for median-priced home	% HH affording median-priced home	HAI	Qualifying income for median-priced home	% HH affording median-priced home	HAI
2016 ^a	0%	251,557	<24%	0.41	255,144	<24%	0.41
2016	75%	221,459	<24%	0.47	175,177	<36% & >23%	0.59
2016	90%	217,748	<24%	0.48	151,331	<36% & >23%	0.69
2016	100%	212,781	<24%	0.49	135,595	<43% & >35%	0.77
2016	110%	210,540	<24%	0.49	120,116	<53% & >42%	0.87
2016	120%	210,339	<24%	0.49	104,375	<53% & >42%	1.00
2025 ^b	0%	258,300	<24%	0.40	268,013	<24%	0.39
2025	75%	222,830	<24%	0.47	195,013	<36% & >23%	0.53
2025	90%	221,452	<24%	0.47	172,725	<36% & >23%	0.60
2025	100%	219,485	<24%	0.47	157,822	<36% & >23%	0.66
2025	110%	215,281	<24%	0.48	143,070	<43% & >35%	0.73
2025	120%	211,452	<24%	0.49	128,629	<43% & >35%	0.81
2040 ^b	0%	232,019	<24%	0.45	228,206	<24%	0.46
2040	75%	219,589	<24%	0.47	168,005	<36% & >23%	0.62
2040	90%	214,291	<24%	0.49	148,170	<43% & >35%	0.70
2040	100%	211,323	<24%	0.49	134,863	<43% & >35%	0.77
2040	110%	210,229	<24%	0.49	121,970	<53% & >42%	0.85
2040	120%	210,213	<24%	0.49	109,032	<53% & >42%	0.95

Note. HH = household; HAI= housing affordability index. Percent of households affording median-priced home is reported as range due to unavailability of a fully disaggregated household income dataset (households' income data is only available as households count in different income brackets or as summary statistics like median or average).

^a current state. ^b baseline scenarios.

3.4. Conclusion

California Senate Bill 827 (2018), *Planning and zoning: Transit-rich housing bonus* has the potential to substantially decrease San Francisco's housing prices, by overriding local zoning

restriction on medium density residential development within the half-mile buffer around major train stations in the Bay Area. Despite this substantial drop in housing prices, the majority (nearly 65%) of households will not yet afford a median-priced home – that is, they will still need to spend more than 25% of their take-home pay on housing. However, we didn't increase residential density along high-quality bus routes (another area where SB 827 seeks to increase residential density) in our analysis due to the complexities of compiling bus route data (including frequencies) for the entire Bay Area. Including areas both around train stations and along high-quality bus routes could demonstrate a more tangible impact on housing affordability.

Finally, our work also showcases the power of geospatial modeling and simulation to support metropolitan-wide land use policy and planning decisions on a spatially disaggregated level. SB 827 offers a one-size-fits-all solution to the housing crisis across the state of California by applying the same requirements to all transit-rich areas. But models and simulations like SimP-R enable planners and policy makers to explore how different locations of a metropolitan area or city might respond to different spatially detailed land use policies and plans. Spatially disaggregated models also facilitate land use planning coordination among different jurisdictions in a metropolitan area by demonstrating how local land use decisions might impact other areas. With the support of spatially disaggregated models, SB 827 could be fine-tuned to account for local peculiarities and local stakeholders' concerns.

CHAPTER 4: A CONCEPTUAL DATA AND SOFTWARE MODEL

The real estate pricing simulation framework presented in Chapter 2 (SimP-R) enables cities and metropolitan areas to explore the impacts of land use change scenarios on home prices on a spatially disaggregated level without going through the costly and complex process of implementing, calibrating, and running full-fledged ILUTMs. While SimP-R is substantially simpler than ILUTMs, its implementation still requires an advanced level of proficiency in machine learning and data analytics.

In this chapter, I first discuss the challenges and opportunities that accompany implementing the proposed simulation framework in a city or metropolitan area, including compiling the required data, training and validating the pricing model, and reporting and the visualization of outputs. I then present a conceptual model for an opensource software application that will simplify SimP-R's implementation by adding integrated data pipelines, partially automating model training and validation, and creating a guided workflow. The purpose of creating such an application is to make it possible for a typical planning agency's GIS staff with proficiency in processing and compiling geospatial data but without prior knowledge of statistics and machine learning to set up SimP-R for their particular metropolitan area. Such a model will also allow decisionmakers including planners, policy makers, the general public to build and explore scenarios through a simple map interface using this already calibrated simulation for their metropolitan area.

4.1. Challenges and opportunities in implementing SimP-R

Historical sales transactions and unit attributes for the study area

SimP-R's real estate pricing model is trained based on housing units' historical sales prices (as the observed price) and attributes (as predictors). Assessor's offices maintain historical

property sales transactions and information at the county level, such as units' age, area, number of bedrooms, and in some cases roof type and structural and mechanical systems. Property records are public domain and available for anyone to view; moreover, they are often distributed in digital format. While assessor's offices do not necessarily build and maintain infrastructure that allows the general public easy access to large numbers of records (e.g., through application programming interfaces [APIs]), they are required to deliver requested data, and they do so at least through static means. For example, the City and County of San Francisco Office of the Assessor-Recorder delivers requested data on compact discs.

Identifying and purging or fixing erroneous records can be a major challenge. Statistical methods for identifying outliers (either univariate or multivariate) can help with identifying potentially erroneous records, but all identified points must still be examined individually (and often manually), and if needed, compared to other sources such as other records, new surveys, satellite imagery, and exterior images. The overall quality of the data (after removing erroneous records) can then be evaluated by creating small random samples from the data and conducting surveys of sample units.

Disaggregated housing and job counts (in influence geography)

The supply-to-demand ratio is the main predictor in SimP-R's pricing model, making it sensitive to changes in the distribution of housing and jobs. Disaggregated housing and job counts are the main component of the proposed supply-to-demand ratio measure (see Chapter 2). A variety of public and proprietary sources with national coverage are available for obtaining housing and job counts on different spatial levels (e.g., CBG, census block, zip code), and many offer APIs for programmatic access to data and its integration into software applications (e.g., American Community Surveys, and Esri Living Atlas of the World).

While American Community Survey data are public, they are several years lagged, and their estimates usually have high margins of error at spatially granular levels due to the small sample size of surveyed households. Combining decennial census and American Community Surveys with other sources, some commercial alternatives (e.g., Esri) offer more recent estimates with smaller margins of error.

Fully disaggregated housing data (containing individual units and workplaces) can also be obtained by geocoding addresses from opensource address books such as OpenAddresses. Commercial sources such as Esri Business Locations are also available for fully disaggregated jobs data. However, computing the supply-to-demand ratio measure using fully disaggregated housing and jobs data can (computationally speaking) be extremely expensive, if not impractical.

Matrix of travel times between the study area and influence geographies

The proposed measure of the supply-to-demand ratio is gravity-based. That is, the impact of each housing unit and job from the influence geography is presented in the study area according to an exponential decay function of travel time (or distance) between the location of the job or unit and where the supply-to-demand ratio is measured (see Chapter 2, Equation 2). Computing the proposed measure requires distances or travel times between all locations where the supply-to-demand ratio will be measured, and locations of housing units and jobs. Creating this matrix of travel times (or distances) requires a road network with the cost (time or distance) for all segments and a shortest path solver to find the shortest time or distance between all origin-destination pairs in the matrix.

Open Source Routing Machine (OSRM) can be used to compute the matrix of travel times or distances based on OpenStreetMap's road network data for metropolitan areas in the US. Proprietary solutions such as TomTom's Road Network Data and Esri's Network Analyst

are also available for computing this matrix. While its creation can be computationally expensive, it can be pre-computed for different spatial levels (e.g., CBGs, census blocks, zip codes) for every metropolitan area because it doesn't require frequent updates.

Specific predictors for the study area

In addition to unit-related attributes such as age, number of bedrooms, and area, some predictors in the real estate pricing model describe the locations of units (e.g., slope and access to public transportation, retail and other services, etc.). While these attributes can be related to the exact unique location of the unit or the building that contains it, they may be computed for a larger area (e.g., city block) in which the unit is located, due to the resolution of the data or computational complexities.

Moreover, the set of location-related predictors can vary from one study area to another given differences in cities' characteristics and unique drivers of the particular real estate market. Since these predictors are not the same in all cities and metropolitan areas and data sources for these predictors are often local (a single source with national coverage is not usually available), they cannot be offered as integrated services in the application. Data related to these predictors should thus be compiled and passed on to the application as input.

Model selection, training, and validation

The selection, training, and validation of the real estate pricing model are interrelated processes that are performed simultaneously. Regardless of the set of candidate predictors, sample size, and choice of learning technique, price is always expressed as a function of the predictors in SimP-R's real estate pricing model. Since the overall formulation of the model is constant, model selection, training, and validation can be partially automated at least with learning techniques that don't require architecture or hyperparameter optimization. For instance,

linear regression does not have any hyperparameter, and in the case of random forest models, there are widely accepted rules of thumb for setting hyperparameters without optimization.

The processes for model validation and selection (i.e., selecting the set of features that defines the model) generally require a thorough understanding of statistical and machine learning concepts. However, within the limited scope of SimP-R's pricing model, these processes can be partially automated by integrating formal model selection methods (e.g., stepwise forward selection or backward elimination) and k-fold cross-validation, and implementing a graphical interface for communicating validation results and model choices to users in simple, non-technical language.

4.2. A Data and Software Model for SimP-R

In this section, I present a model for a cloud-based software application that will simplify the implementation of SimP-R for metropolitan areas, allowing users to easily create and explore housing- and job-related land use scenarios. This software application will be accessible through standard Internet browsers and is intended for two types of users. The first are typical planning agency GIS specialists who would play a software administration role, setting up SimP-R for their study area using semi-automated processes for model selection, training, and validation. The users will compile geospatial data related to the study area such as historical property sales and units' characteristics. The second groups of users include decisionmakers (including land use planners, and the general public), who could use the map-based graphical interface to create scenarios, run simulations, and explore output.

Figure 4.1 shows a low-fidelity model of the proposed software. The front end works as an intermediate interface between users and back-end data and services. Through this interface, administrative users (i.e., GIS staff) could create new projects (i.e., new implementations of

SimP-R for their metropolitan area) in the back end. Projects would serve as containers for the study area’s data (including unit characteristics and their locations) that administrative users would pass on as input. Through this interface, such users could interact with the model training and validation service, obtaining information on the model’s performance and eventually choosing the final model, which would then be stored in the project space. Through this interface, the administrative user could also set the scale of analysis (e.g., census block group, block, etc.), as well as the extent of the study area and influence geography.

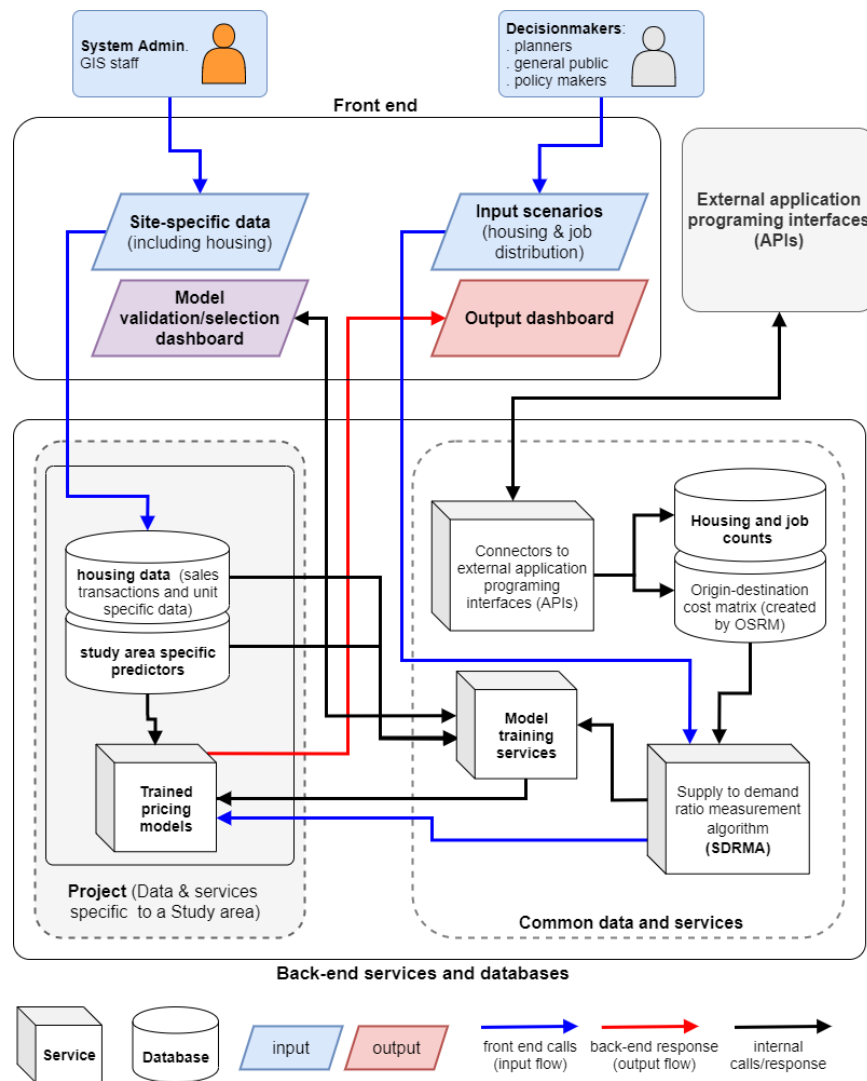


Figure 4.1. A low-fidelity data and software model for SimP-R

Disaggregated housing and job counts and the pre-computed matrix of travel times or distances for all US combined statistical areas will be offered for different levels of analysis as part of the built-in common data and services. The SDRMA back-end service uses the matrix and housing and job counts to compute supply-to-demand ratio values through the training and validation process. Supply-to-demand ratio values cannot be precomputed, as the decay rate for access to housing should be optimized for each metropolitan area.

The front-end interface would also enable the second group of users (i.e., policymakers, planners, and the general public) to create scenarios in a map environment and pass them on to the SDRMA. SDRMA would then compute supply-to-demand ratio values for the given input scenario and pass them on to the already trained pricing model service in its respective project container, which would predict new home prices. Finally, the simulation results would be sent back to the front-end interface and displayed to users.

CHAPTER 5: CONCLUSION

This dissertation presented a simple-to-implement geospatial framework (SimP-R) for simulating the interaction between home price and land use change (here, housing, and job counts) on a spatially disaggregated level, demonstrating its ability to evaluate the effectiveness of land use policies that seek to mitigate the housing affordability crisis. To this end, I implemented the proposed framework for the city of San Francisco, with the Bay Area as the influence geography (see Chapter 2). Using this implementation of SimP-R, I then simulated the price effects of a series of scenarios that represented a range of possible outcomes of land use changes under the California transit housing bill (SB 827) across the San Francisco Bay Area (see Chapter 3). Using the simulation results, I was able to show that an increase in residential density around major train stations as instructed by SB 827 will substantially reduce San Francisco's home prices. Simulation results also showed that despite this substantial drop in home prices, San Francisco's housing affordability crisis will remain largely unresolved given the severity of the problem. Through these simulation experiments, I was also able to capture the spatial variation in the price effect of the land use change scenarios at the census block group level.

The simulation experimentations demonstrated that the proposed simulation framework will enable decisionmakers to evaluate the impact of their land use policies not only on the local housing market, but also on other housing markets in the metropolitan area. The simulation experiments in Chapter 3 showed changing residential density around major train stations across the Bay Area (beyond the city of San Francisco) can significantly affect home prices in the City. This capability will allow local land use planning agencies to explore how their decisions will

affect housing prices in other jurisdictions in the metropolitan area (and vice versa), which will be particularly beneficial for metropolitan-wide intermunicipal coordination of land use policies.

SimP-R's ability to capture the spatial interaction between land use changes in one location and shifts in home prices in other areas within the metropolitan area originates from a novel gravity-based measure of the supply-to-demand ratio I introduced as a predictor in the real estate pricing model (see Chapter 2). Weighting housing units and jobs with an exponential decay function of travel distance (on the road network), our measure specifies the supply effect of each housing unit and demand effect of each job as a continuous surface around them; their effect is at its maximum at the location of the housing unit or job and drops exponentially with distance from them. The pricing model validation showed that the proposed measure of supply-to-demand ratio was a powerful predictor of home prices, both in the random forest and linear implementations of SimP-R's real estate pricing model (and in fact was the most important among all the predictors I examined). The disaggregated representation of supply and demand and their related price effect also comes from the gravity-based approach to specifying supply and demand.

While the gravity-based measure of the supply-to-demand ratio turned out to be a significant predictor of home prices in San Francisco, its performance in predicting housing prices in other cities has yet to be examined. Since in this measure, proximity to jobs is the determinant of demand, the model may not perform the same way in cities where accessibility to jobs is not the main driver of housing price (e.g., cities with large inventories of vacation rental units). This measure also does not account for housing demand by people who work remotely, which might become an important limitation given the growing number of remote workers. This implementation of the supply-to-demand ratio also treats all jobs the same and does not take into

the potential variability in the impact of proximity to high- vs. low-paying jobs on housing prices. Accounting for remote workers' housing demand and variability in wages appear to be an important direction for future improvements in the measure of supply-to-demand ratio given the growing number of remote workers, and the potentially strong relationship between the spatial distribution of wages and home prices. It is also worthwhile to explore including other predictors such as access to parks or schools in the future implementations of the proposed real estate pricing model.

While geospatial simulations can inform land use policies and plans by enabling decisionmakers to explore the impact of alternative land use scenarios, simulation results should be carefully interpreted in scenarios that substantially deviate from the observed (i.e., current) state. Extrapolation inevitably assumes that the observed relationship between the predictors and response variable will hold beyond the observed range of data in the absence of any empirical evidence. I showed that the uncertainty around the simulation output increased (as expected) as the supply-to-demand ratio values deviated further and further from the observed range (see Chapter 2). The increasing difference between the outputs of the two implementations of the pricing model also highlights the limitations of both implementations in scenarios that require extreme extrapolation. However, considering the training and prediction mechanics of both learning methods (i.e., linear and random forest regressions) and comparing the two models' outputs against the theoretically expected response, I concluded that the random forest implementation underpredicted and the linear model overpredicted the magnitude of price changes under extreme scenarios.

The primary purpose of this dissertation was to introduce a real estate pricing simulation framework that would be simple to implement and accessible to a large group of land use

decisionmakers in more cities and metropolitan areas. While the proposed framework is orders of magnitude simpler than any ILUTM, its implementation still requires an advanced level of proficiency in machine learning and data analytics. To make SimP-R more accessible, Chapter 4 presented a conceptual framework for an application with integrated data pipelines and partially automated learning, and discussed limitations on developing such an application. I do not suggest that SimP-R or similar simple single-purpose standalone models can replace ILUTMs. In fact, the housing and job projections I used in Chapter 3 to build the base scenarios were disaggregated using UrbanSim's implementation for the San Francisco Bay Area. ILUTMs are useful (and currently the only solution) for modeling the complex interactions among multiple urban processes (e.g., land use, transportation, real estate market, etc.) while accounting for the iterative feedback from one process to another, as well as macro-level inputs such as national or regional economic growth. However, they are not accessible to many land use planning agencies, given the technical and financial resources that they require. They are also not always suitable for the rapid exploration of many scenarios given their data needs and run time, and developing simple to implement geospatial models like SimP-R can fill this gap.

Finally, the framework presented in this dissertation can be calibrated for simulating changes in rents too. However, since rental data unlike sales transactions are not publicly recorded, compiling a random sample of housing rent data set would take a substantially greater effort. While housing rent and price are often highly correlated, price is speculative, but rent is not. In theory, price of a unit is the sum of its all future discounted cash flow (i.e., rents), and thus driven by speculated future housing demand, but rent at any given time is only related to the actual demand for the physical space housing offers at that time. In fact, since my proposed measure of supply-to-demand ratio does not capture expected future demands, it might be a

better predictor of housing rents than prices. As the housing affordability crisis is affecting renter households more widely than homeowners (50% vs. 23%), calibrating the proposed framework for simulating rents and using it to examine how land use policies like SB 827 might impact rents would be a worthwhile area for future research.

REFERENCES

- Alamo, C. and Uhler B. (2015). *California's High Housing Costs: Causes and Consequences*. Legislative Analyst's Office, Sacramento, CA. Retrieved: <http://www.lao.ca.gov/reports/2015/finance/housing-costs/housing-costs.pdf>
- Alonso, W. (1964). *Location and land use: Toward a general theory of land rent*. Harvard University Press.
- Anselin, L. and Le Gallo, J., 2006. Interpolation of air quality measures in hedonic house price models: spatial aspects. *Spatial Economic Analysis*, 1, 31–52.
- Antipov, E. A., & Pokryshevskaya, E. B. (2012). Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics. *Expert Systems with Applications*, 39(2), 1772-1778.
- Association of Bay Area Governments and Metropolitan Transportation Commission (2017a). *Projections 2040*. Retrieved from: <https://data.bayareametro.gov/Demography/Projections-2040-by-Jurisdiction/grqz-amra>
- Association of Bay Area Governments and Metropolitan Transportation Commission (2017b). *Plan Bay Area 2040: Final Regional Forecast of Jobs, Population and Housing*. Retrieved from: <https://data.bayareametro.gov/Demography/Projections-2040-by-Jurisdiction/grqz-amra>
- Babcock, R. F. (1966). *The zoning game: municipal practices and policies*. The University of Wisconsin Press.
- Basu, S. and Thibodeau, T., 1998. Analysis of spatial autocorrelation in house prices. *The Journal of Real Estate Finance and Economics*, 17, 61–85.
- Beitel, K. (2007). Did Overzealous Activists Destroy Housing Affordability in San Francisco? A Time-Series Test of the Effects of Rezoning on Construction and Home Prices, 1967—1998. *Urban Affairs Review*, 42(5), 741-756.
- Bellotti, A. (2017). Reliable region predictions for automated valuation models. *Annals of Mathematics and Artificial Intelligence*, 81(1-2), 71-84.
- Bonnett, W. (2000). *Build Ships! Wartime Shipbuilding Photographs, San Francisco Bay, 1940-1945*. Windgate Press.
- Bourassa, S., Cantoni, E., and Hoesli, M., 2010. Predicting house prices with spatial dependence: a comparison of alternative methods. *Journal of Real Estate Research*, 32, 139–159.
- Buchanan v. Warley, 245 U.S. 60 (1917)
- Calder, V. (2017). Zoning, Land-Use Planning, and Housing Affordability. Cato Institute Policy Analysis, (823).
- Can, A. (1992). Specification and estimation of hedonic housing price models. *Regional science and urban economics*, 22(3), 453-474.

- Case, B., Clapp, J., Dubin, R., & Rodriguez, M. (2004). Modeling spatial and temporal house price patterns: A comparison of four models. *The Journal of Real Estate Finance and Economics*, 29(2), 167-191.
- Chica-Olmo, J., 2007. Prediction of housing location price by a multivariate spatial method: cokriging. *Journal of Real Estate Research*, 29, 91–114.
- Coppola, P., Ibeas, Á., dell’Olio, L., & Cordera, R. (2013). LUTI model for the metropolitan area of Santander. *Journal of Urban Planning and Development*, 139(3), 153-165.
- Costa, G., Pickup, L., Di Martino, V. (1988). Commuting—a further stress factor for working people: evidence from the European community. *International Archive of Occupational and Environmental Health* (60), pp. 371–376.
- Costonis, J. J. (1972). The Chicago Plan: Incentive Zoning and the Preservation of Urban Landmarks. *Harvard Law Review*, 574-634.
- Curry, B., Morgan, P., & Silver, M. (2002). Neural networks and non-linear statistical methods: an application to the modelling of price–quality relationships. *Computers & Operations Research*, 29(8), 951-969.
- de la Barra, T. (1989). *Integrated land use and transport modelling*. Cambridge, UK: Cambridge University Press.
- Din, A., Hoesli, M., & Bender, A. (2001). Environmental variables and real estate prices. *Urban studies*, 38(11), 1989-2000.
- Dolbeare, C. N. (1991). *Out of reach: Why everyday people can't find affordable housing*. Low Income Housing Information Service.
- Dovey, A. (2018, April 16). *California's Battle Over One Size Fits All Transit-Oriented Zoning*. Next City. Retrieved from: <https://nextcity.org/daily/entry/californias-battle-over-one-size-fits-all-transit-oriented-zoning>
- Dowall, D. E., & Landis, J. D. (1982). Land-Use Controls and Housing Costs: An Examination of San Francisco Bay Area Communities. *Real Estate Economics*, 10(1), 67-93.
- Echenique, M. H., Flowerdew, A. D., Hunt, J. D., Mayo, T. R., Skidmore, I. J., & Simmonds, D. C. (1990). The MEPLAN models of Bilbao, Leeds and Dortmund. *Transport Reviews*, 10, 309–322
- Ellickson, R. (1973). Alternatives to Zoning: Covenants, Nuisance Rules, and Fines as Land Use Controls. *The University of Chicago Law Review*, 40(4), 681-781.
- Elliott, M. (1981). The impact of growth control regulations on housing prices in California. *Real Estate Economics*, 9(2), 115-133.
- Federal Housing Finance Agency (2019). *All-Transactions Indexes*. Retrieved from: <https://www.fhfa.gov/DataTools/Downloads/Pages/House-Price-Index-Datasets.aspx>
- Fischel, W. A. (2004). An economic history of zoning and a cure for its exclusionary effects. *Urban Studies*, 41(2), 317-340.

- Fischel, W. A. (2015). *Zoning rules! the economics of land use regulation*. Lincoln Institute of Land Policy, Cambridge, Massachusetts.
- Fischler, R. (1998) Health, Safety, and the General Welfare: Markets, Politics, and Social Science in Early Land-Use Regulation and Community Design. *Journal of Urban History*, 24, pp. 675–719.
- Frieden, B. (1979). *The Environmental Protection Hustle*. MIT Press. Cambridge, MA.
- Glaeser, E. L., & Gyourko, J. (2003). The impact of building restrictions on housing affordability. *Economic Policy Review*, 9(2).
- Glaeser, E. L., Gyourko, J. & Saks, R. (2005). Why is Manhattan so expensive? Regulation and the rise in housing prices. *The Journal of Law and Economics*, 48.2: 331-369.
- Goodman, A. C. (1978). Hedonic prices, price indices and housing markets. *Journal of urban economics*, 5(4), 471-484.
- Haar, C. M. (1955). In accordance with a comprehensive plan. *Harvard Law Review*, 68(7), 1154-1175.
- Hirt, S. (2014). *Zoned in the USA: The Origins and Implications of American Land Use Regulation*. Cornell University Press.
- HUD's New Rental Affordability Index (n.d.). Retrieved from:
<https://www.huduser.gov/portal/pdredge/pdr-edge-trending-110716.html>
- Hunt, J. D., & Abraham, J. E. (2003). Design and application of the PECAS land-use modelling system. *Paper presented at the 8th International Conference on Computers in Urban Planning and Urban Management*, Sendai, Japan.
- In re Lee Sing, 43 F. 359 (C.C.D. Cal. 1890)
- Isakson, H. R. (1986). The nearest neighbors appraisal technique: an alternative to the adjustment grid methods. *Real Estate Economics*, 14(2), 274-286.
- Jackson, L. E. (2003). The relationship of urban design to human health and condition. *Landscape and urban planning*, 64(4), 191-200.
- Jewkes, M., & Delgadillo, L. (2010). Weaknesses of housing affordability indices used by practitioners. *Journal of Financial Counseling and Planning*, 21(1).
- Johnson, M. S. (1994). *The second gold rush: Oakland and the East Bay in World War II*. University of California Press.
- Karkkainen, B. C. (1994). Zoning: a reply to the critics. *Journal of Land Use & Environmental Law*, 45-89.
- Kamiya, G. (2018, November 23). When WWII Brought Blacks to the East Bay, Whites Fought for Segregation. *San Francisco Chronicle*. Retrieved from:
https://www.sfchronicle.com/chronicle_vault/article/When-WWII-brought-blacks-to-the-East-Bay-whites-13417228.php
- Kauko, T. (2003). On current neural network applications involving spatial modelling of property prices. *Journal of housing and the built environment*, 18(2), 159-181.

- Klosterman, R. E. (1994). Large-scale urban models retrospect and prospect. *Journal of the American Planning Association*, 60(1), 3-6.
- Koschinsky, J., Lozano-Gracia, N., and Piras, G., 2012. The welfare benefit of a home's location: an empirical comparison of spatial and non-spatial model estimates. *Journal of Geographical Systems*, 14, 319–356.
- Kuntz, M., & Helbich, M. (2014). Geostatistical mapping of real estate prices: an empirical comparison of kriging and cokriging. *International Journal of Geographical Information Science*, 28(9), 1904-1921.
- L.A. Tenants Union and Housing Is A Human Right Protest SB 827 (2018, February 13). Housing Is A Human Right. Retrieved from: <https://www.housinghumanright.org/la-tenants-union-housing-human-right-protest-sb-827>
- Lai, W. (1994). The economics of land-use zoning: A literature review and analysis of the work of Coase. *Town planning review*, 65(1), 77.
- Lee Jr, D. B. (1973). Requiem for large-scale models. *Journal of the American Institute of planners*, 39(3), 163-178.
- Leslie, S. W., & Kargon, R. H. (1996). Selling Silicon Valley: Frederick Terman's model for regional advantage. *Business History Review*, 70(4), 435-472.
- Liu, J. G., Zhang, X. L., & Wu, W. P. (2006). Application of fuzzy neural network for real estate prediction. *International Symposium on Neural Networks* (pp. 1187-1191). Springer, Berlin, Heidelberg.
- Luger, M. I. (1991). *Technology in the garden: research parks and regional economic development*. University of North Carolina Press.
- Luxen, D. and Vetter, C. (2011). Real-time routing with OpenStreetMap data. *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. New York, NY.
- Martínez, F. J. (1996) MUSSA: A land use model for Santiago City. *Transportation Research Record*, 1552, pp. 126–134.
- Matlack, J. L., & Vigdor, J. L. (2008). Do rising tides lift all prices? Income inequality and housing affordability. *Journal of Housing Economics*, 17(3), 212-224.
- McMillen, D. P., & McDonald, J. F. (1993). Could zoning have increased land values in Chicago?. *Journal of Urban Economics*, 33(2), 167-188.
- McMillen, D. P., & McDonald, J. F. (1999). Land use before zoning: The case of 1920's Chicago. *Regional Science and Urban Economics*, 29(4), 473-489.
- Mills, E. S. (1967). An aggregative model of resource allocation in a metropolitan area. *The American Economic Review*, 197-210.
- Moeckel, R., Llorca Garcia, C., Moreno Chou, A. T., & Okrah, M. B. (2018). Trends in integrated land use/transport modeling: an evaluation of the state of the art. *Journal of Transport and Land Use*, [S.l.], v. 11, n. 1. Retrieved from: <https://www.jtlu.org/index.php/jtlu/article/view/1205>

- Muth, R. F. (1969). *Cities and Housing; the Spatial Pattern of Urban Residential Land Use*.
- Myers, D., Baer, W. C., & Choi, S. (1996). The Changing Problem of Overcrowded Housing. *APA Journal* 62(1): 66-84.
- National Association of Realtors (n.d.). *Housing Affordability Index*. Retrieved from: <https://www.nar.realtor/research-and-statistics/housing-statistics/housing-affordability-index>
- National Low-Income Housing Coalition (2019). *Out of Reach 2019*. Retrieved from: https://reports.nlihc.org/sites/default/files/oor/OOR_2019.pdf
- Nelson, R.H. (1977) *Zoning and Property Rights: An Analysis of the American System of Land Use Regulation*. MIT Press, Cambridge, MA.
- OpenStreetMap contributors (2018). *California dump, data file (.osm.pbf) from 1 October 2018 of database Geofabrik*. Retrieved from: <https://download.geofabrik.de/north-america/us.html>
- Quigley, J. M., & Raphael, S. (2005). Regulation and the high cost of housing in California. *American Economic Review*, 95(2), 323-328.
- Rosen, S. (1974). Hedonic prices and implicit markets: product differentiation in pure competition. *Journal of Political Economy*, 82, 34–55.
- Salvini, P., & Miller, E. J. (2005). ILUTE: An operational prototype of a comprehensive microsimulation model of urban systems. *Networks and spatial economics*, 5(2), 217-234.
- Schneider, B. (2018, April 18). *YIMBYs Defeated as California's Transit Density Bill Stalls*. CityLab. Retrieved from: <https://www.citylab.com/equity/2018/04/californias-transit-density-bill-stalls/558341/>
- sd11.senate.ca.gov, (April 18, 2018). *Senator Wiener's Bill to Allow More Housing near Public Transportation Stalls in Senate Committee*. Retrieved from: <https://sd11.senate.ca.gov/news/20180417-senator-wiener's-bill-allow-more-housing-near-public-transportation-stalls-senate>
- Selim, H. (2009). Determinants of house prices in Turkey: Hedonic regression versus artificial neural network. *Expert systems with Applications*, 36(2), 2843-2852.
- Sierra Club (2018). *Sierra Club Policy on Transit-Oriented Development*. Oakland, CA. Retrieved from: <https://www.sierraclub.org/press-releases/2018/02/sierra-club-policy-transit-oriented-development>
- Solari, C. D., & Mare, R. D. (2012). Housing crowding effects on children's wellbeing. *Social science research*, 41(2), 464-476.
- Tay, D. P., & Ho, D. K. (1992). Artificial intelligence and the mass appraisal of residential apartments. *Journal of Property Valuation and Investment*, 10(2), 525-540.
- U.S. Bureau of Labor Statistics (2020). *Employment 1990-2020: Metropolitan Statistical Areas*. Retrieved from: <https://beta.bls.gov/dataQuery/search>
- U.S. Census Bureau (2020). *Building Permits Survey*. Retrieved from: <https://www.census.gov/construction/bps/>

- U.S. Census Bureau, (2018). *American community Survey*
- Weiss, M. A. (1987) *The Rise of the Community Builders: The American Real Estate Industry and Urban Land Planning*. New York: Columbia University Press.
- Yang, J. S. (2009). The anti-Chinese cubic air ordinance. *American Journal of Public Health*, 99(3), 440-440.
- Waddell, P. (2002) UrbanSim: Modeling Urban Development for Land Use, Transportation, and Environmental Planning. *Journal of the American Planning Association*, 68:3, 297-314.
- Wang, L., & Waddell, P. (2013, January). A disaggregated real estate demand model with price formation for integrated land use and transportation modeling. *Transportation Research Board Annual Meeting*. Washington, DC. Rederived from:
https://www.researchgate.net/publication/237018513_A_Disaggregated_Real_Estate_Demand_Model_with_Price_Formation_for_Integrated_Land_Use_and_Transportation_Modeling
- Wagner, P., & Wegener, M. (2007). Urban land use, transport and environment models: Experiences with an integrated microscopic approach. *DisP-The Planning Review*, 43(170), 45-56.
- Wegener, M. (2004). Overview of land-use transport models. In David A. Hensher and Kenneth Button (Eds.): *Transport Geography and Spatial Systems. Handbook 5 of the Handbook in Transport*. Pergamon/Elsevier Science, Kidlington, UK, 2004, 127-146.
- Worzala, E., Lenk, M., & Silva, A. (1995). An exploration of neural networks and its application to real estate valuation. *Journal of Real Estate Research*, 10(2), 185-201.
- Zillow Group Research (n.d.). *Zillow Home Value Index (zip code level)*. Retrieved from:
<https://www.zillow.com/research/data/>