Segmented Housing Search*

Monika Piazzesi
Stanford & NBER

Martin Schneider
Stanford & NBER

Johannes Stroebel
New York University

July 2013

Abstract

This paper considers search, trading and valuation in heterogeneous but interconnected housing markets. We use a novel data set on housing search behavior to document stylized facts on buyer search patterns. We then build a quantitative model of the housing market in the San Francisco Bay Area. The model accounts for turnover, inventory and search activity in a large number of market segments that are defined based on the search pattern data. We use the model to infer the distribution of searcher preferences as well as features of the matching technology. We find that buyer interest as well as owner mobility differ substantially, even within narrow geographic areas. As a result, liquidity discounts vary widely across market segments.

1 Introduction

While it is natural to think of housing as a search market, it is not obvious at what level of aggregation market activity should be analyzed. Any individual property is considered by a clientele of searchers with diverse but partially overlapping search ranges. For example, some searchers simply care about affordable houses within some radius of their place of work. Others require a certain minimum size to accommodate their families, or select more narrowly to target certain school districts. The interaction of these different searchers determines turnover, inventory and time on market within housing segments.

This paper considers search, trading and valuation in heterogeneous but interconnected housing markets in the San Francisco Bay Area. We introduce a novel dataset on housing search behavior to document stylized facts on search patterns within the Bay Area. We find

*PRELIMINARY & INCOMPLETE. Addresses: piazzesi@stanford.edu, schneidr@stanford.edu, johannes.stroebel@nyu.edu. We thank Ed Glaeser, Giuseppe Moscarini, Robert Shimer, and Silvana Tenreyro for helpful comments. We are grateful to Trulia for providing data. We also thank seminar participants at the NBER Summer Institute, Conference on Heterogeneous-Agent Models in Macroeconomics in Mannheim, Macroeconomic Dynamics Workshop in London, Northwestern-Tsinghua Conference, Society of Economic Dynamics, Yale University, UCLA, NYU Stern, Chicago Booth and the Minneapolis Fed. This research is supported by a grant from the National Science Foundation.

that search ranges are typically defined by geography, but also by price and home size. When we divide the Bay Area into segments based on observed search patterns, we find substantial heterogeneity in market outcomes, such as time on market, even within zip codes. These segments also differ by clientele; in particular, the typical searcher in a poor urban segment searches more broadly.

We build a model of search with multiple segments that can be quantified with data on search ranges as well as other market activity. We use the model to infer the distribution of searcher preferences as well as the matching technology in the various segments. A key finding is that more expensive neighborhoods are searched by larger clienteles (more searchers per house) who also tend to move less often. This pattern explains why houses in these expensive neighborhoods turn over more slowly but still sell more quickly once they do come on the market. It also implies that houses in more expensive neighborhoods trade at lower liquidity discounts – the present value of both search and transaction costs is lower in expensive neighborhoods.

The real estate website Trulia.com allows searchers to set an alert that triggers an email whenever a house with the searcher's desired characteristics comes on the market. We consider a large sample of such alerts to infer the distribution of search profiles in the Bay Area. Housing search occurs predominantly along three dimensions: geography, price and size, the latter captured by the number of bathrooms. Most searchers look for houses in contiguous areas, though there is significant heterogeneity in the geographic breadth. In particular, searchers who scan expensive urban areas tend to look at substantially more inventory than searchers who look at less expensive or more suburban areas.

To analyze the effect of search behavior on market activity, we divide the Bay Area into a set of 576 distinct segments along the dimensions suggested by the search profiles. We then measure the cross section of turnover, inventory and time on market at the segment level by matching search alert data to deeds records as well as MLS feeds. We find that at least half of the variation in market activity occurs within zip codes, our finest geographic unit. Moreover, the cross sectional means of inventory, turnover, and time on market are all mutually positively correlated, both for our entire sample 2008-2011 and for individual years. In particular, more expensive neighborhoods tend to have lower inventory, turnover and time on market than cheaper neighborhoods.

Our model describes heterogeneous agent types who are subject to moving shocks; they buy and sell houses in a set of market segments. An agent type is identified with a search range - a subset of segments that he searches over in order to find a house. Agents are more likely to match in those segments within their search range where inventory is higher. The steady state equilibrium of the model delivers a mapping from the distribution of preferences into the joint distribution of turnover, inventory, time on market and price by segment, as well as the distribution of buyers by type.

When we quantify the model, each agent type corresponds to one of about 9000 search ranges derived from the alert data. Each search range is a subset of our set of 576 Bay Area segments and may reflect interest along the geography, quality and size dimensions. The distribution of search ranges contains information on how many searchers are interested in

a segment. We define popularity as the weighted number of types that are interested in moving to a segment: a more popular segment has a larger clientele of agents searching over it. Segments also differ in the arrival rate of moving shocks: in a less stable segment, moving shocks occur less frequently.

The instability and popularity of segments can be identified from data on market activity. Indeed, in more popular segments we should see more searchers (and hence more email alerts), higher turnover, lower inventory and lower time on market. Intuitively, a larger potential buyer pool more quickly absorbs any houses that come on the market, thus increasing turnover and leaving less inventory. In less stable segments, houses come on the market at a faster rate, which also leads to more transactions and thus higher turnover. However, when more houses are put up for sale there is higher inventory, leading to higher time on market.

The cross section of market activity depends not only on instability and popularity – both average statistics at the segment level – but on the entire distribution of search ranges. For example, if two market segments are more integrated in the sense that there is a larger share of common buyers, then time on the market should be more similar – common buyers have no reason to wait for inventory in only one of the two markets. The parameters we find thus depend on how interconnected Bay Area segments are in our data.

The above intuition on identification guides our calibration strategy. We select the distribution of preferences to match the model implied distribution of search patterns, together with moments of time on market and volume. The model is tractable enough to infer a high dimensional parameter vector. According to our parameter values, expensive areas tend to be more stable, agents there are less likely to move out. Moreover, expensive urban areas tend to be less popular in the sense that there are at any point in time fewer buyers scanning inventory.

We establish three sets of results. First, we document that the relationship between inventory and buyer interest – the Beveridge curve – is positively sloped within cities and negatively sloped across cities. The negative slope across cities is evidence against the integration of housing markets at the city level. The positive slope within cities is consistent with integration but could also be generated by differences in both instability and popularity of segments within a city.

Second, we use the estimated parameters to infer liquidity discounts for houses in various segments. These discounts are quantitatively large, between 10% and 40% of the frictionless house value (defined as the present discounted value of future housing services by the house.) The liquidity discounts are large in segments that are less stable, where houses turn over more often. They are also large in illiquid segments, where houses take a long time to sell for whatever reason. High turnover and high time on market increase the value of the trading frictions that the current and future buyers face, which amount to the liquidity discount.

Finally, we illustrate the role of search patterns for the transmission of shocks with comparative statics exercises. In particular, we ask how time on market and inventory change if the supply of houses in a segment increases. The answer crucially depends on the number of searchers and what other markets those searchers look at. For example, shocks to

a downtown San Francisco segment with many searchers who search broadly is transmitted widely across the city. In contrast, shocks to a suburban segment close to the San Francisco city boundary has virtually no effect on the market in the city itself.

Related Literature

Our paper provides the first model in which potential buyers search for a house in different segments of the market. Their search patterns may integrate different housing segments and thereby create commonality among these markets. Alternatively, the search patterns may lead to fully segmented housing markets that do not have common features. Our paper contributes to a literature that has investigated the implications of search models for a single market (e.g. Wheaton, 1990; Krainer, 2001; Caplin and Leahy, 2011; Novy-Marx, 2009; Ngai and Tenreyro, 2009; Piazzesi and Schneider, 2009; Burnside et al., 2011; Han and Strange, 2013)[1]

Landvoigt et al. (2012) develop an assignment model to study different housing segments. Their model has implications for the relative volume of various segments, but not for overall volume or the behavior of time on the market. Van Nieuwerburgh and Weill (2010) study the predictions of a dynamic spacial model for the dispersion of wages and house prices across U.S. metropolitan areas. Empirical studies (e.g. Poterba et al., 1991; Bayer et al., 2007; Mian and Sufi, 2009) document the importance of determinants such as credit constraints, demographics, or school quality in different housing markets.

More related to our paper, Genesove and Han (2012) document the number of homes that actual buyers have visited on their house hunt, but without knowing the location or other characteristics of these homes, which are key elements in our work. A number of papers have considered how to divide housing markets into segments (Islam and Asami, 2009, survey the literature). Most of these paper discuss how to split housing markets into mutually exclusive segments based on similarity along a number of characteristics. Goodman and Thibodeau (1998) define housing markets as geographical areas based on a consistent price per unit of housing services. Leishman (2001) argues that housing markets can be segmented both spatially and structurally.

Perhaps the closest paper to ours is by Manning and Petrongolo (2011) who estimate a search and matching model for local labor markets. While the study does not have data on where unemployed workers look for jobs (as we have for home buyers), it uses their home addresses, the addresses of job vacancies in their local area and puts more structure on how workers compare jobs with different commuting times (e.g., workers are indifferent about commuting within some radius.)

[1]Recent models of a single housing market with frictions *other than* search include Piazzesi and Schneider (2012), Floetotto and Stroebel (2012), Favilukis et al. (2010) and Glover et al. (2011). Empirical analyses of frictions in the real estate market include Glaeser and Gyourko (2003), Levitt and Syverson (2008), Garmaise and Moskowitz (2004) and Stroebel (2012).

2 Data Description

To conduct the empirical analysis, we combine a number of key datasets. The first dataset contains the universe of ownership-changing deeds in the Bay Area counties (Alameda, Contra Costa, Marin, San Benito, San Francisco, San Mateo and Santa Clara) between 1994 and 2011. The property to which the deeds relate is uniquely identified via the Assessor Parcel Number (APN). The variables in this dataset that we use for this project include property address, transaction date, transaction price, type of deed (e.g. Intra-Family Transfer Deed, Warranty Deed, Foreclosure Deed), and the type of property (e.g. Apartment, Single-Family Residence). We identify armslength transactions and foreclosure transactions using information on the type of deed and transaction price.

The second dataset contains the universe of tax assessment records in the Bay Area for the year 2009. Properties are again identified via their APN. This dataset includes information on property characteristics such as construction year, owner-occupancy status, lot size, building size, and the number of bedrooms and bathrooms.

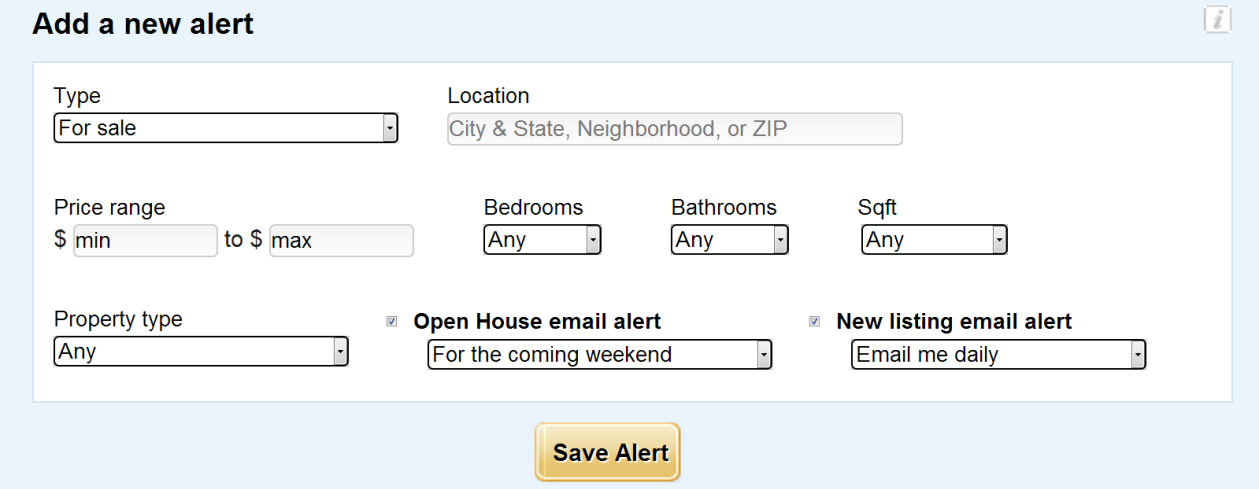
The third dataset includes the universe of all property listings on one of the largest U.S. online real estate search engine Trulia.com (24 million unique monthly users) between October 2005 and December 2011. The key variables in this dataset that we use are the listing date, listing price and the listing address. By using the address of properties to match listings data to deeds data, we can construct a measure of the time on market for each property that eventually sells. In addition, by combining listings data and sales data, we can also construct a measure of the total inventory of houses listed for sale in a particular housing market segment at a particular point in time. The construction of these measures is discussed in more detail in section 4.3.

The final dataset includes information on the search behavior of Trulia.com users. Visitors to Trulia.com can search for houses that are currently listed for sale or have recently been sold by specifying a number of search criteria such as geography, price range, home size and property type. After having refined their search criteria to fit their true preferences, users then have the option to set an email alert, which informs every time a new house with their preferred characteristics comes on the market. In addition, users can directly sign up for an email alert, by imputing information into the search mask shown in Figure 1. We observe a random subset of over 40,525 search alerts set by 24,125 unique Trulia users for the Bay Area counties between March 2006 and April 2012. Of these about 98% request "for sale" alerts, the remaining 2% request "sold" alerts.

3 Email Alerts

In this section we analyze the contents of the search queries in the email alerts set on Trulia.com. This provides unique insights into the search behavior of home buyers in the Bay Area, and is a rare instance in which search behavior is observable for an important market with key search frictions. In later sections we use these email alerts to parametrize

Figure 1: Setting email alerts on Trulia.com



our distribution of buyers across different segments of the housing market.

Every search query contains at least one geographic restriction. Table 1 shows that roughly a third of the queries does not specify any fields in addition to geography. The other fields that are searched by regularly include listing price and the number of bathrooms. Just under a third of queries specifies both price and the number of bathrooms, while another third specifies just a price criterion. The remaining 5% of queries specifies just a bathroom criterion in addition to the geographic restriction.

The other fields in Figure 1 are not often used. For example, only 1.3% of queries specifies square footage while 2.7% of queries specifies the number of bedrooms. We talked to local realtors who confirmed that the number of bathrooms is a commonly used filter to place restrictions on the size of homes. Below, we will thus focus on geography, price and the number of bathrooms as the important dimensions of housing search.

Table 1: Distribution of Alert Parameters

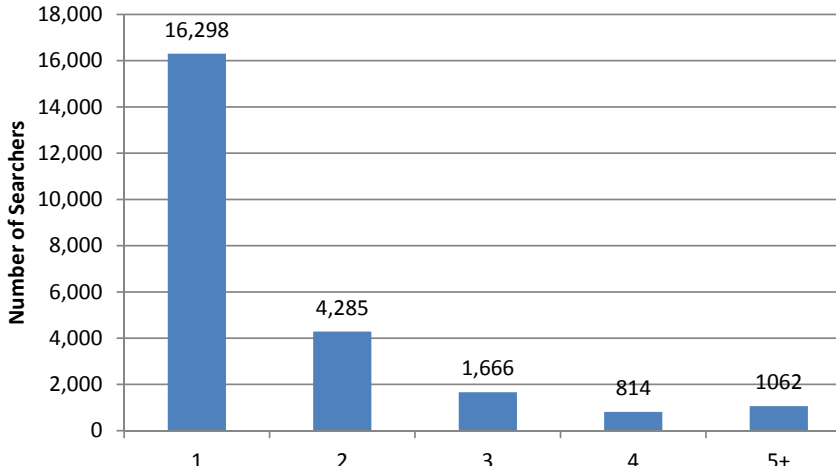
	Price not specified	Price specified	Total
Baths not specified	13,019	13,777	26,796
Baths specified	1,848	11,881	13,729
Total	14,867	25,658	40,525

Note: This figure shows the distribution of which parameters are specified in addition to geography in our query sample.

3.1 Pooling Queries by Searcher

The unit of observation in our search dataset is an email alert, while the unit of economic interest is the entire search range considered by a home buyer. It is possible for each searcher to set more than one separate email alert covering different segments of the housing market, the union of which would define the searcher's total search set. In our email alert dataset, a searcher is identified by a unique (scrambled) email address to which the alerts will be sent. There are a total of 24,125 unique individuals setting search queries; this means that the average searcher sets about 2 email alerts. Almost 70% of individuals only set one search alert, and more than 90% of individuals set 3 or fewer alerts. Figure 2 shows a histogram with the number of search queries by individual.

Figure 2: Histogram of the number of email alerts by searchers



In some the following descriptive analysis of search behavior, we pool across all queries set by the same searcher. For the geography dimension, which is discussed in section 3.2, this pooling is straightforward: we first determine which geographies are covered by each query, and then consider a searcher's geographic range to be the union of all geographies covered by at least one query of the searcher. For the analysis of the joint density of geography and price dimension discussed in section 3.8 such pooling is more complicated. This is because a searcher might search for houses in Palo Alto between $500,000 and $1 million in one query, and in Mountain View between $400,000 and $700,00 in another. For the descriptive analysis of searcher behavior, we will take the widest possible price range considered in the queries. In the example above, we assume the searcher is interested in houses in Palo Alto and Mountain View between $400,000 and $1 million. In section 4.2 we define housing market segments that differ along a geography, price and home size dimension. These segments are used for the subsequent calibration of the model. With these multi-dimensional segments the pooling across different queries by the same searcher is much easier: we first determine which segments are covered by each query and then pool all segments covered by any search alert by a particular searcher to be part of that searchers search set.

3.2 Geography

Each alert defines the desired search geography by selecting one or more city (e.g. Palo Alto, Menlo Park, San Jose), zip code or neighborhood (e.g. the Mission district in San Francisco). About 61% of alerts define the finest geographic dimension in terms of cities, 18% in terms of zip codes and the remaining 21% of alerts specify the finest geographic dimension in terms of neighborhoods. Many queries may include geographies in terms of cities, zip codes and neighborhoods in the same query. Figure 15 in Appendix A.1 shows a number of cross tabulations of the geographic dimension of the alerts. For example, a total of 1,108 alerts specify exactly one zip code and one city. 1,624 alerts specify more than 4 cities, while 2,406 alerts specify more than 4 neighborhoods.

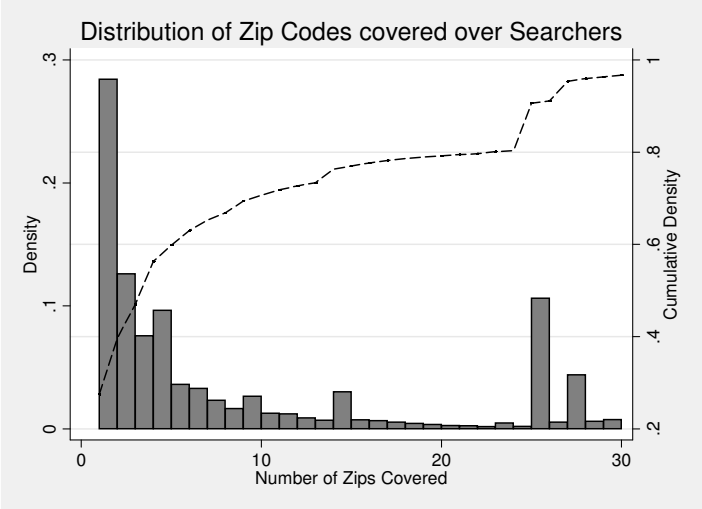
In order to further analyze the geographic dimensions of our queries, we consider a zip code as our common unit of geography. To implement this, we need to deal with alerts that specify geography at a unit that might not perfectly overlap with zip codes. For alerts that select listings at the city level, we assign all zip codes that are at least partially within the range of the city to be covered by the search query (i.e. for a searcher who is looking in Mountain View, we assign the query to cover the zip codes 94040, 94041 and 94043). Neighborhoods and zip codes also do not line up perfectly, and so for each neighborhood we again consider all zip codes that are at least partially within the neighborhood to be covered by the search query (i.e. for a searcher who is looking in San Francisco's Mission District, we assign the query to cover zip codes 94103 and 94110). This provides us, for each search alert, with a list of zip codes that are covered by that search alert. Using this method of assigning queries to zip codes, the search alerts in our dataset cover a total of 233 unique Bay Area zip codes. As discussed in section 3.1, we then pool across all queries set by the same searcher. That is, we add all zip codes that are considered by at least one query of the searcher to that person's search set.

The frequency and cumulative frequency of number of zips selected by each searcher are shown in Figure 3. 6,631 searchers (about 28%) consider exactly one zip code, 2,943 searchers consider exactly 2 zip codes. About 71% of searchers consider 10 zip codes or less. In total, 2,478 searchers consider exactly 25 zip codes (these are alerts that only select San Francisco as a city) and 1,026 searchers consider exactly 27 zip codes (these are alerts that only select San Jose as a city). The maximum number of zip codes considered by a single user is 199, arising from a searcher that selects 78 separate cities. The average number of zip codes covered by a searcher is 9.24.

3.3 Geography Patterns - Distance

We next analyze patterns amongst zip codes that are jointly selected by agents within a a search alert for those 72% of searchers that select more than one zip code. There are a number of ways to analyze these queries. In a first analysis, we consider distances between the zip codes selected by the same searcher. To do this, we begin by determining for each

Figure 3: Distribution of Geography Specifications



Note: This figure shows the frequency and cumulative distribution of the number of Bay Area zip codes covered by the individuals setting search alerts in our datasets.

zip code the geographic and population-weighted zip code centroids.[2] Amongst each set of zip codes selected by a particular searcher, we compute the maximum and mean distance between these zip code centroids. We consider 3 measures of distance: (i) the direct "as the crow flies" geographic distance, (ii) the travel time by car between the zip code centroids and (iii) the commuting time by public transport.[3] Table 2 shows the distribution of these measures across the 17,488 searchers that select more than one zip code.

The maximum distance between two geographic zip code centroids selected by the same search query is 104 miles. This distance is generated from a searcher that select both Petaluma and Los Gatos as cities (the bottom left panel of Figure 18 shows a map of all zip codes selected by one of those searchers). The searcher with the smallest maximum distance between two zip code centroids (just over half a mile) selects the zip codes 94111 and 94133, two zip codes in San Francisco's Financial District (this geographic dimension was actually selected by three different searchers). On average amongst searchers that select more than one zip code, the maximum distance between two geographic centroids among the zip codes selected is 9.8 miles. When considering the average distance between all zip code centroids selected, this ranges from a minimum of 0.6 miles to a maximum of 72.9 miles across

[2]The geographic centroid of each zip code is provided by the U.S. census bureau. To determine the population-weighted centroid of each zip code, we compute the average latitude and longitude of all census blocks within that zip code, weighted by the population of each census block.

[3]Measures (ii) and (iii) are calculated using Google Maps, with an adjusted version of the traveltime.ado file. Commuting times were computed for travel at 8am on Wednesday, March 20, 2013. A few zip code centroids are inaccessible by public transport as calculated by Google. Public transport distances to those zip code centroids were replaced by the 99th percentile of travel times between all zip code centroids for which this was computable. This captures that these zip codes are not well connected to the public transport network.

Table 2: Distribution of Distances across Search Alert Zip Codes

	Geographic Zip Code Centroids					
	Min	Bottom Decile	Median	Top Decile	Max	Mean
Max Geographic Distance	0.6	2.6	6.7	21.2	104.4	9.8
Mean Geographic Distance	0.6	1.8	3.3	9.0	72.9	4.8
Max Car Travel Time	4.0	10.5	21.0	38.5	147.5	23.3
Max Public Transport Time	12.0	40.0	80.5	354.5	558.0	152.8
Mean Public Transport Time	12.0	28.8	50.0	148.8	488.8	78.3
	Population-Weighted Zip Code Centroids					
	Min	Bottom Decile	Median	Top Decile	Max	Mean
Max Geographic Distance	0.5	2.3	6.8	21.1	103.3	9.7
Mean Geographic Distance	0.5	1.8	3.2	8.9	74.0	4.7
Max Car Travel Time	4.0	9.5	20.5	38.5	143.5	22.8
Max Public Transport Time	3.8	8.9	13.1	19.7	132.5	14.0
Mean Public Transport Time	10.5	40.5	79.0	375.0	573.5	140.1

Note: This figure shows the summary statistics across searchers who select more than one zip code (N = 17,488) of travel time and geographic distances between the centroids (geographic and population-weighted) of all zip codes selected by that query. Travel times are measured in minutes. Geographic distances are measured in miles.

searchers. Similar numbers obtain when considering distances between population-weighted zip code centroids rather than between geographic zip code centroids.

Travel times, either by car or by public transport, allow us to measure commuting distances between zip code centroids selected by the same searchers (below we will also determine the "center" for each set of zip codes selected, to consider where people might want to travel to). The average across searchers of the maximum travel time by car between any zip code centroid selected by the same searcher is about 23 minutes. The 10th percentile is 11 minutes, the 90th percentile is 39 minutes. For travel by public transport, the average maximum time between query zip codes is 2 hours and 32 minutes; the 10th percentile is 40 minutes, the 90th percentile is almost six hours. These results suggest that while, on average, zip codes selected by the same searcher are reasonably close together geographically and in terms of commuting, there is significant heterogeneity in the geographic breadth covered by different email alerts. This will have important implications when calibrating our search model.

3.4 Geographic Patterns - Contiguity

A second aspect of the geographic dimension selected by home searchers is whether or not they are contiguous, that is, whether it is possible to drive from each zip code in the search

set to all other zip codes without ever leaving a zip code in the search set. As described in Appendix A.2, we allow travel across one of the six bridges across the San Francisco Bay to provide contiguity between two zip codes either side of the bridge. We then analyze whether or not the zip codes selected by a particular searcher are part of a contiguous set of zip codes. Figures 17 and 18 in the appendix shows examples of contiguous and non-contiguous search sets selected by Bay Area home buyers. To analyze the contiguity of zip codes selected by the same searcher more systematically we again focus on those home buyers that search in more than one zip code. Table 3 shows summary statistics for contiguity measures for searchers that select a certain number of zip codes. In total, 82% of all searchers select contiguous geographies. The maximum number of distinct contiguous segments selected by an individual searcher is 10, while on average a query selects 1.24 distinct contiguous segments. Out of searchers that select only two zip codes, only 9% select two non-adjacent zip codes. The increase in the share of contiguous queries for the group with 21-30 zip codes selected can be explained by the prevalence of searches for "San Francisco" and "San Jose" in that category.

Table 3: Contiguity Analysis – Summary Statistics

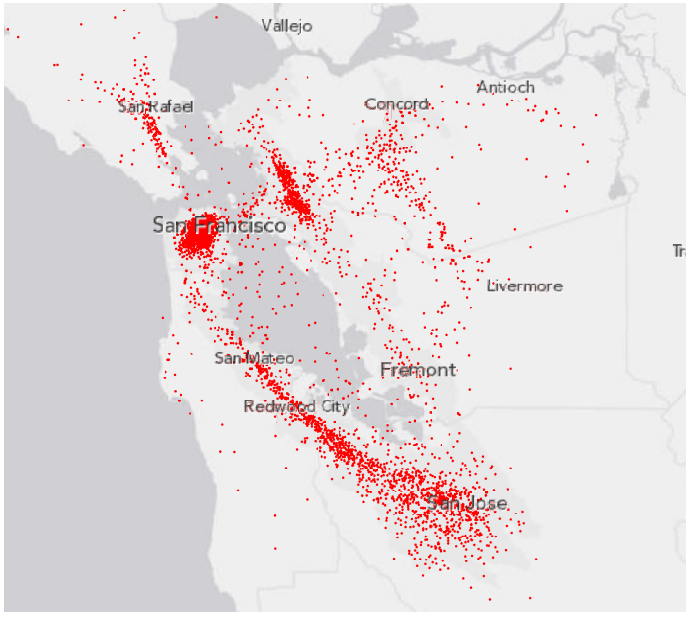
		Contiguous Segments		
Number of Zips Covered	Share contiguous	Mean	Max	Total Number
2	91%	1.09	2	2,927
3	83%	1.18	3	1,761
4	91%	1.10	3	2,248
5	67%	1.37	4	844
6-10	71%	1.38	5	2,612
11-20	74%	1.38	8	2,071
21-30	91%	1.13	10	4,213
30+	48%	1.94	9	798
Total	82%	1.24	10	17,474

Note: This figure shows summary statistics for contiguity measures across queries that select different number of zip codes.

3.5 Geographic Patterns - Circularity

As a third way of evaluating the geographic dimension of housing search patterns we consider to what degree the searcher pursues a strategy of selecting a central point (such as the location of a job, or a good school), and then searches in a circular fashion around that central spot. We proceed in a number of steps. First, for each searcher we find the geographic center of his search set - this takes the average longitude and the average latitude of all (geographic) zip code centroids selected by that searcher. Figure 4 plots these centers. Most of the search set centroids focus either on downtown San Francisco, Silicon Valley, San Jose and Berkeley.

Figure 4: Location - Query Center



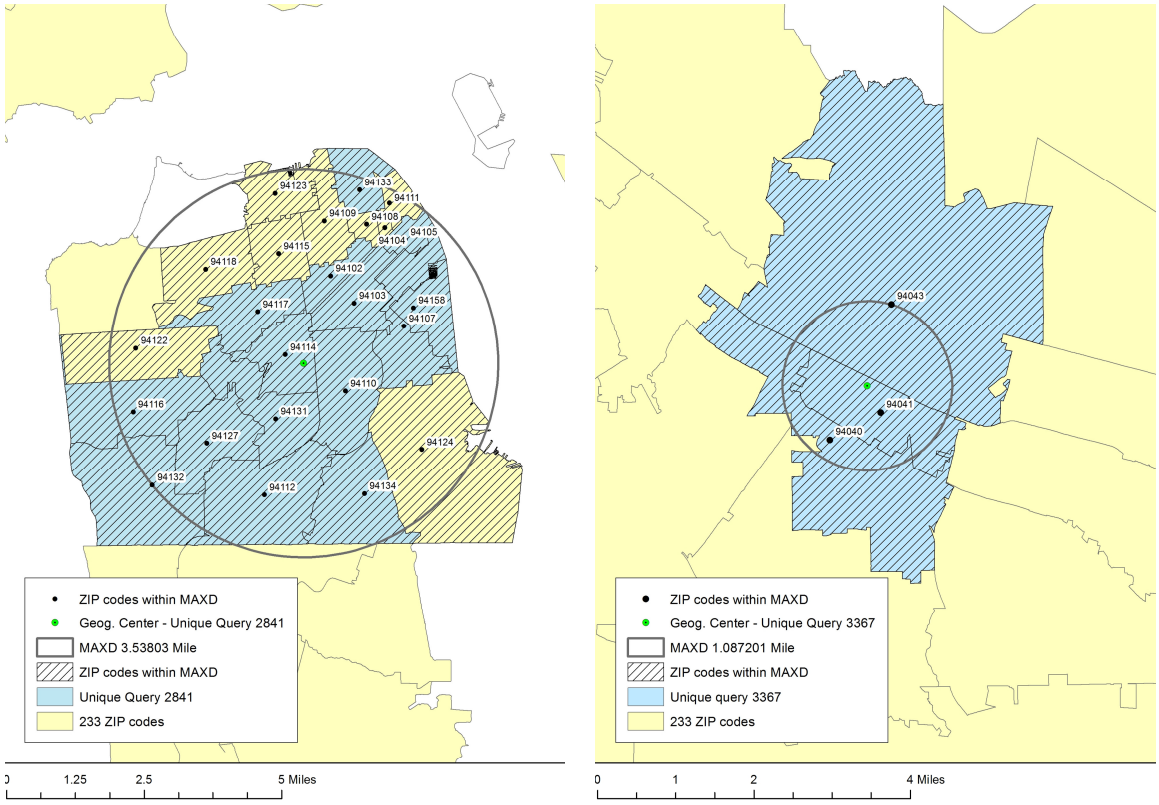
Note: This figure shows the geographic centers of all search alerts that select more than 1 zip code.

We next determine the maximum distance from a zip code centroid in the search set to this center. On average, this maximum distance is 3.95 miles. The 10th percentile is 1.31 miles and the 90th percentile is 12.78 miles. We then consider all zip codes in the Bay Area to see which ones are at least as close as the furthest zip code in the search set to the search set center. We next check how many of those zip codes are actually part of the search set. Figure 5 illustrates this procedure.

We then analyze the degree of "circularity" of the geographic dimension more systematically. Again, we focus on the subset of searchers who select at least two zip codes. On average, 47% of all searchers cover every single zip code centroid that falls within the maximum distance between the geographic search set center and the furthest zip code centroid. Unsurprisingly, this figure is highest, at 83%, for queries that only cover two zip codes, and declines for queries that cover more zip codes. In addition, for search sets with a larger maximum distance between the furthest zip code covered and the search set center, the proportion of searches that cover all zip codes within this maximum distance from the center declines. The searcher covers 78% of all zip codes centroids that are as close or closer to the search set centroid as the farthest covered zip code centroid. Even for the relatively small set of non-contiguous queries, this figure is about 33%.

A key conclusion from the analysis of the geographic dimension of email alerts is that most home searchers in the for the Bay Area search across zip codes with similar accessibility to a particular location (which usually corresponds with to job centers), creating a contiguous set of zip codes considered. However, there is significant and important heterogeneity in how geographically broad the searches are: While about 28% of searchers only consider a single

Figure 5: Explanation of Circularity Test



Note: This figure shows two examples of the circularity analysis. All zip codes that are part of the search set are shown in blue. The geographic center of each search set is given in green. The circle is centered around this geographic center and has radius equal to the furthest distance of any zip code centroid in the search set to the search set center. All zip codes whose center lies within the circle (and who are thus at least as close as the furthest zip code center in the search set) are shaded.

zip code, the 10% most broad search sets cover zip codes that are more than 20 miles apart and can take several hours to travel between by public transport. In addition, there is a sizable number of searchers that select non-contiguous zip codes, sometimes reasonably far apart. Searchers setting such broad queries are going to play an important role in providing a force that equalizes time on market between different Bay Area geographies.

3.6 Price Ranges

The second key dimension regularly selected by home searchers is the listing price. Restricting the price dimension allows households to select homes of differential quality within the same geographic segment. As we discuss below, homes of different qualities (which are listed at different prices) are likely to be part of different housing market segments. At the searcher level about 63% of all searchers specify a price dimension in at least one of their queries (Table 1 showed that about 61% of all search alerts specified a price dimension).

When selecting a price, 7,856 (51%) of searchers specify both an upper and a lower bound of a price range, 7,166 (46.5%) specify only an upper bound, while 386 (2.5%) only select a lower bound. The top panel of Figure 6 shows the distribution of minimum and maximum prices selected in the email alerts. Notice that most minimum and maximum prices are set at multiples of $\$50,000$, with particularly pronounced peaks at multiples of $\$100,000$.

In addition to the heterogeneity across search alerts in the breadth of geography selected, there also appears to be significant heterogeneity in the breadth of the price ranges selected by different searchers. Amongst those individuals that set both an upper and a lower bound, the 10th percentile of queries selects a price range of $100,000, the median a price range of $300,000 and the 90th percentile a price range of $1.13 million. The bottom left panel of Figure 6 shows the distribution of price ranges, both for those agents that select an upper and a lower bound, as well as for those agents that only select an upper bound. In the bottom right panel we also analyze how the breadth of the price range varies with the price segment considered. We bin the price range midprice into 10 groups. The price range considered increases monotonically with the midpoint of the price range. We also consider the hypothesis whether people set the price ranges by considering the price they are willing to pay, and search that price plus/minus a fixed percentage. To do this, we find the ratio of the price range to the mid-point of the price range. This is shown in the bar chart of the bottom right panel of Figure 6.

3.7 Number of Bathrooms

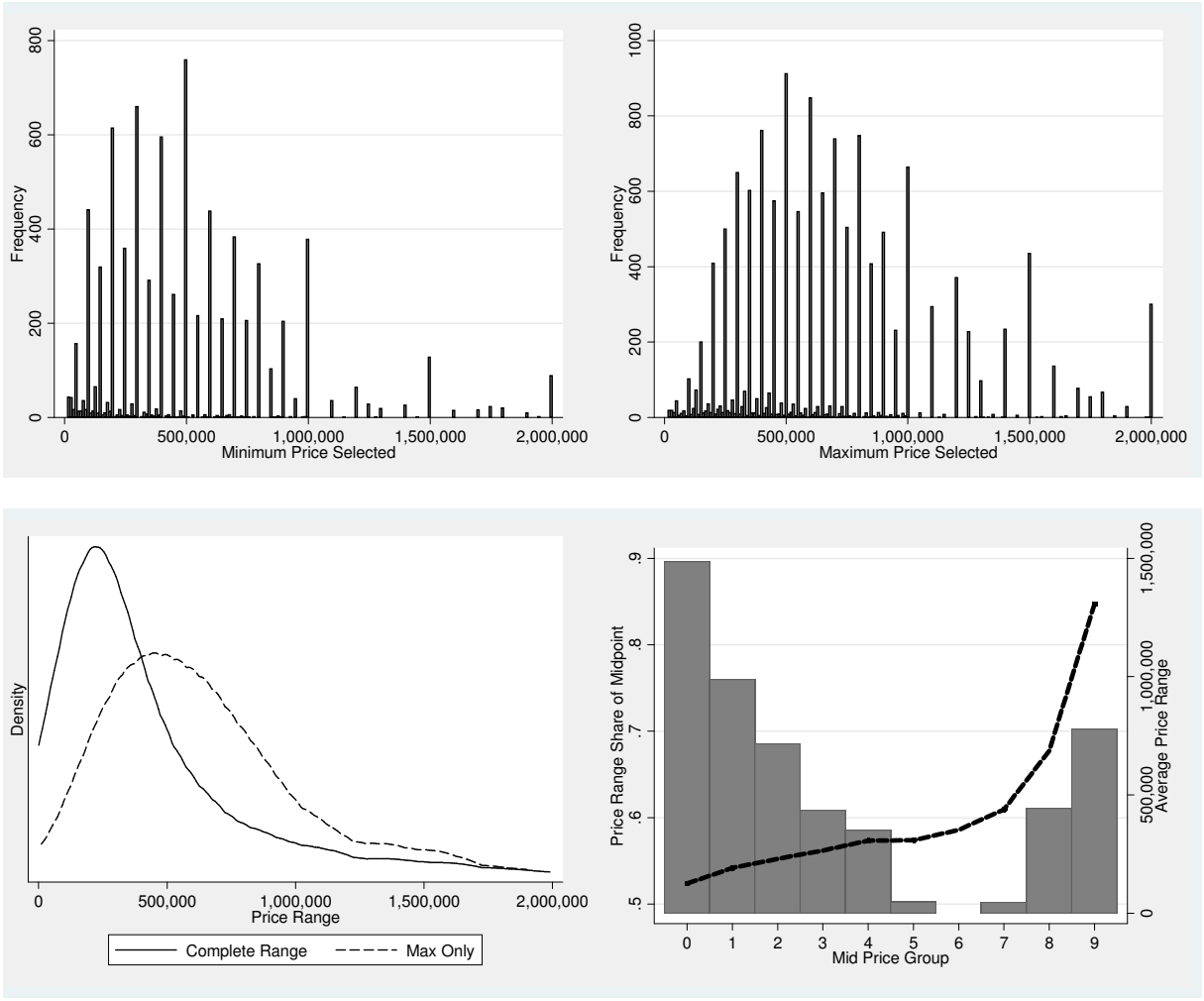
As described above, the third dimension that is regularly populated in the email alerts is a constraint on the number of bathrooms. This appears to be a home searcher's most popular way of selecting homes of a certain size. Figure 7 shows the distribution of bathroom cutoffs selected for the Bay Area. 68% of all bathroom limits are set a value of 2, most of them as a lower bound. This setting primarily excludes 1 and 2 bedroom apartments and very small houses.

3.8 Interaction of Geography, Price Range and Bathrooms

In the previous sections we showed that while there are clear patterns in the way that people search for homes across the geographic and home quality and size dimension, there was significant heterogeneity in how broad the searches were on the price and geography dimension. In this section we analyze the interaction of the breadth of the different search dimensions. The results are show in Table 4.

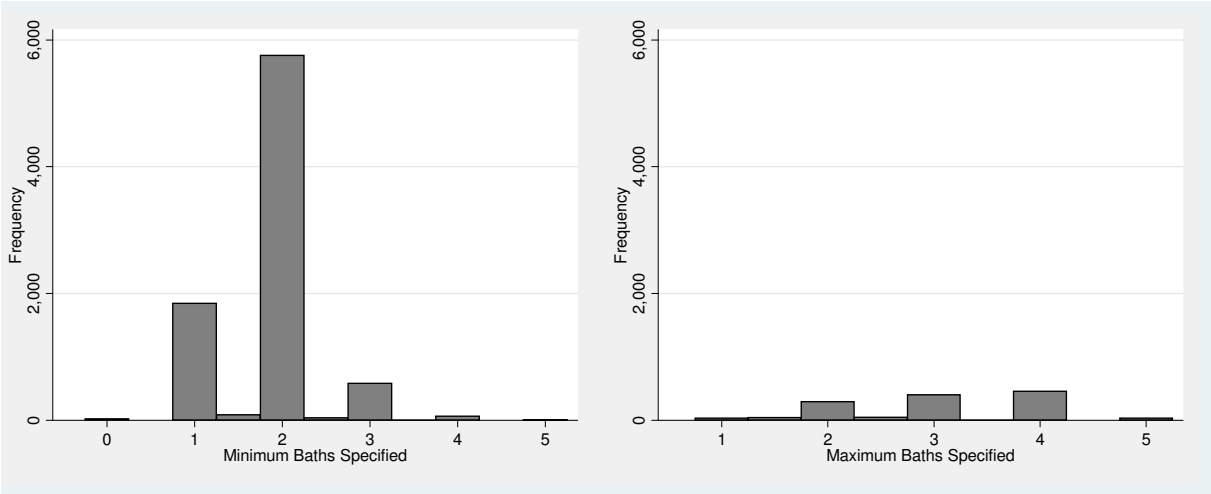
Those search searchers that specify a price restriction cover an average of 10.3 zip codes, while searchers that do not specify a price restriction only cover 7.3 alerts on average. Amongst those search sets that cover more than one zip code, the average maximum geographic distance between zip code centroids is 7.9 miles for those that do not set a price restriction, and 1.06 miles for those that do. The searches with price restrictions also have bigger maximum commuting times. This suggests that searchers trade off between being

Figure 6: Price Cutoff Analysis



Note: This figure shows a histogram in steps of $10,000 of the minimum and maximum listing price parameters selected by home searchers in their email alerts. The bottom left panel of this figure shows the distribution of price ranges across queries both for queries that only select a price upper bound as well as for those queries that select an upper bound and a lower bound. The bottom right panel shows statistics only for those alerts that select an upper and a lower bound. The line chart shows the average price range by for different groups of mid prices, the bar chart shows the average of the price range as a share of the mid price.

Figure 7: Bathrooms Cutoffs Selected



Note: This figure shows a histogram in steps of 0.5 of the minimum and maximum bathroom parameters selected by home searchers in their email alerts.

Table 4: Geography, Price and Bath Parameter Interaction

	No Price		Price		No Bath		Bath	
	Mean	N	Mean	N	Mean	N	Mean	N
# Zips Covered	7.3	8,725	10.3	15,400	8.8	15,716	10.0	8,409
Max Dist. (Mil)	7.9	5,375	10.6	12,113	8.9	10,899	11.1	6,589
Max Car (Min)	20.8	5,375	24.5	12,113	22.5	10,899	24.7	6,589
Max Public Trans. (Min)	75.8	5,375	92.4	12,113	82.1	10,899	95.9	6,589
Is Continguous	54%	8,725	62%	15,400	59%	15,716	60%	8,409

Note: This Table shows summary statistics across queries that cross-tabulate moments across different search parameters.

more selective on the geography that they are willing to consider and being more selective on the price they are willing to pay. A similar tradeoff can be detected between being more specific on home size (selecting a certain number of bathrooms) and being more selective on the geography considered (i.e. selecting fewer zip codes).

3.9 Representativeness of Queries

There are a number of reasons why we believe that Trulia search queries do indeed provide a reasonable approximation of overall housing market search behavior. Firstly, the internet has become the most important tool in the home buying process, with over 90% of homebuyers using the internet in their home search process (National Association of Realtors, 2013). For 35% of home buyers, looking online is the first step taken in the home purchase process. The

fraction of people who found the home they end up purchasing on the internet is the same as the fraction who found that home through real estate agents. This is not just within the younger age group: 86% of home buyers between the ages of 45 and 65 go online to search for a home. The median age of homebuyers using the internet is 42, the median income is $83,700 (National Association of Realtors, 2011). This is only slightly younger than the median of all home buyers (which is 45) and slightly wealthier (the median income of all home buyers was $80,900). In addition to showing that online real estate search is almost universal, this suggests that the online real estate audience is generally rather representative of the overall population of home buyers. Trulia, with approximately 24 million unique monthly visitors (71% of whom report to plan to purchase in the next 6 months), has similar demographics to those of the overall online home search audience (Trulia, 2013).

3.10 Stability of Search Patterns

For the purpose of calibrating our search model, we interpret the observed search sets as representative of the set of houses considered by that particular searcher. An important question (for example, for the interpretation of comparative statics exercises based on the model) is whether these search sets are invariant to changes in market conditions. For example, one might worry that during periods with high market activity with many homes listed, searchers narrow the range of houses that they consider to a subset of their previous search set. To test whether this is the case, we analyze whether average search parameters change alongside the strong seasonality of the housing market. Figure 3.10 shows that both volume and search activity exhibit seasonal movements, with more activity during the summer months.

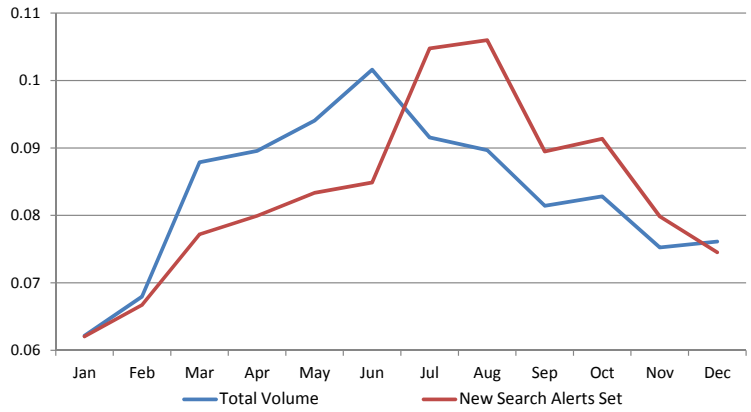
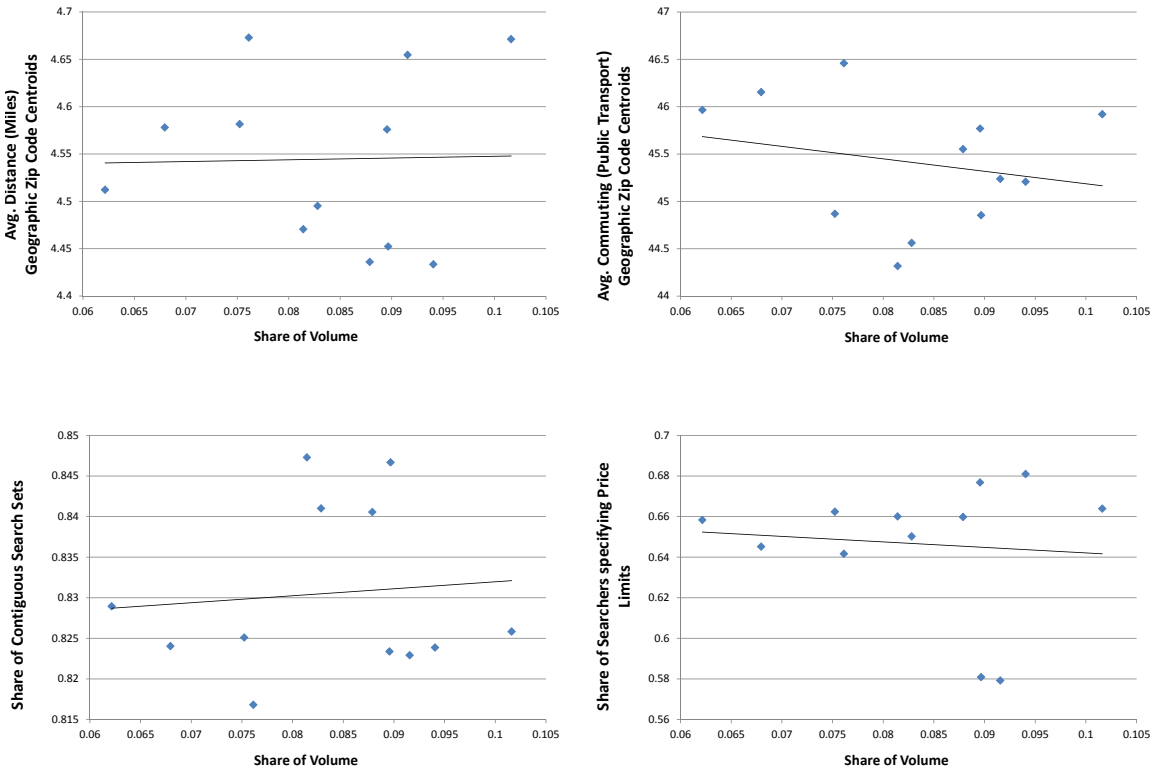


Figure 8: Seasonality of the Housing Market – share of annual total transactions and annual total search listings set in each month of the year.

To test whether search parameters adjust to this market activity, we test whether the geo-

graphic range considered during the summer months is narrower than in the winter months. In each panel of Figure 9 we show a scatter-plot between the share of total volume in a month, and a particular search dimension. In particular, we include (clockwise from top-left) the average distance between geographic zip code centroids, the average commuting time by public transport between geographic zip code centroids, the share of searches that yield contiguous search sets and the share of searches that include a price dimension. We can see that none of these dimensions exhibit any meaningful seasonality, which makes us confident in our interpretation of search parameters ares time-invariant preference sets.

Figure 9: Non-Seasonality of Search Parameters



Note: This figure shows the correlation of search parameters with the share of annual volume in a particular month on the horizonal axis.

4 Model setup

The model describes a small open economy, such as the San Francisco Bay Area. Time is continuous and the horizon is infinite. Agents live forever and discount the future using the riskless rate r.

Segments, search ranges and clienteles

Let H denote a finite set of market segments. The data counterpart of a model segment is a homogenous set of houses defined by location, size and quality (for example, all houses in Berkeley with 2 or more bathrooms that cost less than \$400K). The details of the construction of these segments are described in section 4.2. The measure μ^H on H counts the number of houses in each segment. We normalize the total number of houses in the economy to one:

$$\bar{\mu}^H = \sum_{h \in H} \mu^H(h) = 1.$$

Let Θ denote a finite set of agent types. Agents have quasilinear utility over two goods: numeraire ("cash") and housing services. Agents own at most one house. When an agent moves into a house, he obtains housing services $v(h) > 0$ until the house falls out of favor, which happens at the rate $\eta(h)$. After the house falls out of favor, the agent no longer receives housing services from that particular house. The agent can then put the house on the market in order to sell it and subsequently search for a new favorite house. We assume that the search for a new house is costless, whereas putting the a house on the market in segment h involves costs $c(h)$ per period.

Agent type θ is identified by a *search range*, a subset $\tilde{H}(\theta) \in H$ of market segments that he is interested in. The data counterpart of a search range is a search pattern derived from our Trulia search alert data – it will typically span multiple segments. Search ranges are part of the description of preferences – an agent will never move into a house outside $\tilde{H}(\theta)$. We use a measure μ^Θ on Θ to count the number of agents of each type. The total number of agents is

$$\bar{\mu}^\Theta = \sum_{\theta \in \Theta} \mu^\Theta(\theta) > 1.$$

Since there are more agents than houses and agents own at most one house, some agents are always searching. The idea is that these $\bar{\mu}^\Theta - 1$ agents rent or stay in a hotel while they search for a house to buy.

Our search query data tell us for each type θ and segment h whether θ is interested in h or not. The resulting "who-searches-where"-matrix allows us to measure the interconnectedness of segments. We define the *clientele* of segment h as the set of all agents who are interested in segment h

$$\tilde{\Theta}(h) := \left\{ \theta \in \Theta : h \in \tilde{H}(\theta) \right\}.$$

It is helpful to consider two extremes. The market is fully segmented if every segment is searched by a single type who is interested only in that segment. In contrast, the market is fully integrated if there is only one type who searches all segments.

We can combine search query data and inventory data to look at the inventory that a given type θ is interested in. Let $\mu^S(h)$ denote the number of houses for sale in segment h. The *inventory scanned* by type θ is then

$$J(\theta) := \sum_{h \in \tilde{H}(\theta)} \mu^S(h) \tag{1}$$

The clientele of the typical segment h consists of multiple types scanning different inventories. Those inventories partly overlap, at least because everyone scans $\mu^S(h)$.

Matching

Matching in the housing market involves searchers scanning inventory, identifying suitable properties and making contact with sellers. We capture this process by a random matching technology. We make two key assumptions. First, we assume that searchers flow into segments within their search range in proportion to segment inventory. This assumption is natural if searchers are equally likely to find their favorite house anywhere in their search range. Formally, let $\tilde{\mu}^B(\theta)$ denote the number of buyers of type θ. We define the number of buyers in segment h as

$$\mu^B(h) = \sum_{\theta \in \tilde{\Theta}(h)} \frac{\mu^S(h)}{J(\theta)} \tilde{\mu}^B(\theta) \tag{2}$$

For the given segment h, buyers can belong to any type in the clientele $\tilde{\Theta}(h)$. If a type θ searches only segment h, then $J(\theta) = \mu^S(h)$ and all buyers $\tilde{\mu}^B(\theta)$ of type θ are in fact buyers in h. If segments have roughly the same inventory, searchers are equally likely to be buyers in any of the segments in their search range. More generally, the more inventory is available in h relatively to other segments in type θ's search range, the larger the share of type θ buyers who flow into h.

Our second assumption is the presence of a matching function. The match rate in segment h is given by

$$m(h) = \tilde{m}\left(\mu^B(h), \mu^S(h), h\right),$$

where $\tilde{m}$ is increasing in the number of buyers and sellers and with $\tilde{m}\left(0, \mu^S, h\right) = \tilde{m}\left(\mu^B, 0, h\right) = 0$. At this point, we do not make further assumptions on the functional form of the function $\tilde{m}$. What is important is that it is allowed to depend on the segment h directly (that is, other than through the number of buyers and inventory). For example, the process of scanning inventory could be faster in a segment because the properties are more standardized, or because more open houses are available to view properties.

Once a buyer and seller have been matched, the seller makes a take-it-or-leave-it offer.[4] If the buyer rejects the offer, the seller keeps the house and the buyer continues searching. If the buyer accepts the offer, the seller starts to search, whereas the buyer moves into the house and begins to receive utility $v(h)$.

Equilibrium

In equilibrium, agents make optimal decisions taking as given the distribution of others' decisions. In particular, owners decide whether or not to put their houses on the market, sellers choose price offers and buyers choose whether or not to accept those offers. In what follows, we focus on steady state equilibria in which (i) owners put their house on the market

[4]This assumption on the bargaining protocol is sufficient to ensure that agents find the strategies optimal, which we discuss below.

if and only if their house fell out of favor, so that the owners do not receive housing services from it, and (ii) all offers are accepted.

Since the model has a fixed number of agents and houses, the steady state distributions of agent states can be studied independently of the prices and value functions. We need notation for the number of agents who are in different states. Let $\mu^H(h;\theta)$ denote the number of type θ agents who are homeowners in segment h, and let $\mu^S(h;\theta)$ denote the number of type θ agents whose house is listed in segment h. Finally, let $\tilde{\mu}^B(\theta)$ the number of type θ agents who are currently searching to buy a house. In steady state, all these numbers are constant. We now derive a set of equations to determine these numbers.

Since $\mu^S(h;\theta)$ must be constant in steady state, the number of houses newly put on the market by type θ agents in segment h must equal the number of houses sold by type θ agents in segment h:

$$\eta(h)\left(\mu^H(h;\theta)-\mu^S(h;\theta)\right)=\frac{\mu^S(h;\theta)}{\mu^S(h)}\tilde{m}\left(\mu^B(h),\mu^S(h),h\right) \tag{3}$$

The left-hand side shows houses coming on the market, given by the rate that houses fall out of favor multiplied by the number of houses by owners of type θ not on the market. The right-hand side shows houses for sale by owners of type θ that are matched with buyers and thus leave the inventory, given by the share of type θ seller in segment h multiplied by the match rate.

Since the number $\mu^H(h;\theta)$ of houses that are put on the market by owners θ is also constant over time, the share of houses sold by type θ agents in segment h must equal the share of houses bought by type θ agents in market h (both as a fraction ot total volume.)

$$\frac{\mu^S(h;\theta)}{\mu^S(h)}=\frac{\mu^S(h)}{J(\theta)}\frac{\tilde{\mu}^B(\theta)}{\mu^B(h)}. \tag{4}$$

On the right hand, the share of type θ buyers in segment h equals the number of type θ buyers that flow to h in proportion to inventory, as in (2), divided by the total number of buyers in segment h.

The number of agents and the number of houses must add up to their respective totals:

$$\begin{aligned}\mu^H(h)&=\sum_{\theta\in\tilde{\Theta}(h)}\mu^H(h;\theta),\\ \mu^\Theta(\theta)&=\tilde{\mu}^B(\theta)+\sum_{h\in\tilde{H}(\theta)}\mu^H(h;\theta).\end{aligned} \tag{5}$$

Equations (2), (3), (4) and (5) jointly determine the unknown numbers $\mu^H(h;\theta)$, $\mu^S(h;\theta)$, $\mu^B(h)$ and $\tilde{\mu}^B(\theta)$, a system of $\#H+\#\Theta(1+2\#H)$ equations in as many unknowns.

4.1 Understanding the cross section of observables

The model identifies three forces that determine market activity and prices in the cross section: the rates $\eta(h)$ at which houses fall out of favor, the distribution of search ranges

$\tilde{H}(\theta)$ and the number of agents $\mu^{\Theta}(\theta)$ of type θ, and the segment-specific effects on match rates summarized by $\tilde{m}(.,.,h)$. These forces correspond, roughly, to differences in supply, demand and market frictions, respectively. We now develop some intuition for how these exogenous forces drive the cross section of observables in equilibrium.

Observables

Given a measure of housing stock, we define the *inventory share* in a segment as the share of all houses for sale $I(h) = \mu^S(h)/\mu^H(h)$. Similarly, the *turnover rate* in a segment is defined as $V(h) = m(h)/\mu^H(h)$. In a steady state equilibrium, the average time a house is on the market is given by

$$T(h) = \frac{\mu^S(h)}{m(h)} = \frac{I(h)}{V(h)}. \tag{6}$$

It follows that a measure of inventory share can be recovered from measures of turnover and time on the market.

The model also suggests a simple way to measure buyer interest at the segment level. Indeed, the inventory scanned by a searcher $J(\theta)$ can be measured using data on inventory together with alert data. Since the alerts are a sample of buyers, we also have a measure of the relative (but not the absolute) frequency of a search profile $\tilde{H}(\theta)$ in the total buyer pool:

$$\beta(\theta) := \frac{\tilde{\mu}^B(\theta)}{\bar{\mu}^{\Theta} - 1}.$$

Given our knowledge of $\beta(\theta)$, $\mu^S(h)$ and $J(\theta)$, we observe the number of buyers $\mu^B(h)$ at the segment level up to a constant. We can therefore define a directly observable measure of *relative buyer interest* in a segment as

$$D(h) = \frac{\mu^B(h)}{\mu^H(h)(\bar{\mu}^{\Theta} - 1)}. \tag{7}$$

Here we normalize by the segment housing stock in order to take out segment scale. The measure of buyer interest can be directly computed from data; the only assumption from the model used is that searcher flow to segments in proportion to inventory as in equation (2).

Perfect segmentation

The mapping from parameters to observables depends on the nature of the search patterns. For example, parameters that are "local" to segment h, such as the rate at which houses come on the market there, will matter less for local inventory and time on the market if segment h is more integrated with similar segments. To provide some intuition on the role of integration and its interaction with other forces, we first consider the extreme case of perfect segmentation.

Suppose there are exactly as many types as segments and each type scans exactly one segment. We use the label $\theta = h$ for the type scanning segment h and otherwise drop θ arguments. From (3), equilibrium inventories are determined segment by segment by

$$\eta(h)\left(\mu^H(h) - \mu^S(h)\right) = \tilde{m}\left(\mu^{\Theta}(h) - \mu^H(h), \mu^S(h), h\right). \tag{8}$$

The left-hand side is the rate at which houses come on the market in segment h. It is strictly decreasing in inventory: higher inventory means that fewer agents are living in their favorite house and thus fewer houses can come on the market each instant. The right-hand side describes the rate at which houses are sold. It is strictly increasing in inventory: higher inventory means that buyers are more likely to be matched with a house. It follows that there is a unique equilibrium level of inventory μ^S – if inventory is too low, then too many houses come on the market whereas if inventory is too high, then too many houses are sold.

Consider how segments differ in the cross section. If we divide by $\mu^H(h)$, the right side of equation (8) shows the steady state turnover rate; time on market is given by $T(h) = \mu^S(h)/m(h)$. If houses come on the market more quickly (higher $\eta(h)$), then turnover increases together with inventory and time on the market. In contrast, if there are more buyers for the same number of houses (higher $\mu^\Theta(h)$), then turnover also increases, but inventory and time on the market decline. Intuitively, an increase in either demand or supply increases volume. The difference is that an increase in supply also make the market clear more slowly so inventory is higher.

Perfect integration

Consider now the opposite polar case of perfect integration: all segments are exclusively scanned by a single type. Both the number of buyers and the number of sellers are then proportional to inventory $\mu^S(h)$. In other words, the "queue length" $\mu^B(h)/\mu^S(h)$ is equated across segments. According to equation (2), the buyers $\mu^B(h)$ in segment h are the fraction of total buyers $\bar{\mu}^\Theta - 1$ that flow into segment h; it is determined by the share of inventory in the overall inventory, $\mu^S(h)/\sum_{h\in H}\mu^S(h)$. Equilibrium inventories again adjust equate the flow of houses coming on the market to the volume of sales:

$$\eta(h)\left(\mu^H(h) - \mu^S(h)\right) = \tilde{m}\left(\frac{\mu^S(h)\left(\bar{\mu}^\Theta - 1\right)}{\sum_{h\in H}\mu^S(h)}, \mu^S(h), h\right) \tag{9}$$

The effect of an increase in $\eta(h)$ in segment h is qualitatively similar to the case of full segmentation: turnover and inventory will both increase. However, the effect on inventory will typically be weaker because more searchers flow into h as more houses come on the market there. In other words, an increase in supply endogenously gives rise to an offsetting increase in demand. If the matching function does not depend on the segment but has constant returns in μ^B and μ^S, then time on the market is also equated across segments. More generally, segment-specific frictions that affect the speed of matching would lead to differences in time on market even in a fully integrated area.

Equation (9) suggests that we can learn about the importance of integration by looking at a scatter plot of inventory share $I(h)$ and our buyer-interest measure $D(h)$. If all segments are perfectly integrated, then all points $(D(h), I(h))$ must lie on a straight line with a positive slope. More generally, suppose that not all segments are perfectly integrated, but there is a subset of perfectly integrated segments that do not share common buyers with any other segment. The points for that subset should then be on a straight line.

While upward sloping pieces of a $(D(h), I(h))$ are indicative of integration, the latter cannot be inferred from segment level data alone. Indeed, any cross section of buyer interest and inventory at the segment level they could in principle also be generated by a perfectly segmented economy. For example, consider equation (8). An increase in $\eta(h)$ increases $I(h)$, leaving $D(h)$ unchanged. An increase in the number of buyers $\mu^{\Theta}(h)$ increases $D(h)$ and decreases $I(h)$. A perfectly segmented economy in which, for reasons exogenous to the model, houses come on the market more quickly in areas in which there are more buyers can thus also generate an upward sloping line in the $(D(h), I(h))$ plane.

4.2 Defining segments

The ideal way to define market segments from data would be to consider the partition of the universe of Bay Area houses that is implied by joining all our search queries. Any division of houses into segment would then be motivated by the preferences of at least one searcher. Moreover, the preferences of any one searcher could be expressed exactly through a subset of the set of all segments. The problem with the ideal approach is sample size: the number of houses per segment and the number of searchers per search range would be too small to accurately measure moments such as time on the market, inventory and buyer interest.

Our approach is essentially to work towards the ideal partition, but subject to the constraint that segments must be sufficiently large in terms of volume and housing stock. This leads us to a set H of 576 segments as well as a set Θ of 9091 search ranges that can each be represented as a subset of H. We provide a detailed description of the algorithm in the Appendix. In what follows we only sketch the main steps.

We start from our earlier result that people search mostly according to (i) quality, by specifying price ranges (ii) geography, specifying zip code as the finest unit and (iii) size, by specifying the number of bathrooms, typically either "up to 2" or "more than 2" bathrooms. Facts (ii) and (iii) lead us to first divide the Bay Area by zip code and then divide each zip code into size categories.

To accommodate search according to quality, we further divide – zip code by zip code – each size group into four price groups. Here we start from a set of candidate price cutoffs: \$200K, \$300K, \$400K, \$500K, \$750K and \$1mn.[5] We then select three cutoffs from these candidates that are most often close to price cutoffs appearing in our email alerts. The idea is that the resulting set of segments is close to the ideal partition implied by the alerts. In particular, high prices zip codes will typically have higher cutoffs than lower priced zip codes.

At this point, we have divided each zip code into eight size-price groups. It is possible, however, that some of the groups are too small to provide accurate measures of segment level moments. Our criteria here is that a segment must have enough number of transactions as well as a sufficiently large housing stock. If this is not the case, we merge candidate segments

[5]We convert all prices in 2010 dollars using zip code level repeat sales price indices. This allows us to compare alerts set at different points in time as well as measure moments such as monthly volume in a segment by pooling transactions over time.

to form a larger joint segment. As a result, some zip codes that have very thin housing markets might have very few segments.

Given a final set of segments H, we express the distribution of search patterns derived from alerts as a distribution of subsets of H. For each raw search pattern, we consider the range specified along the dimensions quality, size and geography, ignoring other dimensions. We then determine the set of segments that is approximately covered by the specified range. Some detail is lost at this step: the number of distinct patterns drops from about 30K to about 9K.

We can also calculate, for each segments, average monthly volume and time on the market for each segment from our deeds and listings data, based on data for the period 2008-2012. In addition, we use 2010 ACS data to compute housing stock by segment. Combining these numbers, we obtain the key data moments for the analysis to follow, the inventory share $I(h)$, buyers per house $D(h)$ and turnover $V(h)$.

4.3 Segment level facts

For each segment and month, we determine (1) the median sales price, (2) the total volume of transactions, (3) the average time on market for houses sold in that month and (4) the total inventory of listings per segment and month. For each month, price by segment is determined by finding the median observed transaction price in each segment. Volume for month is constructed by calculating the total number of transactions in each segment. We normalize volume in a segment by dividing it by the total stock of residential housing in the segment in 2009, as determined from the tax assessment records. Time on market for each property is calculated by measuring the time period between the initial listing date and the final sale. We then calculate the average time on market for all properties sold in that particular month.

Table 5 shows summary statistics for our key time series variables. The average segment had a transaction volume of about 0.34 percent of the housing stock transacting each month. The average across our segments in terms of the median sales price was about $615,000. The average segment had about 3 percent of the total housing stock listed for sale, and houses that eventually sell would be on the market for about 5 months.

There is substantial heterogeneity across segments. The standard deviations of the variables in Table 5 across segments are between roughly one half and two thirds of their means. There are segments in which median house prices are below $100k, while others have median house prices that are 30 times as high. In some segments, houses sit on the market for almost a year, while houses in other segments sell in a month. Volume and inventory can be low in some segments, and 100 times higher in other segments.

Table 5 also illustrates that aggregation to the zip code or city omits a large fraction of this heterogeneity at the segment level. A regression of house prices in each segment on house prices in the zip code (and a constant) has an R^2 of 49%. This R^2 drops further to 1% when we use median prices in the city. The R^2s in the regressions on time on market,

volume and inventory at the zip code are also around 50%, while R^2s drop to 8-16% at the city level. These statistics stress the importance of working with disaggregated data – half of the variation in these market variables is lost when we use data aggregated in zip codes, and more than 84% is lost at the city level.

Table 5: Market Activity in Segments

	House Price (Dollars)	Time on Market (months)	Turnover (percent)	Inventory Share (percent)	Buyer Interest
Mean	614,891	4.77	0.34	2.97	1.04
Std.Dev.	412,756	1.45	0.20	1.99	0.92
Min	79,839	1.53	0.01	0.31	0.03
Max	2,480,551	11.27	1.75	15.59	8.61
% Zip Code	0.49	0.40	0.49	0.48	
% City	0.01	0.16	0.16	0.08	
Cross Sectional Regressions on Zip Code Price and Segment−Zip Code Price					
Zip code Price		−1.24	−0.0018	−0.0155	0.35
(t-stat)		(−13.6)	(−9.2)	(−9.1)	(5.5)
Segment-Zip Price		−1.55	−0.0014	−0.0143	−0.68
(t-stat)		(−12.3)	(−8.5)	(−9.2)	(−9.8)
R^2		0.43	0.30	0.26	0.22
Cross Sectional Regressions on Segment Price with Zip Code Fixed Effects					
Segment		−1.16	−0.003	−0.0146	−0.80
(t-stat)		(−15.5)	(−9.5)	(−10.9)	(−16.7)
R^2		0.71	0.61	0.62	0.67

Note: This Table shows summary statistics of the market activity in housing segments. The "% Zip Code" (or "% City") is the R^2 of a cross sectional regression of the segment level variable on a constant and its zip code (or city level) mean. The 'Cross Sectional Regressions on Zip Code Price and Segment−Zip Code Price' run the segment level variable on a constant, the zip code house price and the difference between the house price in the segment and the price in the zip code. The 'Cross Sectional Regressions on Segment Price and Zip Code Fixed Effects' run the segment level variable on a constant, the house price in the segment and zip code fixed effects. The table reports t-statistics in brackets.

Figure 10 illustrates these cross sectional facts on a map of the Bay Area. The dots in the map represent the individual segments. Since there can be multiple segments within a zip code, we plot the segment-dots in a zip code in a circle. The cheapest segment in the zip code is on top of the circle and house prices increase going clockwise around the circle.

The left panel shows that inventory is low (light blue) in expensive areas. These include

the downtown area of San Francisco, with a few northeast segments that have somewhat more inventory (darker blue), Berkeley (the light blue area across the Bay) and much of Silicon Valley (between San Francisco and San Jose). Inventory is higher in the cheaper areas, such as San Jose, the East Bay around Oakland, and the Sacramento Delta in the northeast of the map. Table 5 confirms this visual impression with the cross sectional regressions, where the coefficient of inventory on zip code house prices is negative.

The map also reveals variation within zip codes. In most circles, the top dot is darker indicating that inventory in the cheapest segment within zip codes is higher. Table 5 confirms this pattern with the negative coefficient on the house price in a segment, whether we control for zip code price or for zip code fixed effects.

The right panel shows buyer interest for the various Bay Area segments. Buyer interest is high in expensive areas – downtown San Francisco, in Marin (across the Golden Gate Bridge), Berkeley, and Silicon Valley. Interestingly, buyer interest within zip codes is highest for the cheapest segment. Indeed, the regression of buyer interest on house prices in the zip code and the segment has a positive coefficient on the zip code house price; buyer interest in more expensive zip codes is higher. But the coefficient on the difference between the segment price and the zip price is negative, which means that buyer interest for the cheaper houses within zip code is higher.

5 The cross section of housing markets

We now describe how we infer the parameters η and μ^{Θ} as well as features of the matching function $\tilde{m}$ from market activity data. The goal is to assess the relative importance of the three forces – supply, buyer interest and market frictions – in accounting for the cross section of housing markets.

Exact identification of parameters

We assume a matching function such that the model can exactly match numbers on (i) the cross sections of turnover $V(h)$ and inventory $I(h)$ by segment and (ii) for each search range $\tilde{H}(\theta)$, the share $\beta(\theta)$ of the total number of searchers scanning that range (iii) the average time spent searching, T^b. This requires that the direct effect of the segment on the match rate is sufficiently flexible.[6]

As long as the matching function is such that we can match moments (i)-(iii), inference about the parameters η and μ^{Θ} is independent of the exact shape of $\tilde{m}$. From the system of equations that determines the equilibrium, we have that, for each set of observables I, V, $\tilde{\beta}$

[6]For example, we could start from a matching function $\hat{m}\left(\mu^B, \mu^S\right)$ that depends on the number of buyers $\mu^B(h)$ and inventory $\mu^S(h)$ and then add a segment-specific proportional factor, i.e.

$$\tilde{m}\left(\mu^B, \mu^S, h\right) = \bar{m}(h)\,\hat{m}(\mu^B, \mu^S, h).$$

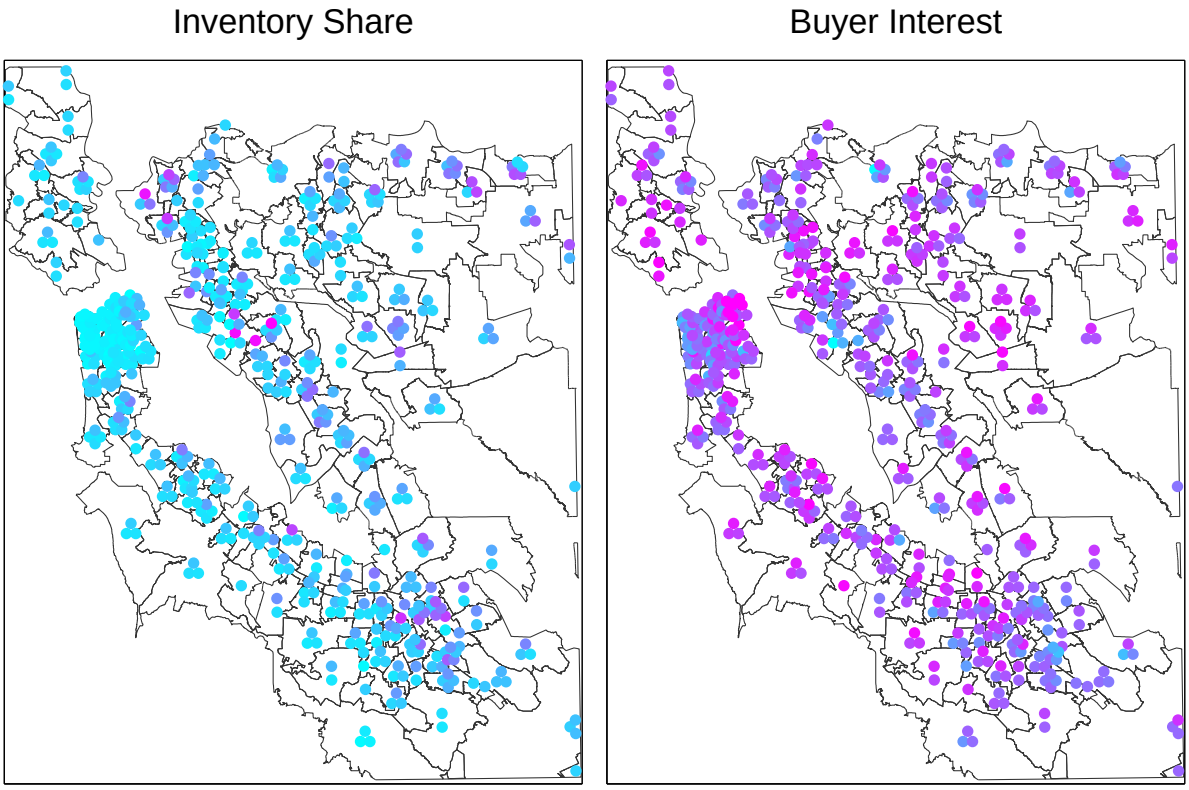


Figure 10: Map of the Inventory Share and Buyer Interest. Left panel shows the distribution of inventory share $I(h)$. Right panel shows buyer interest $D(h)$. Colors move from light blue (low values) to pink (high values). The cheapest segment in the zip code is on top of the circle and house prices increase going clockwise around the circle.

and T^b, there is a unique set of parameters η and μ^Θ as well as (endogenous) buyer matching probabilities

$$\alpha^B(h) := \frac{\tilde{m}\left(\mu^B(h), \mu^S(h), h\right)}{\mu^B(h)}. \tag{10}$$

Moreover, the parameters $\mu^\Theta(\theta)$ and the observable moments $I(h)$, together with the observable exogenous housing stocks $\mu^H(h)$, imply unique numbers of buyers and sellers $\mu^B(h)$ and $\mu^S(h)$, respectively. As a result, if we knew the functional form of the matching function, we could back out the relevant parameters from (10). For now, we simply report the properties of supply, buyer interest and frictions in terms of $\eta(h)$'s, $\mu^\Theta(\theta)$'s and $\alpha^B(h)$'s, respectively. Numbers for $V(h)$, $I(h)$ and $\tilde{\beta}(\theta)$ come directly from segment level transaction and listing data as well as search pattern data. We set the target for the average search time to 6 months.

Instability

The rate at which owners put houses on the market can be inferred from inventory and

turnover alone, independently of the distribution of search patterns. Indeed, in steady state, dividing (3) by the segment housing stock $\mu^H(h)$, the share of houses that come on the market per unit of time in a segment must equal the turnover rate

$$\eta(h)(1 - I(h)) = V(h).$$

The share $1 - I(h)$ of the segment housing stock that is not currently for sale is typically close to one. As a result, the rates η closely tracks turnover in the cross section. We refer to a segment with higher η as more unstable, since homeowners move more often.

While on average monthly turnover rates, and hence $\eta(h)$'s are small, the differences in $\eta(h)$'s by segment are substantial. For example, the 25th percentile, $\eta(h) = 0.0021$ whereas at the 75th percentile $\eta(h) = 0.0042$. In unstable segments houses come on the market at more than double the rate than in stable segments. Since η tracks turnover, we also know from the previous section that it is strongly negatively correlated with price. In other words, cheaper segments tend to be less stable. The left panel of Figure 11 maps instability by segment (light blue represents a lower η, pink represents higher η) - this confirms that more expensive areas are more stable.

In our model, instability captures the mobility of the population of a segment. It is natural to ask whether the backed out parameters $\eta(h)$ comove with demographic variables that are related to the frequency of moving. Table 6 shows that instability is highly correlated with the fractions of households in various age groups. To establish this fact, we proceed in two steps. We first run regressions of $\eta(h)$ on zip code fixed effects. Second, we correlate the zip code fixed effects with demographic variables as reported in the American Community Survey. The results show that the percentage of older households is negatively correlated with $\eta(h)$. In other words, areas with a larger fraction of older households are more stable. Similarly, zip codes with higher unemployment rates and lower median income are more unstable.

Popularity

The distribution of buyer types contains information on how popular the segment is, in the sense that many searchers are interested in it. We would like to summarize popularity at the segment level, while taking into account that a searcher who scans a bigger housing stock is less interested in a particular segment within that stock. We thus split agents who scan multiple segments in proportion to the housing stock in those segments. We then define popularity as the weighted number of interested agents per house:

$$\delta(h) = \sum_{\theta \in \tilde{\Theta}(h)} \frac{\mu^{\Theta}(\theta)}{\sum_{h \in \tilde{H}(\theta)} \mu^H(h)}.$$

If the market is perfectly segmented, then $\delta(h) = \mu^{\Theta}(h)/\mu^H(h)$, where $\mu^{\Theta}(h)$ is the number of types who uniquely search segment h. In this case, every $\delta(h)$ is larger than

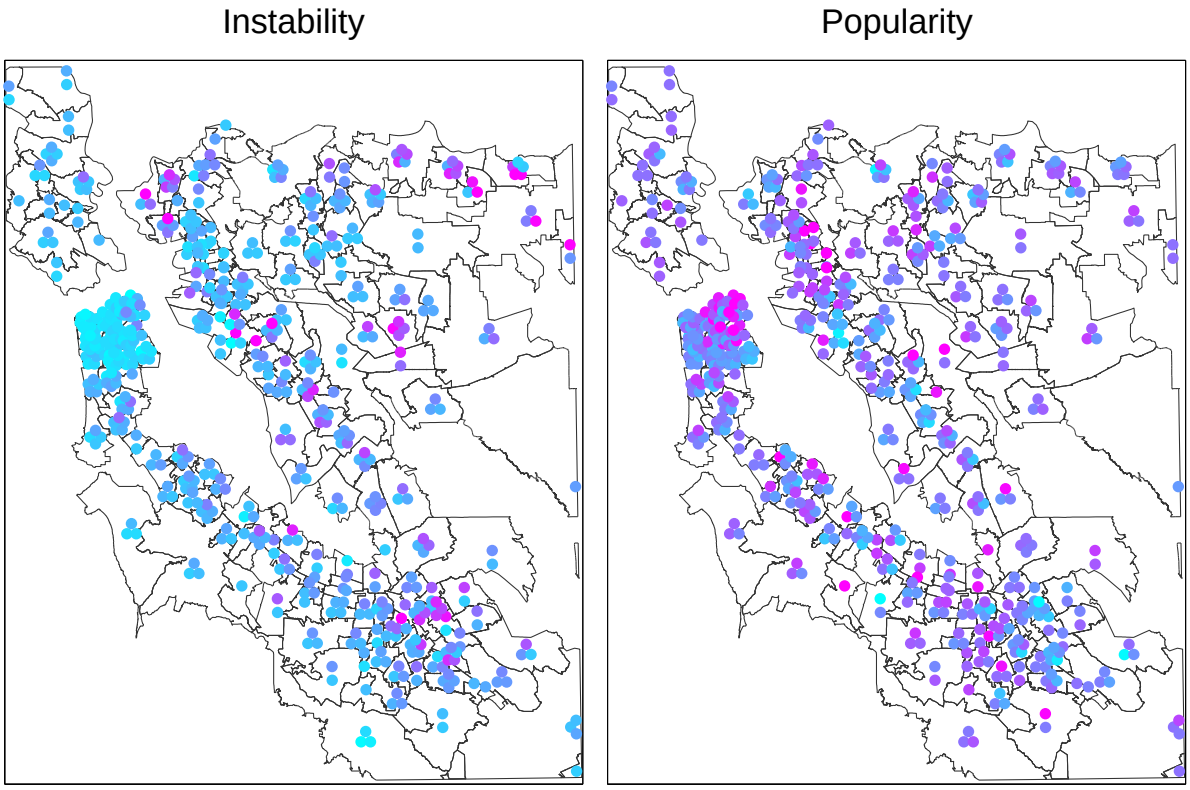


Figure 11: Map of Instability and Popularity. This figure shows the distribution of instability, $\eta(h)$, in the left panel, and popularity, $\delta(h)$, in the right panel. Colors move from light blue (low values) to pink (high values). The cheapest segment in the zip code is on top of the circle and house prices increase going clockwise around the circle.

one – some agents must be searching in equilibrium in every segment because volume is positive. Differences in $\delta(h)$ across segments reflect differences in local interest per house. In contrast, if the market is perfectly integrated, then the $\delta(h)\, s$ reflect the preferences of a single searcher. We will have $\delta(h) = \sum \mu^{\Theta}(\theta)$. We thus again obtain a number larger than one for all h. Table 6 shows that those segments with a higher income and a lower unemployment rate are more popular and hence attract more interest from buyers. There is only a weak correlation of popularity with the age composition of the households. The right panel of Figure 11 shows the geographic distribution of popularity: Light blue segments are less popular, pink segments are more popular.

5.1 Equilibrium prices

Since sellers make a take-it-or-leave offers, they charge a price equal to the buyers' continuation utility. Denote by $V^F(h;\theta)$ the utility of a type θ agent who obtains housing services from a house in segment h. At the equilibrium actions, the Bellman equations of that agent

Table 6: Correlations between Estimated Model Parameters and Demographics

	Instability	Popularity	Buyer Match Rate
Unemployment	0.48	−0.36	0.60
Median Income	−0.29	0.22	−0.47
Fraction of Households Aged			
15-24 years	0.42	−0.15	
25-34 years	0.42	0.01	
35-44 years	0.38	0.01	
45-54 years	0.23	−0.12	
55-59 years	−0.18	−0.05	
60-64 years	−0.42	0.16	
65-74 years	−0.46	0.16	
75-84 years	−0.38	−0.06	
above 85 years	−0.32	−0.05	

Note: This table reports correlations between demographic characteristics at the zip code level with the fitted values from a regression of the segment variables indicated on top (instability $\eta(h)$, popularity, buyer match rate) on zip code fixed effects.

as well as that of a seller are

$$rV^F(h;\theta) = v(h) + \eta(h)\left(V^S(h;\theta) - V^F(h;\theta)\right),$$
$$rV^S(h;\theta) = \frac{m(h)}{\mu^S(h)}\left(p(h)(1-c) - V^S(h;\theta)\right),$$

where $p(h)$ is the price of houses in segment h. We can combine these equation and solve for the price

$$p(h) = \frac{v(h)}{r} - \frac{\eta(h)}{r + m(h)/\mu^S(h) + \eta(h)}\frac{v(h)+c}{r} \tag{11}$$

The first term is the present value of a permanent flow of housing services. This price obtains if houses never fall out of favor ($\eta = 0$) or if the market is frictionless in the sense that matching is infinitely fast ($m/\mu^S \to \infty$). More generally, the price incorporates a liquidity premium – the second term – that reflects foregone utility flow during search as well as the cost of search itself. The liquidity premium is larger if houses fall out of favor more quickly (η higher) and if it is more difficult to sell a house in the sense that time on market $\mu^S(h)/m(h)$ is longer.

5.2 Beveridge curves within and across cities

Figure 12

Figure 12: Beverage curves within and across cities.

5.3 Liquidity discounts

We now ask how market frictions identified by our estimation affect the dispersion of house prices across segments. The price formula (11) shows how the price is determined as the difference between a "fundamental" price $v(h)/r$ and a *liquidity discount* that capitalizes the present value of search and transaction costs. The latter are segment-specific: the popularity and instability properties of a segment derived above affect both the average time on the market (and hence search costs) as well as turnover (and hence the frequency at which transaction costs arise). What is as yet missing to evaluate the formula is a measure of fundamental value.

To estimate both fundamental value we can use the cross section of median prices together with our estimation results. In particular, for each cross section of prices and parameter vector, we can back out from (11) the vector of mean utility values $v(h)$ such that the model exactly matches the cross section of transaction prices. We postulate a real interest rate of 2% and set the transaction cost such that the average sale costs 6% of the resale value of the house, a standard number in the literature.

The results are summarized in Figure 13. The left hand panel plots median price against the liquidity discount, stated as a percentage of price. The right hand panel shows the geographic distribution of liquidity discounts. There are two notable results here. First, liquidity discounts are large – they can be up to 40% of the sales prices. Second, liquidity discounts differ widely by segment, oftentimes within the same zip code. In poor segments with high volume and high time on market, both search and transaction costs are high; as a result, prices are significantly lower than they would be in a frictionless market. In rich segments discounts are still significant, but they are considerably smaller.

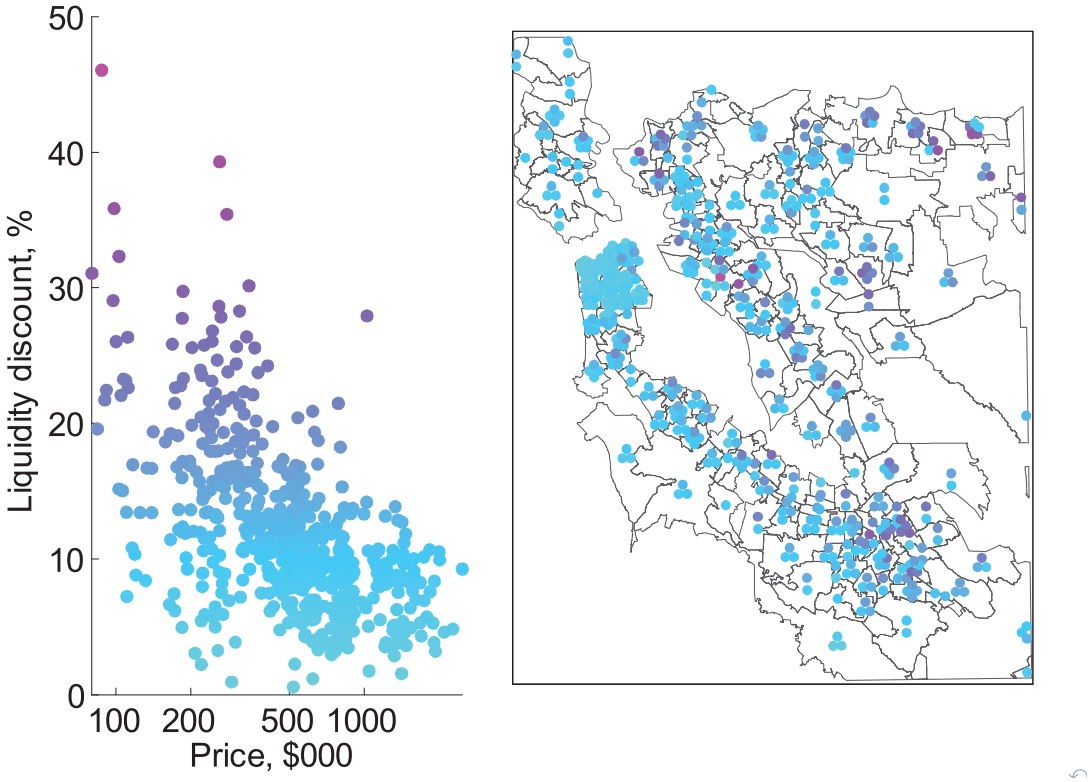


Figure 13: Liquidity Discounts. Left panel: mean zip code price vs zip code liquidity discount as a percentage of mean price; color coding reflects liquidity discount. Right panel: zip codes colored by same code as in right panel.

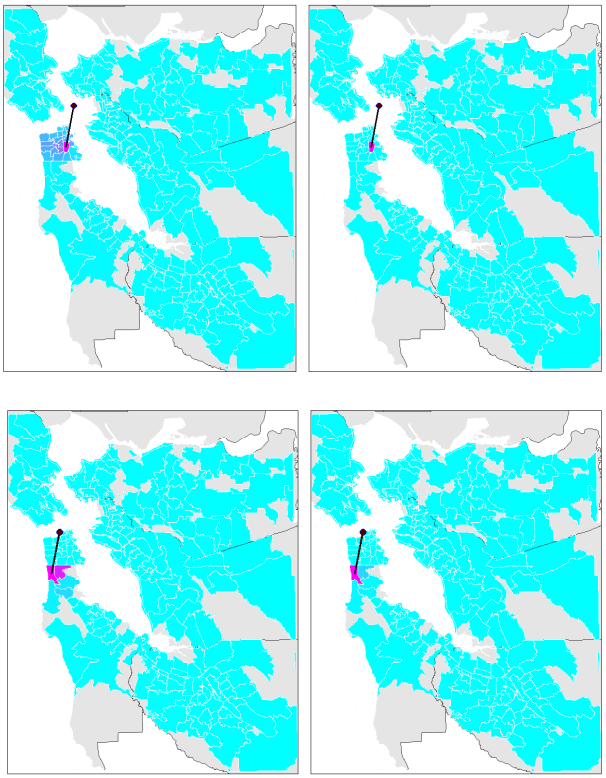
5.4 Comparative statics

Figure 14 shows how the steady state equilibrium changes if the supply of houses in a zipcode is increased by one percent of the existing housing stock in one zipcode only. Formally, we recompute the steady state using the same parameters as above, but we increase $\mu^H(h)$ by one percent in one particular segment h. All panels are maps of only the tip of the San Francisco peninsula. The left two panels assume that the hypothetical change in the

housing stock occurs in zipcode 94102 in downtown San Francisco, marked by a pin. The leftmost panel shows the change in time on market, and the second panel shows the change in inventory. Changes increase from blue to pink.

The result is that a change in 94102 has spillover effects: it increases time on market and inventory all over San Francisco as well as in the suburbs. This is because a large share of searchers scan all these segments jointly. In contrast, the two panels on the right show panels assume that the housing stock increases in the suburb of Daly City (zipcode 94015), again marked with a pin. Here the spillover effects are confined to neighboring zipcodes to the south and west. Remarkably, essentially nothing happens to the north, in the city of San Francisco itself. These results show that search patterns introduce asymmetries in the transmission of shocks.

Figure 14: Responses to one-percent increases in the housing supply.



Note: Panels 1 and 2: responses in time on market and inventory, respectively. The responses are to a one-percent increase in the housing supply in downtown San Francisco (zipcode 94102, marked by a pin). Panels 3 and 4: responses in time on market and inventory, respectively. The responses are to a one-percent increase in the housing supply in the suburb of Daly City (zipcode 94015).

References

Bayer, Patrick, Fernando Ferreira, and Robert McMillan, "A unified framework for measuring preferences for schools and neighborhoods," *Journal of Political Economy*, 2007, *115* (4), 588–638.

Burnside, Craig, Martin Eichenbaum, and Sergio Rebelo, "Understanding booms and busts in housing markets," 2011.

Caplin, Andrew and John Leahy, "Trading frictions and house price dynamics," *Journal of Money, Credit and Banking*, 2011, *43* (s2), 283–303.

Favilukis, Jack, Sydney C Ludvigson, and Stijn Van Nieuwerburgh, "The macroeconomic effects of housing wealth, housing finance, and limited risk-sharing in general equilibrium," 2010.

Floetotto, Max and Johannes Stroebel, "Government Intervention in the Housing Market: Who Wins, Who Loses?," *University of Chicago*, 2012.

Garmaise, Mark J and Tobias J Moskowitz, "Confronting information asymmetries: Evidence from real estate markets," *Review of Financial Studies*, 2004, *17* (2), 405–437.

Genesove, David and Lu Han, "Search and matching in the housing market," *Journal of Urban Economics*, 2012, *72* (1), 31–45.

Glaeser, Edward L. and Joseph Gyourko, "The Impact of Building Restrictions on Housing Affordability," *FRB New York: Economic Policy Review*, 2003, *9* (2), 21–39.

Glover, Andrew, Jonathan Heathcote, Dirk Krueger, and José-Víctor Ríos-Rull, "Intergenerational redistribution in the great recession," 2011.

Goodman, Allen C and Thomas G Thibodeau, "Housing market segmentation," *Journal of housing economics*, 1998, *7* (2), 121–143.

Han, Lu and W Strange, "What is the Role of the Asking Price for a House," *Working Paper*, 2013.

Islam, Kazi Saiful and Yasushi Asami, "Housing market segmentation: A review," *Review of Urban & Regional Development Studies*, 2009, *21* (2-3), 93–109.

Krainer, John, "A theory of liquidity in residential real estate markets," *Journal of Urban Economics*, 2001, *49* (1), 32–53.

Landvoigt, Tim, Monika Piazzesi, and Martin Schneider, "The Housing Market(s) of San Diego," 2012.

Leishman, Chris, "House building and product differentiation: An hedonic price approach," *Journal of Housing and the Built Environment*, 2001, *16* (2), 131–152.

Levitt, Steven D and Chad Syverson, "Market distortions when agents are better informed: The value of information in real estate transactions," *The Review of Economics and Statistics*, 2008, *90* (4), 599–611.

Manning, Alan and Barbara Petrongolo, "How Local are Labor Markets? Evidence from a Spatial Job Search Model," 2011.

Mian, Atif R and Amir Sufi, "House prices, home equity-based borrowing, and the US household leverage crisis," 2009.

National Association of Realtors, "2011 Profile of Home Buyers and Sellers," Technical Report 2011.

_ , "Digital House Hunt," Technical Report 2013.

Ngai, L Rachel and Silvana Tenreyro, "Hot and cold seasons in the housing market," 2009.

Nieuwerburgh, Stijn Van and Pierre-Olivier Weill, "Why has house price dispersion gone up?," *The Review of Economic Studies*, 2010, *77* (4), 1567–1606.

Novy-Marx, Robert, "Hot and cold markets," *Real Estate Economics*, 2009, *37* (1), 1–22.

Piazzesi, Monika and Martin Schneider, "Momentum Traders in the Housing Market: Survey Evidence and a Search Model," *The American Economic Review*, 2009, *99* (2), 406–411.

_ **and** _ , "Inflation and the price of real assets," 2012.

Poterba, James M, David N Weil, and Robert Shiller, "House price dynamics: The role of tax policy and demography," *Brookings Papers on Economic Activity*, 1991, *1991* (2), 143–203.

Stroebel, Johannes, "The impact of asymmetric information about collateral values in mortgage lending," *University of Chicago*, 2012.

Trulia, "The Largest Audience of Real Estate Consumers," Technical Report 2013.

Wheaton, William C, "Vacancy, search, and prices in a housing market matching model," *Journal of Political Economy*, 1990, pp. 1270–1292.

A Data Appendix

A.1 Geographic Specifications in the Queries

Figure 15: Distribution of Geography Specifications

Number of Cities	0	1	2	3	4	5+	Total
0	5,654	5,796	211	130	84	239	12,114
1	23,833	1,108	63	28	8	39	25,079
2	616	46	12	8	7	15	704
3	469	39	5	3	9	12	537
4	428	22	5	3	4	5	467
5+	1,431	104	28	14	10	37	1,624
Total	32,431	7,115	324	186	122	347	40,525

	Number of Neighborhoods						
Number of Cities	0	1	2	3	4	5+	Total
0	5,749	3,533	216	227	220	2169	12,114
1	23,631	1,306	23	19	11	89	25,079
2	524	107	17	10	5	41	704
3	418	87	8	5	1	18	537
4	358	84	4	4	3	14	467
5+	1,149	340	30	18	12	75	1,624
Total	31,829	5,457	298	283	252	2406	40,525

	Number of Neighborhoods						
Number of Zip Codes	0	1	2	3	4	5+	Total
0	24,690	4,799	259	249	224	2210	32,431
1	6,431	462	23	23	21	155	7,115
2	247	53	11	5	3	5	324
3	130	40	2	1	2	11	186
4	90	23	1	2	0	6	122
5+	241	80	2	3	2	19	347
Total	31,829	5,457	298	283	252	2406	40,525

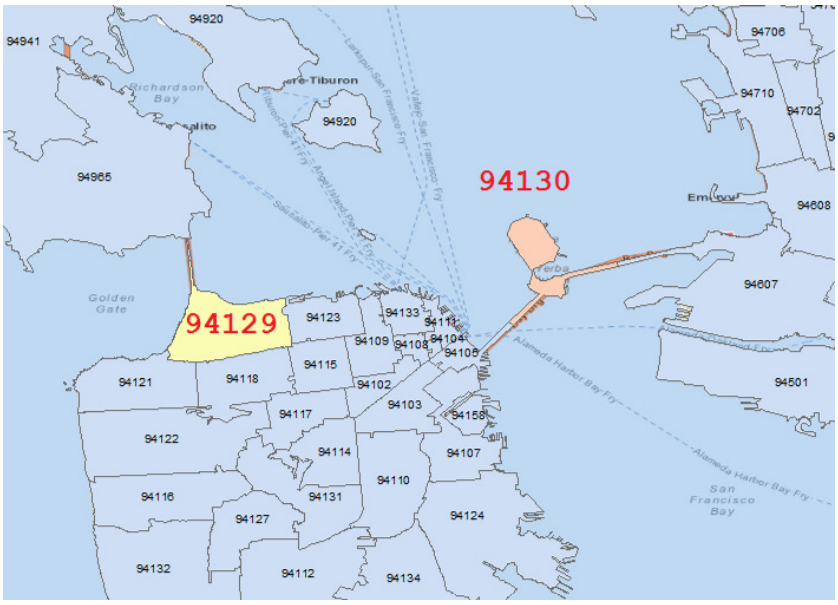
Note: This figure shows the distribution of geographic search parameters selected in our sample.

A.2 Constructing Contiguity Measures

To analyze whether all zip codes are contiguous, one challenge is provided by the San Francisco Bay. The location of this body of water means that two zip codes with non-adjacent borders should sometimes be considered as contiguous, since they are connected by a bridge such as the Golden Gate Bridge. Figure 16 illustrates this. Zip codes 94129 and 94965 should be considered contiguous, since they can be traveled between via the Golden Gate Bridge. To take the connectivity provided by bridges into account, we manually adjust the

ESRI shape files to link zip codes on either side of the Golden Gate Bridge, the Bay Bridge, the Richmond-San Rafael Bridge, the Dumbarton Bridge and the San Mateo Bridge. In addition, there is a further complication in that the bridgehead locations are sometimes in zip codes that have essentially no housing stock, and are thus never selected in search queries. For example, 94129 primarily covers the Presidio, a recreational park, that contains only 271 housing units. Similarly, 94130 covers Treasure Island in the middle of the SF Bay, again, with only a small housing stock. These zip codes are very rarely selected by search queries, which would suggest, for example, 94105 and 94607 would not be connected. This challenge is addressed by manually merging zip codes 94129 and 94130 with the Golden Gate and Bay bridge respectively. This ensure, for example, that 94118 and 94955 are connected even if 94129 was not selected.

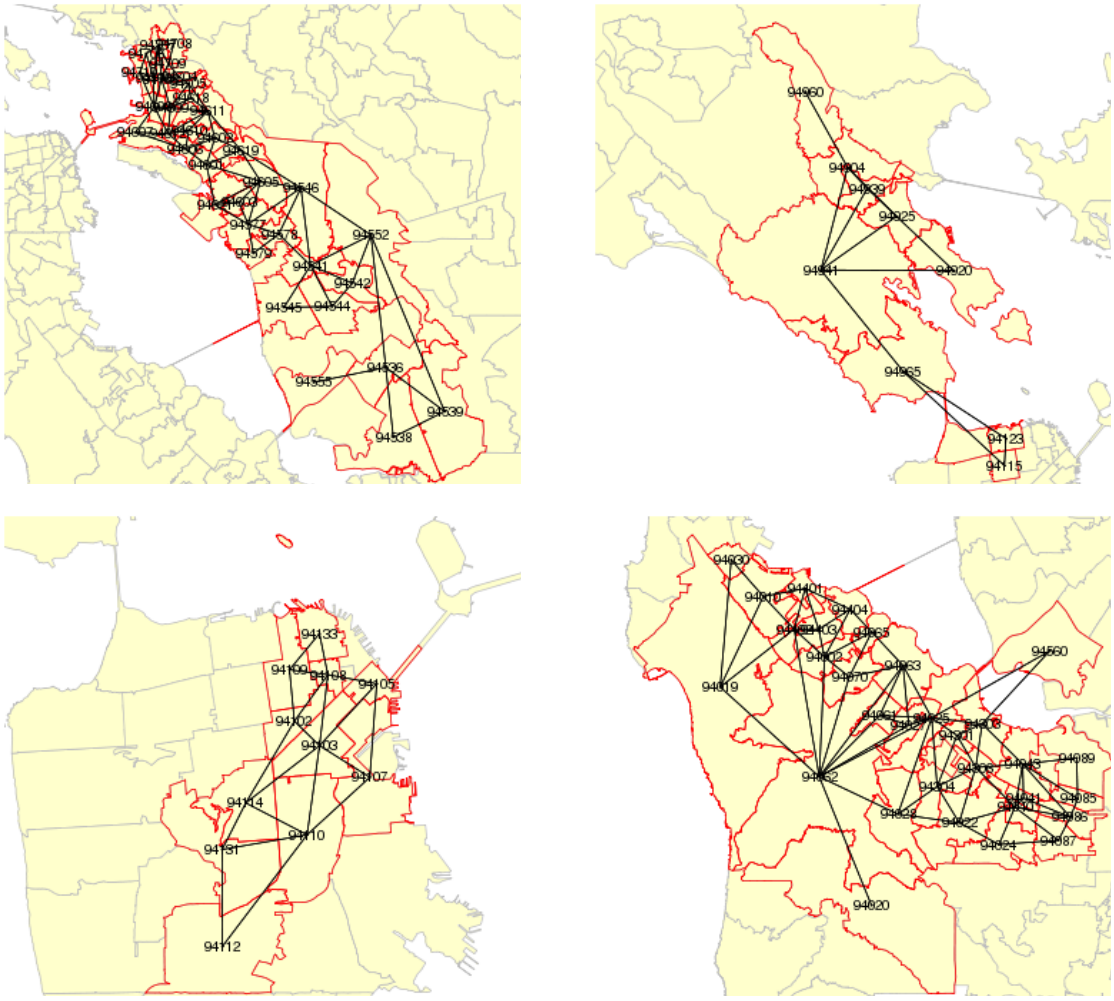
Figure 16: Bridge Adjustments - Contiguity Analysis



Note: This figure shows how we deal with bridges in the Bay Area for the contiguity analysis.

In the following we provide examples of contiguous and non-contiguous search sets. The top left panel of Figure 18 shows all the zip codes covered by a searcher that searched for homes in Berkeley, Fremont, Hayward, Oakland and San Leandro. This is a relatively broad set, covering most of the East Bay. The top right panel shows a contiguous set of jointly searched zip codes, with connectivity derived through the Golden Gate Bridge. The searcher queried homes in cities north of the Golden Gate Bridge (Corte Madera, Larkspur, Mill Valley, Ross, Kentfield, San Anselmo, Sausalito and Tiburon), but also added zip codes 94123 and 94115. The bottom left panel shows the zip codes covered by a searcher that

Figure 17: Sample Contiguous Queries

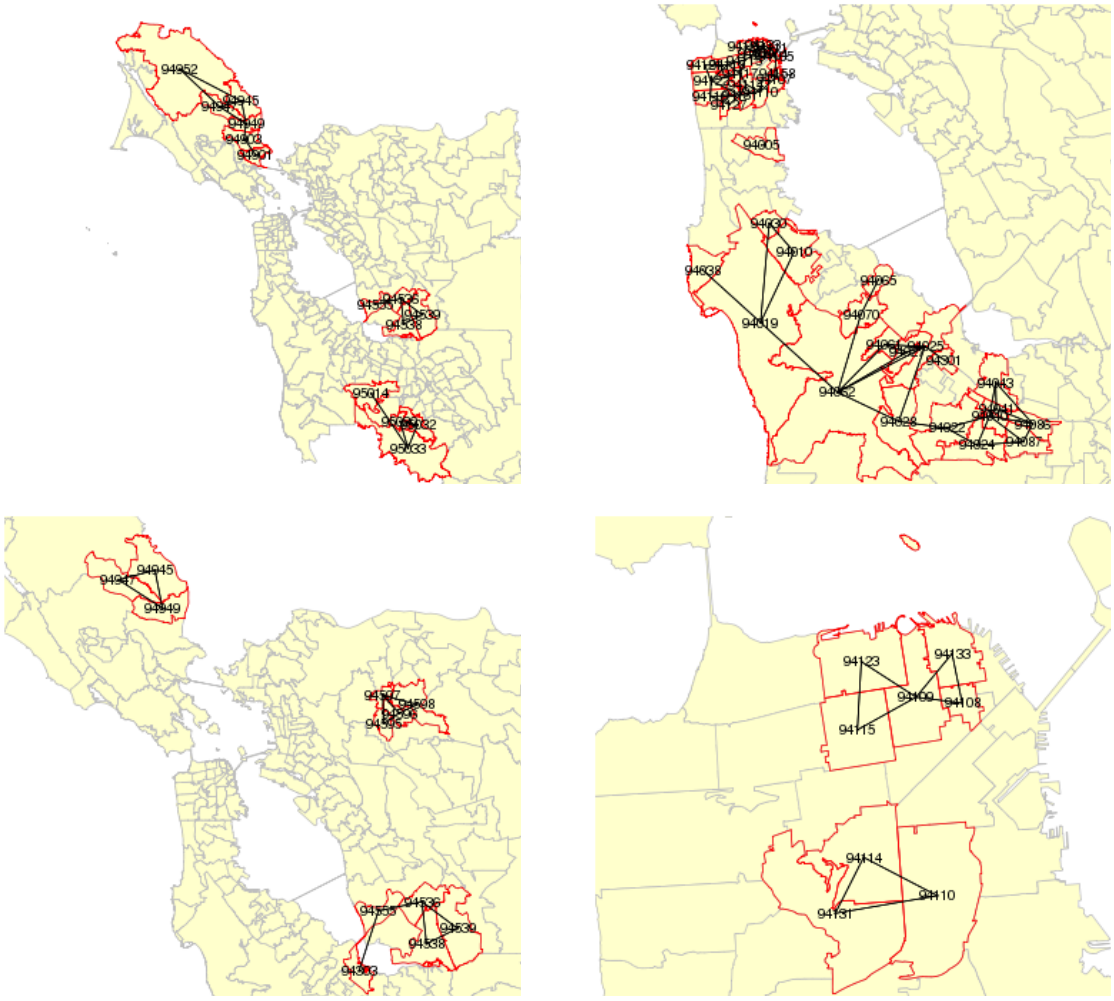


Note: This figure shows a sample of contiguous search sets. The zip codes selected by the searcher are circled in red. Zip code centroids of contiguous zip codes are connected.

selected a number of San Francisco neighborhoods. The final contiguous search set (bottom right panel) was generated by a searcher that selected a significant number of South Bay cities.[7] These are all locations with reasonable commuting distance to the tech jobs in the Silicon Valley. Notice how the addition of Newark adds zip code 94550 on the East Bay, which is connected to the South Bay via the Dumbarton Bridge. Not all email alerts generate sets of zip codes that are contiguous. In Figure 18 we show four actual non-contiguous search sets. The top left panel shows the zip codes covered by a searcher that selects the cities of Cupertino, Fremont, Los Gatos, Novato, Petaluma and San Rafael. This generates three

[7] Atherton, Belmont, Burlingame, El Granada, Emerald Hills, Foster City, Half Moon Bay, Hillsborough, La Honda, Los Altos Hills, Los Altos, Menlo Park, Millbrae, Mountain View, Newark, Palo Alto, Portola Valley, Redwood City, San Carlos, San Mateo, Sunnyvale, Woodside.

Figure 18: Sample Non-Contiguous Queries



Note: This figure shows a sample of non-contiguous search sets. The zip codes selected by the searcher are circled in red. Zip code centroids of contiguous zip codes are connected.

contiguous set of zip codes, rather than one large, contiguous set. The zip codes in the bottom right belong to a searcher that selected zip code 94109 and the neighborhoods Nob Hill, Noe Valley and Pacific Heights. Again, this selection generates more than one set of contiguous zip codes.

A.3 Segment Construction

This section describes the process of arriving at the set of 576 distinct housing market segments for the San Francisco Bay Area. As before, we select the geographic dimension of segments to be a zip code. Since we will compute average price, volume, time on market and

inventory for each segment, we restrict ourselves to zip codes with at least 800 armslength housing transactions between 1994 and 2012. This leaves us with 191 zip codes with sufficient observations to construct these measures.

We next consider how to further split these zip codes into segments based on a quality (price) and size dimension. Importantly, we will need to observe the total housing stock in each segment in order to appropriately normalize moments such as turnover and inventory. The residential assessment records do contain information on the universe of the housing stock. However, as a result of Proposition 13, the assessed property values in California do not correspond to true market value, and it is thus not adequate to divide the total zip code housing stock into different price segments based on this assessed value.[8] To measure the housing stock in different price segments we use the U.S. Census Bureau's 2011 American Community Survey 5-year estimates, which report the total number of owner-occupied housing units per zip code for a number of price bins. We combine a number of these bins to construct the total number of housing units in each of the following price bins: $< \$200\text{k}, \$200\text{k}-\$300\text{k}, \$300\text{k}-\$400\text{k}, \$400\text{k}-\$500\text{k}, \$500\text{k}-\$750\text{k}, \$750\text{k}-\$1\text{m}, > \$1\text{m}$. These bins provide the basis for selecting price cut-offs to delineate quality segments within a zip code. One complication is that the price boundaries are reported as an average for the sample years 2006-2010. Since we want segment price cut-offs to capture within zip code time-invariant quality segments, we need to adjust for average market price changes of the same-quality house over time. To do this, we adjust all prices and price boundaries to correspond to 2010 house prices.[9]

Not all zip codes have an equal distribution of houses in each price (quality) bin. For example, Palo Alto has very few homes valued at less than \$200,000, while Fremont has very few million-dollar homes. Since we want to avoid cutting a zip code into too many quality segments with essentially no housing stock to allow us measure segment-specific moments such as time on market, we next determine a set of three price cut-offs for each zip code by which to split that zip code. To determine which of the seven census price bin cut-offs should

[8] Allocating homes that we observe transacting into segments based on value is much easier, since this can be done on the basis of the actual transaction value, which is reported in the deeds records.

[9] This is necessary, because the Census Bureau only adjusts the reported values for multi-year survey periods by CPI inflation, not by asset price changes. This means that a \$100,000 house surveyed in 2006 will be of different quality to a \$100,000 house surveyed in 2010. We choose the price that a particular house would fetch in 2010 as our measure of that home's underlying quality. To transform the housing stock by price bin reported in the ACS into a housing stock by 2010 "quality" segment, we first construct zip code specific annual repeat sales price indices. This allows us to find the average house price changes by zip code for each year between 2006 and 2010 to the year 2010. We then calculate the average of these 5 price changes to determine the factor by which to adjust the boundaries for the price bins provided in the ACS data. Adjusting price boundaries by a zip code price index that looks at changes in median prices over time generates very similar adjustments.

constitute segment cut-offs, we use information from the search queries. This proceeds in two steps: First we change the price parameters set in the email alerts to account for the fact that we observe queries from the entire 2006 - 2012 period. This adjusts the price parameters in each alert by the market price movements of homes in that zip code between the time the query was set and the year 2010.[10] Second, we determine which set of three ACS cutoffs is most similar to the distribution of actual price boundaries selected in search queries that cover a particular zip code. For each possible combination of three (adjusted) price cut-offs from the list of ACS cut-offs, we calculate for every email alert the minimum of the absolute distance from each of the (adjusted) search alert price restrictions to the closest cut-off.[11] We select the set of segment price cut-offs that minimizes the average of this value across all queries that cover a particular zip code. This ensures, for example, that if there are many queries that include a high limit such as $1 million, $1 million is likely to also be a segment boundary.

To determine the total housing stock in each price by zip code segment, one additional adjustment is necessary. Since the ACS reports the total number of owner-occupied housing units, while we also observe market activity for non owner-occupied units, we need to adjust the ACS-reported housing stock for each price bin by the corresponding homeownership rate. To do this, we use data from all observed armslength ownership-changing transactions between 1994 and 2010 as reported in our deeds records. We first adjust the observed transaction price with the zip code level repeat sales price index, to assign each house for which we observe a transaction to one of our 2010 price (quality) bins. For each of these properties we also observe from the assessor data whether they were owner-occupied in 2010. This allows us to calculate the average homeonwerhisp rate for each price segment within a zip code, and adjust the ACS-reported stock accordingly.[12]

The other search dimension regularly specified in the email alerts, and that we hence

[10]This ensures that the homes selected by each query correspond to our 2010 quality segment definition. Imagine that prices fell by 50% on average between 2006 and 2010. This adjustment means that a query set in 2006 that restricts price to be between $500,000 and $800,000 will search for homes in the same quality segment as a query set in 2010 that restricts price to a $250,000 - $400,000 range.

[11]For example, imagine testings how good the the boundaries 100k, 300k and 1m fit for a particular zip code. A query with an upper bound of 500k has the closest absolute distance to a cut-off of $\min\{|500-100|, |500-300|, |500-1000|\} = 200$. A query with an upper bound of 750k has the closest absolute distance to a cut-off of 250. A query with a lower bound of 300k and an upper bound of 600k has the closest absolute distance to a cut-off of 0. For each possible set of price cut-offs, we calculate for every query the smallest absolute distance of a query limit to a cut-off, and then find the average across all search alerts.

[12]For example, the 2010 adjusted segment price cutoffs for zip code 94002 are $379,079, $710,775 and $947,699. This splits the zip code into 4 price buckets. The homeownerhip rate is much higher in the higher bucket (95%) than in the lowest bucket (65%). This shows the need to have a price-bucket specific adjustment for the homeownership rate to arrive at the correct segment housing stock.

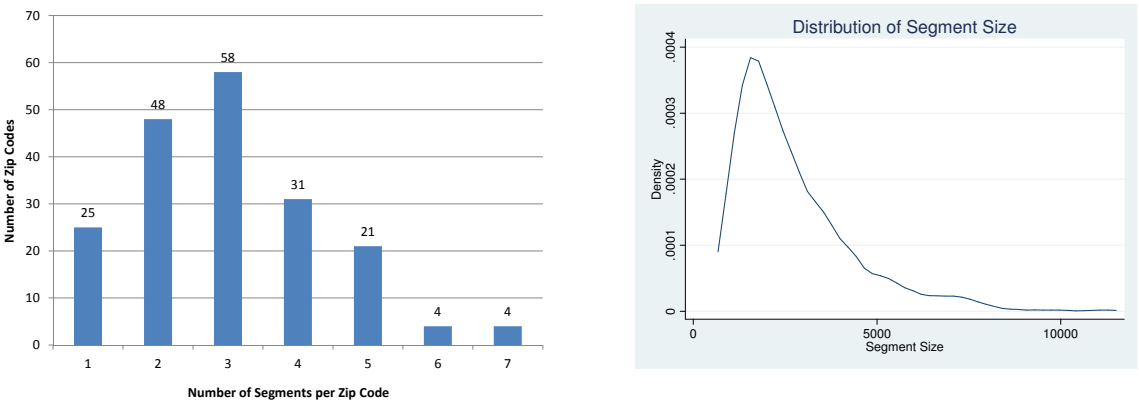
wanted to incorporate in our segment definition, is the number of bathrooms as a measure of the size of a house conditional on its location and quality. Since section 3.7 showed that the vast majority of constraints on the number of bathrooms selected homes with either more or fewer than two bathrooms, we further divide each zip code by price bucket group into two segments: homes with less than two bathrooms, and homes with at least two bathrooms. Unfortunately the ACS does not provide a cross-tabulation of the housing stock by home value and the number of bathrooms. To split the housing stock in each price and zip code segment into the two groups by home size, we apply a similar method as above to control for homeownership rate. We use the zip code level repeat sales price index to assign each home transacted between 1994 and 2010 to a 2010 price (quality) bin. For these homes we observe the number of bathrooms from the assessor records. This allows us to calculate the average number of bathrooms for transacted homes in each zip code by price segment. We use this share to split the total housing stock in those segments into two bathroom size groups.

The approach described above splits each zip code into eight initial segments along three price cutoffs and one size cutoff. For each of these segments, we have an estimate of the total housing stock. Since we need to measure specific moments such as the average time on market with some precision, we need to ensure that each segment is sufficiently large, and has a housing stock of at least 1,000 units. If this is not the case the segment is merged with a neighboring segment until all remaining segments have a housing stock of sufficient size. For price segments where either of the two size subsegments have a stock of less than 1,000, we merge the two size segments. We then begin with the lowest price segment, see whether it has a stock of less than 1,000, and merge it with the next higher price segment. This procedure generates 576 segments. Figure 19 shows how many segments each zip code is being split into. 25 zip codes are not split up further into segments. 48 zip codes are split into two segments, 58 zip codes are split into 3 segments. 418 segments only have a geography and price limitation, and include homes of all sizes falling into those price categories. The right panel of figure 19 shows the distribution of housing stock across segments. On average, segments have a stock of 2,717, with a median value of 2,271. The largest segment has a housing stock of 11,178.

A.4 Assigning segments to search alerts

As a next step, to analyze segments in terms of their search clientele we need to analyze each query and determine which segments are covered by that query. In section 3.2 we describe how we determine which zip codes are covered by each query. In this section we describe how

Figure 19: Segment Overview



we deal with the price and bathroom dimensions to determine the set of segments covered by each query. The challenge is that price ranges selected by queries will usually not overlap perfectly with the price cutoffs of the individual segments. For those queries that specify a price dimension, we assign a query to cover a particular segmented in one of three cases:

1. When the query completely covers the segment (that is, when the query lower bound is below the segment cutoff and the query upper bound is above the segment cutoff).

2. When the segment is open-ended (e.g. \$1 million +), and the upper bound of the query exceeds the lower bound (in this case, all queries with an upper bound in excess of \$1 million).

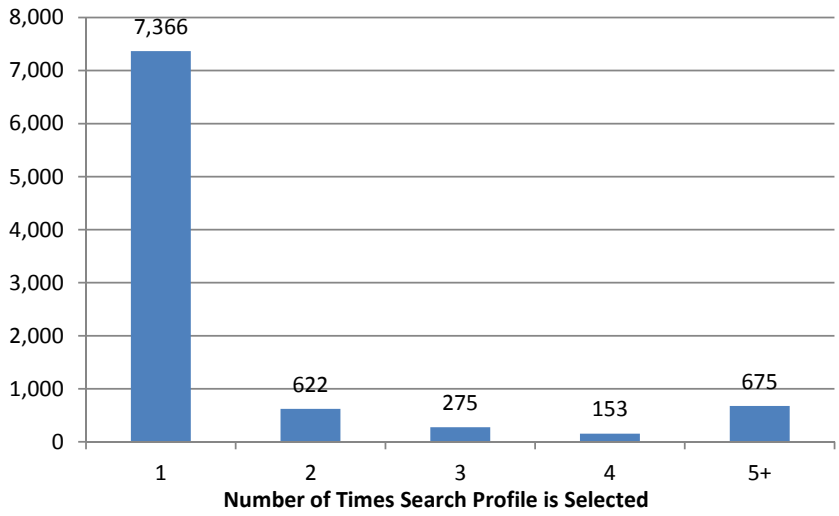
3. For queries that partially cover a non-open ended segment, we determine the share of the segment price range covered by the query. For example, for a segment \$300k - \$500k, the query 0-\$250k covers 25%, the query \$300k - \$700k covers 50% of the segment. We assign all queries that cover at least 50% of the price range of a segment to cover that segment.

To deal with the bathroom dimensions, we let a query cover a segment unless it is explicitly excluded. For example, queries that want at least two bathrooms will not cover the < 2 bathroom segments and vice versa.

We argued above that it was hard to pool across different alerts set by the same individual, since alerts differed along a number of key dimensions including geography, home size and home quality. The housing market segments constructed above allow us to pool all segments

selected by the same searcher. In particular, we begin by determining the subset of the segments that are covered by each individual search alert. This process described in more detail in Appendix A.4. After pooling all segments covered by at least one email alert set by each searcher, we arrive at a total of 9,091 unique search profiles. A total of 7,366 search profiles are selected by only a single user. Another 622 search profiles are selected twice. A total of 338 search profiles are selected more than 10 times each, with the two most commonly selected search profile showing up 1,017 and 416 times. Figure 20 shows the distribution of how often each search profile is selected. We also analyze the total housing stock covered by each searcher. We find that the average (median) searcher covers a total stock of 57,483 (33,807) housing units.

Figure 20: Number of Searchers per Unique Profile



Note: This figure shows how often each of the 9,090 individual search profiles is selected.

A.5 Construction of Segment Moments

Our model links the characteristics of search patterns to segment specific moments such as price, volume, time on market and inventory. In this section we describe how we construct these moments at the segment level. We begin by identifying a set of armslength transactions, which are defined as transactions in which both buyer and seller act in their best economic interest. This ensures that transaction prices reflect the market value (and hence the quality) of the property. We include all deeds that are one of the following: "Grant Deed,"

"Condominium Deed," "Individual Deed," "Warranty Deed," "Joint Tenancy Deed," "Special Warranty Deed," "Limited Warranty Deed" and "Corporation Deed." This excludes, for example, intra-family transfers. We drop all observations that are not a Main Deed or only transfer partial interest in a property (see Stroebel (2012) for details on this process of identifying armslength transactions).

We can then calculate the total number of transactions per segment between 2008 and 2011, and and use this to construct annual volume averages. In order to allocate houses to particular segments, we adjust transaction prices for houses sold in years other than 2010 by the same price index we used to adjust listing price boundaries (see appendix A.3). We arrive at our measure of Volume Share by dividing the annual transaction volume by the segment housing stock. Inventory levels are first constructed at the monthly level. To do this we use the dataset on all home listings on Trulia.com, beginning in January 2006. We assign each listed property to a segment using its location, size and adjusted listing price. Each month we add all newly listed properties in a segment to the inventory observed in the previous month. In addition, all listings that result in a sale as observed in the deeds data get removed from the inventory. We then construct the average of these monthly inventory levels for the period 2008-2011.[13] Inventory Shares are determined by dividing inventory levels in a segment by the total housing stock in the segment. We also construct a second inventory measure, "cold inventory", which is the fraction of the housing stock that is listed, and has been on the market for more than 30 days. In constructing inventory measures, one empirical challenge is that we do not observe when listings that do not result in a sale get removed from the market. We remove all listings for which we do not observe a sale from the inventory 270 days after the initial listings (as a reference point, note that the 90th percentile of time on market for houses that do eventually get sold is about 190 days). Of course, if a house sells that was listed for more than 270 days, we record that as a sale. A second challenge in measuring inventory levels arises from the fact that Trulia's coverage of listings is not 100% (for example, there are properties that are "for sale by owner" and hence do not show up in MLS feeds), and has increased over the time period we consider. However, we do have the universe of all transactions - this allows to construct, for every segment, a measure of how many homes we observe transacting over the sample period without having

[13]Many properties that are sold as REO resales (i.e. mortgage lenders selling properties that are acquired through a foreclosure) do not get listed through an MLS, and hence do not show up in Trulia's listing database. We thus need to construct REO resale inventory in a different way. In the deeds data we observe when a foreclosure occurs, since a foreclosure involves an ownership transfer to the bank. For those REO properties that do show up in the listings data, we calculate the median time between the foreclosure and the listing, which is 20 days. We henceforth add every foreclosed property to the inventory 20 days after we observe the foreclosure, and remove it when we observe an REO resale.

previously observed a listing. We can then scale our measure of inventory by the "share sale without listing" measure for that particular segment.

To calculate the average time on market, we match home listings in the listings database with final transactions from the deeds database.[14] We find segment-specific measures of time on market by averaging the time on market across all transactions that sold between 2008 and 2011. We also calculate the average time on market conditional on the time on market exceeding 30 days, which will be used in our stock-flow model of the matching process. Finally, we calculate the share of "hot sales", i.e. transactions that are recorded within 30 days of the initial listing.

[14] In the very few instances when the listing price and the final sales price would suggest a different segment membership for a particular house – i.e. cases where the house is close to being at a segment boundary and sells for a price different to the listing price, we allocate the house to the segment suggested by the sales price, not the listing price.