Unequal Impact of Zestimate on the Housing Market

Runshan Fu[†], Yan Huang[‡], Nitin Mehta[§], Param Vir Singh[‡], Kannan Srinivasan[‡*]

[†]New York University, [‡]Carnegie Mellon University, [§]University of Toronto

[†]runshan@nyu.edu, [‡]{yanhuang, psidhu, kannans}@cmu.edu, [§]Nitin.mehta@Rotman.Utoronto.Ca

We study the impact of Zillow's Zestimate on housing market outcomes and how the impact differs across socio-economic segments. Zestimate is produced by a Machine Learning algorithm using large amounts of data and aims to predict a home's market value at any time. Zestimate can potentially help market participants in the housing market as identifying the value of a home is a non-trivial task. However, inaccurate Zestimate could also lead to incorrect beliefs about property values and therefore suboptimal decisions, which would hinder the selling process. Meanwhile, Zestimate tends to be systematically more accurate for rich neighborhoods than poor neighborhoods, raising concerns that the benefits of Zestimate may accrue largely to the rich, which could widen socio-economic inequality. Using data on Zestimate and housing sales in the United States, we show that Zestimate overall benefits the housing market, as on average it increases both buyer surplus and seller profit. This is primarily because its uncertainty reduction effect allows sellers to be more patient and set higher reservation prices to wait for buyers who truly value the properties, which improves seller-buyer match quality. Moreover, Zestimate actually reduces socio-economic inequality, as our results reveal that both rich and poor neighborhoods benefit from Zestimate but the poor neighborhoods benefit more. This is because poor neighborhoods face greater prior uncertainty and therefore would benefit more from new signals.

Key words: Algorithms; Social impact; Economics of machine Learning; Housing markets

1. Introduction

Machine Learning (ML) algorithms are now used in our daily life to facilitate important decision making. They impact our lives in myriad ways including access to credit, education, jobs and various other areas (Kleinberg et al. 2018, Agrawal et al. 2019, Lambrecht and Tucker 2019, Hansen

* Runshan Fu is the first author; the other four authors have equal contribution and are listed in alphabetical order.

et al. 2021, Calvano et al. 2020, Fu et al. 2021, Zhang et al. 2021). A major strength of ML algorithms is their capability of analyzing large amount of data quickly and identifying patterns that humans may not be able to identify easily. Therefore, they tend to makes better predictions than humans in many applications (Kleinberg et al. 2018, Fu et al. 2021).

The goal of this paper is to study how a popular ML pricing algorithm affects the housing market. Housing is usually the key component of household wealth and the major collateral for bank lending, and has the most significant long term impact on wealth to income ratios (Piketty and Zucman 2014). Despite its importance, the housing market is often considered inefficient for several reasons. First, houses are heterogeneous assets. Each house is unique in its features (e.g., structure, floor plan, and build quality), amenities, and locations. Second, even for the same property, buyers are heterogeneous in their valuations. This makes it difficult to define and measure the value of a property. Third, there are non-trivial market frictions, such as search costs and various transaction costs, including agent fees, taxes, and moving costs. While realtors are usually more knowledgeable and could help inexperienced buyers and sellers, these factors still make accurately pricing a property difficult, leading to a housing market with large uncertainty.

The emergence of algorithms that predict market value of properties have the potential to reduce uncertainty in a housing market (Yu 2020). Several online real-estate marketplaces have proprietary algorithms that estimate property values. These marketplaces display their estimates for free on their websites. Utilizing large amounts of data on millions of properties and massive computational power, these sophisticated algorithms are able to produce reasonable estimates of property values. In this paper, we focus on the impact of Zillow's pricing algorithm. Zillow is the most popular real estate website in the United States by number of visits, with approximately 36 million unique monthly visitors,[1] or 27.2% of the market share of visits.[2] It was also one of the first websites to publish algorithm-generated property value estimates for properties nationwide. These estimates are called "Zestimates" and are available for more than 100 million U.S. homes.

[1] https://www.statista.com/statistics/381468/most-popular-real-estate-websites-by-monthly-visits-usa/

[2] https://ipropertymanagement.com/research/zillow-statistics

According to Zillow, for on-market properties, 82.2% of Zestimates are within 5% of selling price, 95.1% are within 10% of selling price, and 98.8% are within 20% of selling price.[3] Since these estimates that signal property values are easily available to buyers and sellers alike, they should work to reduce uncertainty in the housing market and, consequently, make it more efficient.

However, uncertainty reduction is only part of the potential effects. As it is difficult to achieve 100% accuracy, these algorithms generate estimates of property values with errors, and they could potentially shift buyers' and sellers' beliefs about property values in the direction of these estimates. While it is common for signals to contain noise and we would expect the bias in beliefs about property values to shrink as one receives more and more independent signals, the problem with these property value estimates as signals is that the estimates in one period are highly correlated with the estimates in the previous periods, especially if the estimates are generated from the same algorithm without major updates. Thus, the shift or bias in beliefs is likely to persist over time.

Both undervalued and overvalued Zestimate could be problematic, as an undervalued Zestimate could lead to lower belief in the property value and may result in a lower selling price, while an overvalued Zestimate could lead to incorrectly high expectations of the property value and result in longer selling time. There have been multiple news articles and online discussions on the inaccuracy of these algorithms and how they cause issues for both buyers and sellers in the market.[4] Thus, despite the fact that Zestimate could reduce uncertainty in the housing market, it is unclear how Zestimate affects buyer surplus and seller profit. This motivates our first objective, which is to examine how Zestimate affects the housing market in terms of market outcomes, including listing price, sales price, time on market, buyer surplus and seller profit.

The second objective of this paper is to examine how the impact of Zestimate differs across neighborhoods. Arguably housing matters more for the poor than for the rich, as studies have

[3] https://www.zillow.com/z/zestimate/, accessed on June 10, 2022.

[4] For a few examples, see:
https://www.courthousenews.com/wp-content/uploads/2017/04/Zillow.pdf,
https://www.nytimes.com/2018/09/14/realestate/why- zillow-addicts-cant-look-away.html,
https://www.reddit.com/r/RealEstate/comments/af414e/how_accurate_is_the_zestimate_number_do_you_put_a/

shown that the share of income spent on housing among homeowners steadily increases as we move from the highest income group to the lowest income group.[5] However, we notice in our analysis that Zestimate is less accurate in poor neighborhoods compared to in richer neighborhoods. A less accurate Zestimate provides a noisier signal and reduces uncertainty to a lesser extent. Thus, Zestimate could potentially benefit poor neighborhoods less than other neighborhoods. This particular observation of Zestimate being less accurate in poor neighborhoods motivates us to study the disparate impact of Zestimate across socio-economic segments.

We address these two objectives as follows. Recall that Zestimate has two potential effects: first, it reduces uncertainty about a property's value; second, it shifts the belief about a property's value. To tease out these two effects and evaluate buyer surplus and seller profit under counterfactual scenarios, we build a structural model of the housing market. Sellers and buyers are uncertain about property values and make decisions based on their beliefs about property values. Zestimates provide signals of property values and update these beliefs under the Bayesian learning framework, which captures the mean shift (in beliefs) effect and uncertainty reduction effect.

We estimate the model using data on properties in Pittsburgh that were listed between February and October 2019. There are 4,023 properties in total across 140 neighborhoods. Our estimation results show that the average information to noise ratio of Zestimate [6] is 0.2906, which suggests that Zestimate has non-trivial effect on buyers' and sellers' beliefs about property values. In addition, the variance of the prior belief about property values in the absence of Zestimate is larger in poor neighborhoods than in other neighborhoods, which suggests that people in poor neighborhoods face greater uncertainty about property value without Zestimate, and therefore could potentially benefit more from an accurate signal of property value.

To examine the effect of Zestimate, in our first counterfactual analysis, we simulate the market outcomes and calculate buyer surplus and seller profit when Zestimate is removed, and compare

[5] https://www.apartmentlist.com/research/housing-markets-and-income-inequality

[6] The variance of buyers' and seller's prior beliefs about property values divided by the variance of the signals of property values provided by Zestimate

them with those in the case where Zestimate is present. To examine how much poor neighborhoods are losing because of the less accurate Zestimate, in our second counterfactual analysis, we increase the Zestimate accuracy in poor neighborhoods up to the Zestimate accuracy in rich neighborhoods, and compare the simulated outcomes with the ones under the current Zestimate values.

We find that Zestimate overall benefits both buyers and sellers in the housing market. On average, it leads 5.35% increase in buyer surplus and 4.16% increase in seller profits. Interestingly, the sellers of 58% of the properties with "undervaluing" Zestimate (i.e., Zestimate that is lower than buyers and sellers' original beliefs about property values before they learn from Zestimates) and the buyers of 63% of the properties with "overvaluing" Zestimate still benefit from Zestimate. The benefit of Zestimate mainly comes from the uncertainty reduction effect of Zestimate that improves seller-buyer match quality. When Zestimate reduces uncertainty, buyers are less affected by the negative signals of a property staying on the market, thus sellers can be more patient to stay on the market and wait for buyers who truly value the properties. This improves match quality, and therefore benefits both buyers and sellers. While the belief shift effect means that "undervaluing" Zestimate would lower buyers' and seller's beliefs about property values and lead to lower expected selling prices, and "overvaluing" Zestimate would increase buyers' expectation about property values and result in wasted buyer visits and longer time on the market, the uncertainty reduction effect dominates the belief shift effect in most of the cases.

Second, we find that Zestimate actually benefits poor neighborhoods more than rich neighborhoods, despite of being less accurate in poor neighborhoods. The average seller profit increase by Zestimate in poor neighborhoods is 4.96%, compared to 3.74% and 4.02% in mid-range and rich neighborhoods; the average buyer surplus increase by Zestimate in poor neighborhoods is 8.27%, compared to 3.16% and 6.01% in mid-range and rich neighborhoods. The reason why Zestimate benefits poor neighborhoods more than rich neighborhoods is because buyers and sellers in poor neighborhoods face greater prior uncertainty before they learn from Zestimate, meaning that buyers and sellers in poor neighborhoods are more uncertain about the market than those in other

neighborhoods to start with. As a result, they would benefit more from additional signals, and even a noisier signal could be more helpful. Additionally, if on average Zestimates in poor neighborhoods were as accurate as Zestimates in rich neighborhoods, then the average total surplus increase would be 6.62%. Compared to the total surplus gain of 5.05% with the current Zestimate accuracy in poor neighborhoods, the potential positive impact of Zestimate on total surplus in poor neighborhoods could further increase by 31.09% with higher accuracy.

In the next section, we review the related literature. Section 3 describes the data and the model-free evidence. Section 4 presents our structural model, and Section 5 describes the estimation process and results. Counterfactual analysis is presented in Section 6. Section 7 concludes.

2. Related Literature

This paper is related to four streams of literature. The first is the literature on ML and human decision making. As ML algorithms are increasingly used to facilitate decision making, an important question to ask is whether algorithms can improve human decisions. Kleinberg et al. (2018) showed that a ML model can improve bail decisions, as the model can reduce crime rates with no change in jailing rates or reduce jailing rates with no change in crime rates, and these benefits can be achieved with reduced racial disparities. Similar results have been reported in the context of crowd lending (Fu et al. 2021) and resume screening (Cowgill 2018). This paper adds to the literature by showing that a popular pricing algorithm can benefit the housing market and increase both seller profit and buyer surplus.

The second is the literature on the impact of ML algorithms on social inequality. Arising with the popularity of ML algorithms is the concern about the potential disparate impacts these algorithms may create. Previous literature has shown how and why some popular algorithms leave disparate impacts across different race or gender groups, including Airbnb's smart-pricing algorithm (Zhang et al. 2021), Facebook advertisement targeting algorithm (Lambrecht and Tucker 2019), and Google search algorithm (Lambrecht and Tucker 2020). This paper adds to the literature by showing that Zestimate may reduce socio-economic inequality as it leads to highest surplus increase in poor neighborhoods, despite of being least accurate in poor neighborhoods.

The third is the literature on the micro-structure of housing market. This body of literature focuses on modeling the home selling process. Broadly speaking, there are two types of models. One is the search models in which buyers and sellers are searching for and matching with each other, and these models typically have one buyer meet with one seller in each period (Carrillo 2012, Chen and Rosenthal 1996b,a, Merlo et al. 2015). Another is the auction models in which multiple buyers may come in the same period and compete for the same house, and this process usually leads to bidding wars (Quan 2002, Han and Strange 2014). In a hot market, bidding wars are common, especially in popular cities, but during a normal time in a normal city, such as the sample we use, the selling process of most of the transactions is better characterized by a search model. Therefore, we focus on a search model in this paper, and we will explain how our model differs from the search models used in the previous literature in Section 4.

The last is the literature on racial differentials in housing markets. Several papers have shown that blacks pay higher price than whites for equivalent units in housing markets (King and Mieszkowski 1973, Ihlanfeldt and Mayock 2009, Straszheim 1974). Lu (2019) and Yu (2021) show that Zestimate might reduce racial disparities in housing market by providing less biased information. This paper adds to the literature by showing that Zestimate might reduce inequality because people in poor neighborhoods face the greatest uncertainty and therefore could benefit most from a pricing algorithm, yet they are still missing out some of the benefit due to the less accurate estimates.

3. Data and Model-free Evidence

3.1. Data

In this paper, we focus on on-market properties in Pittsburgh, Pennsylvania. Our sample consists of properties that were listed between February and October 2019 in Pittsburgh. There are 4,023 such properties in total and they are spread across 140 different neighborhoods. Each property has an individual page on Zillow, where Zestimate is displayed on the top along with some other property important information, including listing price, address, and size of the property, as shown in Figure 1. In addition, there is a section on the property page that shows more detailed

information about Zestimate, including Zestimate range and Zestimate history, as shown in Figure 2. Zestimate is the estimated market value (i.e., selling price) of a property, while Zestimate range is the range in which the selling price is predicted to fall. According to Zillow, a wider Zestimate range "generally indicates a more uncertain Zestimate, which might be the result of unique home factors or less data available for the region or that particular home".[7]

Since housing market condition keeps changing, the market value of a property also changes over time. Thus, in addition to reporting the current Zestimate and the Zestimate Range, Zillow also reports the "Zestimate history" of a property. Zestimate history shows Zestimate values for a property in historical time points. For example, Figure 2 shows that the estimated market value of the property back in October 2020 is about $383.2K, while the "current" (as of June 2021 when the screenshot was taken) estimated market value of the property is $481.1K. The line plot shows the trend of Zestimate values of this property, and by placing the cursor on different parts of the plot, users can read the historical Zestimate in each month in the last 10 years.

Note that these "historical Zestimates" are not necessarily the Zestimates that were displayed in the past. When Zillow implements major Zestimate algorithm updates, it will recalculates historical values using the updated algorithms retroactively. In other words, these historical Zestimates are the estimated market values in the past based on the current Zestimate algorithm. While they are calculated based on the most updated algorithm, Zillow does not allow future information to influence a historical Zestimate, thus a sales in 2020 does not affect a historical Zestimate in 2019.

For each property, we observe its Zestimate, Zestimate range and Zestimate history at roughly two-week intervals for the entire period when the property was on the market. We also observe its on-market activities, including listing time, initial listing price, listing price updates, selling time and final selling price. In addition, we observe a rich set of property features, including detailed location information, property structure, size, number of bedrooms, number of bathrooms, flooring, appliances, roof type, parking, basement and many other home characteristics. Table 1 shows the summary statistics of our data.

[7] https://www.zillow.com/z/zestimate/

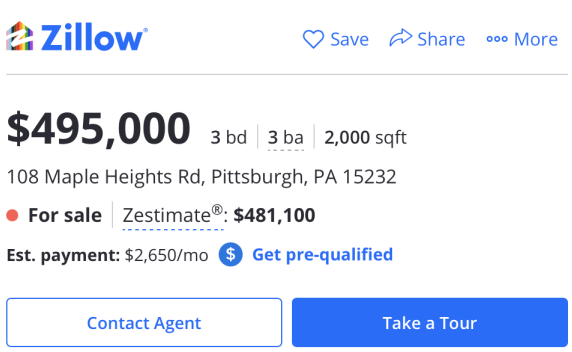

Figure 1 **Screenshot of the top of a property page on Zillow**

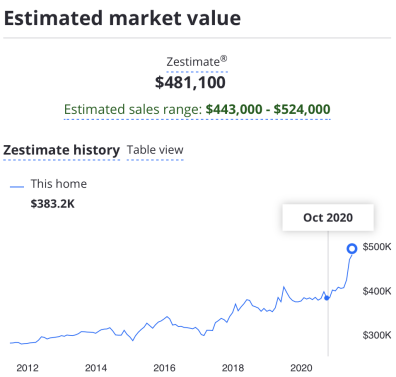

Figure 2 **Screenshot of the Zestimate hisotry section of a property page on Zillow**

Table 1 **Summary Statistics**

Variable	Mean	Std. Dev	Min	Median	Max
(Initial) Listing price	251,707.04	147,763.01	16,000	210,000	995,000
selling price	235,983.44	136,551.06	16,000	195,000	965,000
Time On Market (in days)	51.03	62.17	1	26	300
Number of listing price update *	1.91	1.23	1	1	10
Ratio of selling price over last listing price	0.9674	0.0517	0.5948	0.9777	1
Ratio of selling price over initial listing price	0.9369	0.0789	0.4837	0.9607	1
Proportion of selling price = listing price	0.3541				
Proportion with listing price update	0.3640				
Zestimate at listing time	242,678.49	141,285.87	20,631	200,633	996,126
Ratio of Zestimate over selling price	1.0440	0.1270	0.4037	1.0190	2.0409
Ratio of Zestimate over initial listing price	0.97020	0.0734	0.3104	0.9779	1.6722
Neighborhood Statistics	Poor	Mid-range	Rich		
Average selling price	140,551.39	215,101.98	376,210.21		
Average Zestimate at listing time	143,091.01	223,110.06	390,556.85		
Average normalized ZestRange	0.1525	0.1256	0.1305		
Average abs(ZestDev)	0.1088	0.0802	0.0743		

* among properties with price updates

3.2. Disparate Zestimate Accuracy

To examine the potential disparate impact of Zestimate across neighborhoods, we divide the 140 neighborhoods in our sample into three groups – poor neighborhoods, mid-range neighborhoods and rich neighborhoods – based on median Zestimate value in each neighborhood (including both

on-market and off-market properties).[8] Specifically, the neighborhoods whose median Zestimate is less than the 130 thousands (25 percentile of Zestimates of all properties) are considered as poor neighborhoods, the neighborhoods whose median Zestimate is greater than the 280 thousands (75 percentile of Zestimates of all properties) are considered as rich neighborhoods, and other neighborhoods are considered as mid-range neighborhoods.

We then calculate the normalized Zestimate Range for each property. Normalized Zestimate Range is calculated as the difference between the high value and the low value in a Zestimate Range, divided by the Zestimate value of the same property. As mentioned before, a Zestimate Range indicates the uncertainty in the corresponding Zestimate, and the normalization makes the measure comparable across properties with different value scales. We find that the average normalized Zestimate range is larger in poor neighborhoods compared to in mid-range and rich neighborhoods, suggesting that Zestimate is less precise in poor neighborhoods.

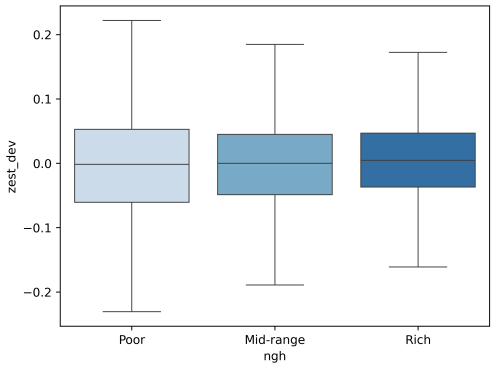

Figure 3 The distributions of Zestimate Deviation (from selling price) in percentage across neighborhoods

This is further confirmed when we compare Zestimates to final selling prices. Zillow measures Zestimate accuracy by the percentage difference between the Zestimate and the selling price of the same property,[9] and we follow Zillow's approach to calculate this Zestimate Deviation (from selling price) for each property in our sample. Figure 3 shows the distributions of Zestimate Deviation in different neighborhoods. We can see that Zestimate Deviation has the largest spread in poor neighborhoods, suggesting that Zestimate is less accurate in poor neighborhoods.

[8] We tried other measures, including median selling price and median listing price, and obtained very similar results.

[9] Note that selling price is endogenous and may be affected by Zestimate, therefore this is a rough measures of Zestimate accuracy. We will employ other measure of Zestimate accuracy later in the paper.

3.3. Algorithm updates

Zestimate has been available for individual properties since 2006 when Zillow was launched. Over the years, Zillow has been working on improving the algorithm to generate more accurate Zestimate. According the published Zestimate accuracy statistics, the median Zestimate Deviation (from the final selling price) has decreased from 6.9% in 2014 to 2.0% in 2019.[10]

During our observational period, there were two major algorithm updates, one in April 2019 and the other in October 2019. These algorithm updates incorporate the use of more unstructured information (e.g. image and video data) and improved Machine Learning techniques, which increase the overall prediction accuracy. The updates suddenly changed Zestimate values and Zestimate Ranges significantly for most of the properties without the anticipation from buyers or sellers. Moreover, these updates also changed historical Zestimate values, providing more accurate estimates of property market values in the past.

3.4. Descriptive Evidence

In this section, we provide evidence of how Zestimate affects the observed market outcomes, including listing price, selling price and time on market (TOM).[11] The main challenge in this set of analysis is that we do not observe the "true market value" of a property, which could be a major confounding factor. For example, if we observe that Zestimate is positively correlated with selling price, we do not know if it is because Zestimate has positive impact on selling price, or just because "true market value" is positively correlated with both Zestimate and selling price.

We overcome this challenge by leveraging the algorithm updates. As mentioned before, the algorithm updates changed not only the current Zestimate, but also the historical Zestimate values. That is, upon updating the algorithm, Zillow also recalculated the estimates of property values in the past, using the data available at times in the past and the updated (more accurate) algorithm. Therefore, for any time point before an algorithm update, we observe two different Zestimate

[10] https://web.archive.org/web/*/https://www.zillow.com/zestimate/

[11] Time on market (TOM) is the time that a property stays on the market. For sold properties, this is the time difference between the selling date and the listing date; for withdrawn properties, this is the time difference between the withdrawn date and the listing date.

values: the original Zestimate that was shown at that time, and the updated Zestimate that became available with the launch of a new algorithm. The updated Zestimate is a more accurate estimate of market value at that time, and it would not affect any event that happened before the algorithm update since it became available only after the algorithm update. Therefore, we use updated Zestimate based on the algorithm launched in October 2019 at the time when a property was listed as a proxy for the true market value of the property, and we calculate the Zestimate error for each property at the time when it was listed as follows:

$$\text{ZestError} = \frac{\text{Original Zestimate} - \text{Updated Zestimate}}{\text{Updated Zestimate}}. \tag{1}$$

This Zestimate error measures the extent to which the displayed Zestimate value deviated from (a proxy of) the true market value. To estimate the impact of Zestimate on market outcomes, we use this Zestimate error instead of the Zestimate value as the regressor to address the concern that the Zestimate value is correlated with true market value, which affects market outcomes.

It is worth emphasizing that when Zillow retroactively updates historical Zestimate with a new algorithm, it uses only information available at the historical time.[12] This means that the updated Zestimate at the listing time would not be directly affected by the final selling price of a property which became available later. However, there could still be concerns that the updated Zestimate at the listing time could potentially be indirectly affected by the final selling price through the updated algorithm, as the final selling price could influence some parameters in the updated algorithm. If this was the case, then the final selling price and the Zestimate error should be negatively correlated by construction, yet our results suggest otherwise. In other words, this concern would strengthen our results shown below.

We examine how Zestimate error affects the market outcomes.[13] Table 2 shows the regression results. In all the regressions, we control for the true property value using the updated Zestimate,

[12] From Zestimate Q&A: "(W)e never allow future information to influence a historical Zestimate (for example, a sale in 2019 could not influence a 2018 Zestimate). Historical Zestimates only use information known prior to the date of that Zestimate." https://www.zillow.com/z/zestimate/

[13] To avoid the results being affected by the extreme values, in the regression analysis we removed properties whose ZestError is greater than 2, or ZestError is less than 0.5, or ZestRange is greater than 0.8.

a rich set of home facts (e.g. number of bedrooms, number of bathrooms, parking, flooring and lot size) and neighborhood fixed effects. We use negative binomial regression when dependent variable is Time On Market (TOM; Column 3), logistic regression when dependent variable is the binary indicator of whether the listing price has ever been updated (PriceUpdate; Column 4) or the binary indicator of whether the selling price is equal to the listing price at the time of selling (SoldAtListPrice; Columns 5), and linear regression for other dependent variables. Since each property is only sold once and we only observe one set of market outcomes for it, we rely cross-sectional variation in ZestError to estimate its impact on the dependent variables.

Columns 1 - 3 in Table 2 show that as Zestimate error increases (which implies higher Zestimate conditional on property value), the initial listing price is higher, the final selling price is higher, and the property stays on the market for longer time. We next examine the effects of Zestimate error on the probability of adjusting listing price while the property is on the market and on the probability of being sold at the listing price. Since more than 98% of the listing price adjustment in our sample is downward, the price adjustment is a signal of the initial listing price being too high. Column 4 shows that as Zestimate error increases, the property is more likely to have downward listing price adjustment, suggesting that the initial listing price is likely to be overly high. In contrast, a property being sold at the listing price suggests that the buyer find the listing price reasonable and therefore is willing to purchase the property at the listing price without further bargaining. Column 5 shows that as Zestimate error increases, the property is less likely to be sold at its listing price (at the time of being sold), suggesting that buyers are less likely to accept the listing price and more likely to bargain with the seller for a lower price. We will discuss how these results conform to our structural model later in Section 6.2.

Next, we examine the effect of Zestimate uncertainty. While Zestimate error measures the deviation of a realized Zestimate value from a proxy of "true property value", Zestimate uncertainty refers to the noise level of Zestimate signals. As mentioned before, Zestimate Range indicates the uncertainty of Zestimate and provides information about the anticipated accuracy of Zestimate.

Table 2 Reduced-form Regression Results

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	ln(ListPrice)	ln(SoldPrice)	TOM	PriceUpdate	SoldAtListPrice	ln(ListPrice)	ln(SoldPrice)	ln(SoldPrice)
ZestError	0.355***	0.274***	0.567**	1.333**	-1.524**	0.548***	0.468***	0.537***
	(0.0171)	(0.0273)	(0.286)	(0.551)	(0.614)	(0.0287)	(0.0458)	(0.0594)
ZestRange						0.100***	-0.00413	-0.124***
						(0.0150)	(0.0242)	(0.0335)
ZestError*ZestRange						-0.686***	-0.855***	-1.150***
						(0.103)	(0.167)	(0.229)
ln(UpdatedZest)	0.980***	0.996***	0.519***	1.032***	-0.685	0.984***	0.993***	1.010***
	(0.00518)	(0.00836)	(0.0874)	(0.170)	(0.567)	(0.00517)	(0.00848)	(0.0103)
Constant	0.228***	0.152	-1.926*	-11.36***	2.472**	0.178***	0.195*	-0.00482
	(0.0688)	(0.110)	(1.161)	(2.256)	(0.972)	(0.0687)	(0.112)	(0.149)
Observations	3,963	3,704	3,963	3,930	3,972	3,963	3,704	2,065
R-squared	0.985	0.966				0.986	0.966	0.973
Home facts	YES	YES	YES	YES	YES	YES	YES	YES
Neighborhood	YES	YES	YES	YES	YES	YES	YES	YES

*** p<0.01, ** p<0.05, * p<0.1. Standard errors in parentheses.

Zestimate Range is defined by a low estimated value and a high estimated value, and we calculate the following value as a measure of Zestimate uncertainty: [14]

$$\text{ZestRange} = \ln(\text{High Estimated Value}) - \ln(\text{Low Estimated Value}). \tag{2}$$

Columns 6 and 7 show the impact of Zestimate uncertainty. First, as Zestimate range (uncertainty) increases, the effects of Zestimate error on listing price and selling price become weaker. This suggests that Zestimate range moderates the effect of Zestimate error. Second, a higher Zestimate range leads to higher listing price, but does not have significant impact on selling price. A significant fraction of the properties were sold at listing price in our sample. If we exclude these properties and focus on properties sold below listing price (column 8), then higher Zestimate range leads to lower selling price among these properties. These results suggest that when Zestimate uncertainty is higher, the listing price would be higher, while the bargained selling price when a

[14] We take the difference of the log values instead of the difference of the absolute values to address the scale problem: a range of $10,000 would be minimal for a million-dollar property, but would be substantial for a $100,000 property. We have tried an alternative measure, i.e., the difference between the High Estimated Value and the Low Estimated Value, divided by the Zestimate value, and obtained similar regression results. We keep the difference of log values here to be consistent with our structural model.

property is not sold at the listing price tends to be lower. Again, we will discuss how these results conform to our structural model later in Section 6.2.

The regression analysis provides evidence that Zestimate affects the initial listing price, the final selling price and the time on market. In addition, the effects of Zestimate are moderated by Zestimate uncertainty, and Zestimate uncertainty itself leads to higher listing price and lower selling price when a property is sold below listing price. Based on the evidence, we build a structural model of the housing market, which we present in the next section.

4. Model

As mentioned in Section 2, search models of housing market are usually used to describe transactions where selling prices are less than or equal to listing prices (Carrillo 2012, Chen and Rosenthal 1996a,b). In the data sample we use, the vast majority (more than 95%) of the transactions do not involve bidding wars and end up with selling prices less than or equal to listing prices. Therefore, we choose to use a search model. We adapt a basic search model from the previous literature that describes the home selling process, and introduce uncertainty and learning about property values in our model. We present our model in Section 4.1, and describe buyer decision and seller decision in Section 4.2 and Section 4.3, respectively. As we are interested in the potential disparate impact of Zestimate across neighborhoods, we allow neighborhood heterogeneity in a number of factors. To keep the presentation of the model clear, in Section 4 to 4.3, we describe how sellers and buyers make decisions under uncertainty in general (i.e., without neighborhood heterogeneity). We then explain how we incorporate the heterogeneity across different groups of neighborhoods (i.e., poor, mid-range and rich) into our model in Section 4.4.

4.1. The Main Model

The home-selling process is modeled as a discrete-time, infinite horizon problem, where each period spans 2 weeks.[15] As assumed in the previous literature that has used a search model, the

[15] We choose 2-weeks as the length of a period because most of the updates (e.g. listing price change and listing status change) happen within 2 weeks and the Zestimate-related data is collected bi-weekly.

housing market consists of risk-neutral,[16] infinitely lived agents. There are two types of agents: households that actively search for a property (buyers) and households who list a property for sale (sellers).

A seller, as a homeowner of property i, would derive certain utility from keeping his property (e.g. by living in the property or renting it out). We call such utility the "holding value" of property i for the seller, denoted as v_i^h, and it is the value that the seller preserves if he does not sell the property. When the seller lists his property, he sets a listing price. Recall that search models capture the scenario where the final selling price is less than or equal to the listing price. Therefore, the listing price is a commitment that the seller uses to attract buyers, as buyers know that the final selling price will be capped at the listing price. As per search models, the seller has a reservation price (referred to as "reservation price" hereafter for brevity), which is the lowest price that the seller is willing to accept. Both the listing price and the reservation price can be changing during the course of selling. While the property stays on the market, the seller can update the listing price and the reservation price at the beginning of each period. We denote the listing price of property i at time t as L_{it} and the reservation price as R_{it}, and describe how the seller determines L_{it} and R_{it} in Section 4.3.

Potential buyers' valuation about the property may be heterogeneous; thus, for a given property, there is a distribution of buyer valuations. In each period t, with probability δ, a potential buyer arrives and views the listing of property i. We refer to this potential buyer as "the buyer" hereafter. The parameter δ captures the market thickness: a larger δ means more frequent buyer arrivals, which suggests a "thicker" market. From the listing, the buyer observes a set of home characteristics as well as the listing price L_{it}. As in the previous housing literature, the buyer does not know exactly how much she would value the property yet (since she has not visited and examined the property), but she knows that her own valuation would be a draw from the distribution of buyer valuations of this property i. We assume that the distribution of buyer valuations of property i is

[16] For a discussion about this assumption, please refer to Appendix A.

$$v_{it} \sim \mathcal{LN}(\lambda_i, \sigma_v^2), \tag{3}$$

where v_{it} denotes the valuation of property i for the buyer who arrives in period t, $\mathcal{LN}$ refers to log normal distribution, λ_i is the expected value of the natural logarithm of buyer valuations (referred to as "log buyer valuations" hereafter), which is property-specific, and σ_v^2 is the variance of buyer valuations' natural logarithm, which is shared across properties.[17]

To examine the role of Zestimate, we incorporate uncertainty and learning about property values in our model. Thus, unlike the previous literature that assumes perfect knowledge of this distribution of buyer valuations for both sellers and buyer, we assume that sellers and buyers know the variance of log buyer valuations, σ_v^2, but they are uncertain about the mean of log buyer valuations, λ_i,[18] which is determined by a rich set of property features (such as floor plans, build quality, and location) as well as how buyers in the market value these features. Therefore, sellers and buyers form beliefs about λ_i and may update their beliefs upon receiving new information from two possible sources: Zestimate and previous buyers' realized valuation (when the property is not sold). We will describe the beliefs and the update process in details later.

With the belief about λ_i (and therefore the belief about v_{it}), the buyer decides whether to visit the property or not (the mathematical formulation of the decision will be explained in Section 4.2). If she chooses to visit the property, then she would incur a visiting cost c_{it}, and her own valuation of the property (v_{it}) is revealed to her after her touring the property. With this realized v_{it}, she engages in a bargaining process with the seller. The property will be sold if and only if the buyer's valuation v_{it} is greater than the seller's reservation price R_{it}. Following Chen and Rosenthal (1996b), we assume a Nash bargaining (Nash 1950, Zhang and Chung 2020) between the seller and the buyer, and the bargaining solution is a weighted average of the buyer's valuation (v_{it}) and the seller's reservation price (R_{it}). If the bargaining solution is higher than the listing price

[17] The distribution of buyer valuations of a property could potentially change over time. Here, we assume it is time-invariant during the course of selling, which is usually a relatively short time period and the changes in property value are arguably negligible.

[18] For a detailed discussion about this assumption and the potential learning about the variance of log buyer valuations, please refer to Appendix C.

(L_{it}), then the selling price will be the listing price, since the listing price is a price commitment by the seller as the maximum price that will be charged. Thus, the final selling price for property i, denoted as p_i, is given as

$$p_i = \min\{\theta v_{it} + (1-\theta)R_{it}, L_{it}\}, \tag{4}$$

where θ is the bargaining power parameter. If no potential buyer arrives, or if a potential buyer arrives but does not choose to visit, or if the buyer visits but realizes a valuation for the property that is lower than the seller' reservation price, then no transaction happens and the seller stays on the market in the next period.

In what follows, we describe the beliefs about λ (the mean of the distribution of the log buyer valuation of property i) in details, including the formation of the prior beliefs as well as how the beliefs get updated because of Zestimate and realized buyer valuations. We decompose λ_i into three components. The first component is driven by features that buyers and sellers observe and we as researchers also observe, and this part is denoted as $\gamma \mathbf{X_i}$, where $\mathbf{X_i}$ is a vector of the observed features and γ is a vector of parameters. The second component is driven by features that buyers and sellers observe but we as researchers do not observe, and this component is modeled as a random variable, denoted as u_i.[19] The last component is the mean value of log buyer valuations that cannot be explained by $\mathbf{X_i}$ and u_i, and is what sellers and buyers are uncertain about. This component is denoted as α_i. While sellers and buyers observe a rich set of property features, they may still be uncertain about the current market condition or the market value of different features. We capture such uncertainty in α_i. Formally, λ_i can be written as

$$\lambda_i = \alpha_i + \gamma \mathbf{X_i} + u_i, \quad u_i \sim \mathcal{N}(0, \sigma_u^2), \tag{5}$$

where α_i captures the part of the property value that is unobserved by sellers and buyers (as well as the researchers).

[19] Here, we basically assume that buyers and sellers observe the same set of home characteristics and have symmetric information. In Appendix D, we relax this assumption and show using simulations that the main mechanisms of our findings remain qualitatively the same.

While the seller and the buyers do not know the true value of α_i for a given property i, they know the true distribution of α_i across all the properties, denoted as $\alpha_i \sim \mathcal{N}(\alpha_0, \sigma_0^2)$. We assume that they have rational expectations and use this true distribution ($\mathcal{N}(\alpha_0, \sigma_0^2)$) as the initial prior belief about α_i for all the properties.[20] This initial prior belief captures information from all the sources except Zestimate at the beginning of the listing, which included the information provided by the realtors, friends and family, as well as the sellers' and buyers' own estimation and calculation. Also note that buyers and sellers share the same initial prior belief, meaning that they are equally uncertain about the distribution of buyer valuations at $t = 1$. We denote the belief about α_i in period t as $\hat{\alpha}_{it}$, and for period $t = 1$ we have $\hat{\alpha}_{i1} \sim \mathcal{N}(\alpha_0, \sigma_0^2)$.

Since α_i is the only uncertain component of λ_i, the uncertainty about α_i is equivalent to the uncertainty about λ_i. For the ease of understanding, we will focus on the uncertainty about λ_i in the rest of the paper. We denote the belief about λ_i at the beginning of period t as $\hat{\lambda}_{it}^0 \sim \mathcal{N}(\mu_{it}^0, (\sigma_{it}^0)^2)$, and it is easy to see that

$$\mu_{i1}^0 = \alpha_0 + \beta \mathbf{X_i} + u_i, \quad \sigma_{i1}^0 = \sigma_0. \tag{6}$$

The seller and the buyers update their belief about λ_i based on signals from two sources: Zestimate and realized buyer valuations (when the property is not sold).

Zestimate is an algorithm generated estimate of a property' market value, or the expected selling price. We denote the Zestimate for property i in period t as Z_{it}. In our model, Zestimate provides a new signal about property value in the first period when a property is listed or when there is an Zestimate algorithm update.A property's Zestimate actually changes every few days, but such changes are mostly minor fluctuations. Such fluctuations hardly provide any new information about the property value, thus we do not treat them as new signals.[21] An exception is when the

[20] In Section 4.4, we will discuss how we allow the initial prior belief to be neighborhood-group specific.

[21] These micro adjustments in absence of major updates mostly reflect periodic re-estimates of weights on comparable homes, as prices of newly sold properties become available over time. They could also reflect minor updates of the algorithm, but we do not have information on when those minor updates happened. In any case, without major updates, the fluctuations in zestimate are very minor, which suggests that these zestimates will be very strongly correlated. In our analysis, we assume these zestimate signals for the same property in absence of major updates to be perfectly correlated. That is, one zestimate signal is as informative as multiple zestimate signals in absence of major updates.

Zestimate algorithm is updated. As mentioned in 3.3, there are two Zestimate algorithm updates in our observation period. Since the updated Zestimates are generated by new algorithms, we model Zestimates before and after the updates as independent signals. Note that the updated algorithms overall improve the accuracy of Zestimates, thus Zestimates before and after the updates are not identically distributed as they may not have different levels of noise.

The expected selling price depends on the mean of log buyer valuations (λ_i), which is what the seller and the buyer are uncertain about and trying to learn about. Since Zestimate is an estimate of the expected selling price, the seller and the buyers cannot directly use Zestimate to learn about λ_i. Instead, they infer an implied signal about λ_i from Zestimate (Z_{it}), and update their beliefs about λ_i with this implied signal. We denote this implied (unbiased) signal[22] about λ_i as ζ_{it}:

$$\zeta_{it} \sim \mathcal{N}(\lambda_i, (\sigma^z_{it})^2) \tag{7}$$

We use a log-linear function as a reduced-form approximation of what buyers and sellers believe to be the mapping from Z_{it} to ζ_{it}, where a_z, b_z are two parameters:

$$\zeta_{it} = a_z + b_z \ln(Z_{it}) \tag{8}$$

The standard deviation of Zestimate signal, σ^z_{it}, is not directly observed, but as mentioned previously, Zillow shows a "Zestimate Range" along with a Zestimate for each property, which describes "the range in which a sale price is predicted to fall, including low and high estimated values".[23] This Zestimate Range can be roughly viewed as a confidence interval, and it tells us how precise the Zestimate value is. Thus, we first calculate the range of log Zestimate in the interval as $ZR_{it} = \ln(Z^H_{it}) - \ln(Z^L_{it})$, where Z^L_{it} and Z^H_{it} denote the low estimated value and the high estimated value in a Zestimate Range, respectively. Then we use a linear function as a reduced-form approximation of what buyers and sellers believe to be the mapping from the range to σ^z_{it}:

$$\sigma^z_{it} = a_s + b_s ZR_{it}, \tag{9}$$

[22] Even if Zestimates are overall biased, rational buyers and sellers would be able to adjust for the bias when transforming Zestimates into signals about the λ_i, as they observe how much Zestimates overvalue or undervalue properties in general. This would be captured by the constant term a_z in Equation 8.

[23] https://www.zillow.com/z/zestimate/

where a_s, b_s are another two parameters.

If a new Zestimate signal is available in a period, then the seller and the potential buyer in this period update their belief from the Zestimate signal under the standard Bayesian learning framework; otherwise, the original belief remains. Formally, we denote the posterior belief about λ_i in period t after learning from the new Zestimate signal (if any) as

$$\tilde{\lambda}_{it} \sim \mathcal{N}(\mu_{it}, \sigma_{it}^2), \tag{10}$$

and the updating from a Zestimate signal follows

$$\mu_{it} = \begin{cases} \mu_{it}^0 + \frac{(\sigma_{it}^0)^2}{(\sigma_{it}^0)^2+(\sigma_{it}^z)^2}(\zeta_{it} - \mu_{it}^0), & \text{if } t = 1 \text{ or Zestimate algorithm updated;} \\ \mu_{it}^0, & \text{otherwise.} \end{cases} \tag{11}$$

$$\sigma_{it}^2 = \begin{cases} (\frac{1}{(\sigma_{it}^0)^2} + \frac{1}{(\sigma_{it}^z)^2})^{-1} & \text{if } t = 1 \text{ or Zestimate algorithm updated;} \\ (\sigma_{it}^0)^2, & \text{otherwise.} \end{cases} \tag{12}$$

It is based on this posterior belief $\tilde{\lambda}_{it}$ that buyers and sellers make their decisions in period t, and we will describe their decision processes in details in the next two sections.

The second source of signal is the realized buyer valuations when a property is not sold after buyer visits. Recall that if a buyer visits a property and realizes a valuation lower than the seller's reservation price, then no transaction would happen and the seller would stay on the market in the next period. Although the seller does not close the deal in this case, he and the future buyers can observe the realized buyer valuation (v_{it}), potentially through negotiation and information from realtors, and learn more about the mean of log buyer valuations (λ_i) from the realized buyer valuation (v_{it}).[24] The realized buyer valuation is a draw from the true distribution of buyer valuations (as shown by Equation 3) and its natural logarithm provides a signal about λ_i: $\ln(v_{it}) \sim \mathcal{N}(\lambda_i, (\sigma_v)^2)$. The learning from the realized buyer valuation happens at the end of a

[24] Even if the exact value of the realized buyer valuation is unknown to the seller and the future buyers, they could still easily know that a previous buyer visited but did not purchase the property, which would provide a signal that the realized buyer valuation was lower than the reservation price. Modeling such a signal would lead to mathematically more complicated but conceptually similar results as in our main model.

period. We denote the posterior belief about λ_i after learning from the realized buyer valuation (if any) as $\tilde{\lambda}'_{it} \sim \mathcal{N}(\mu'_{it}, (\sigma'_{it})^2)$ and we have the learning from the buyer's valuation as

$$\mu'_{it} = \begin{cases} \mu_{it} + \frac{(\sigma_{it})^2}{(\sigma_{it})^2 + (\sigma_v^n)^2}(\ln(v_{it}) - \mu_{it}), & \text{if a buyer visits and the property is unsold;} \\ \mu_{it}, & \text{otherwise.} \end{cases} \tag{13}$$

$$\sigma'_{it} = \begin{cases} (\frac{1}{(\sigma_{it})^2} + \frac{1}{(\sigma_v^n)^2})^{-1} & \text{if a buyer visits and the property is unsold;} \\ \sigma_{it}^2, & \text{otherwise.} \end{cases} \tag{14}$$

This belief about λ_i at the end of period t, $\tilde{\lambda}'_{it}$ carries over to the next period and becomes the prior belief at the beginning of period $t+1$, that is, $\tilde{\lambda}^0_{i(t+1)} = \tilde{\lambda}'_{it}$.

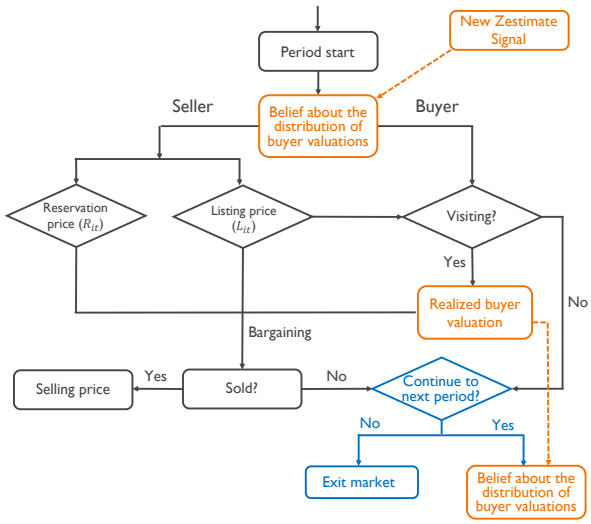

Figure 4 The structure of the main model

Figure 4 shows the overall structure of our main model. The sequence of the events is as follows: in each period, the seller first updates his belief about λ_i from Zestimate (in the first period or when there is an Zestimate algorithm update) and/or the realized buyer valuation from the previous period (if any), and then set or update the listing price and the reservation price. Next, a stochastic process determines whether a potential buyer arrives. If a potential buyer arrives, then the buyer, with the updated belief about λ_i, decides whether to visit or not. Since the seller and the buyers share the same initial prior belief about λ_i and observe the same of signals (both Zestimate and previous buyers' valuations), they have symmetric information sets at all times and always share

the same belief about λ_i.[25] If the buyer chooses to visit, then she and the seller decide the selling outcomes jointly. If the property is not sold at the end of the period, then the seller compares the expected profit in the next period and his holding value: the seller moves to the next period if the expected profit in the next period is higher, and exits the market otherwise. Note that in this model, both sellers and buyers make decisions based on their beliefs about the distribution of buyer valuations. Therefore, as Zestimate influences their beliefs about the distribution of buyer valuations, it affects the decisions on reservation price, listing price and buyer visiting, and these decisions further influence market outcomes, such as selling prices and selling time. In the next two section, we characterize buyer decisions and seller decisions.

4.2. Buyers' Problem

In each period, with probability δ, a potential buyer for property i arrives and updates her belief about λ_i with information of previous buyer valuations (if any)[26] and Zestimate signals. Note that in the search stage, the buyer does not know her own valuation of the property, but she knows that her valuation is drawn from the true distribution of buyer valuations. With the belief about λ_i after learning from all the previous signals from Zestimate and previous buyers' valuations (i.e., $\tilde{\lambda}_{it}$), the buyer effectively has a belief about the distribution of buyer valuations (including her own valuation), denoted as $\tilde{v}_{it}$. Combining Equation 3 and Equation 10, we have

$$\tilde{v}_{it} = \mathcal{LN}(\mu_{it}, \sigma_{it}^2 + \sigma_v^2) \tag{15}$$

With this belief and other available information, the buyer decides whether to visit the property or not. If she chooses to visit, there are three potential outcomes: First, she may purchase the property at the bargaining outcome (below the listing price); second, she may purchase the property at the listing price; last, she may not purchase the property and simply walk away.[27] The

[25] For a discussion about the case when buyers and sellers have asymmetric information, please refer to Appendix D.

[26] The buyer could obtain such information from the real estate agents or other sources about market information.

[27] We do not model the sequential search and assume that the buyer would exit the market if no purchase is made after the visit. This is common in housing market literature and primarily because of data limitation: we do not observe what properties a specific buyer considers or visits and in what sequence the buyer visits them. For a detailed discussion about this assumption, please refer to the Appendix B.

final outcome depends on her realized valuation (which she has a belief about before visiting), and regardless of the outcome, she would incur a visiting cost.

The buyer's visiting cost includes includes monetary cost, time cost, and opportunity cost associated with visiting a property. Since people usually search for properties that match with their wealth level, potential buyers of more valuable properties tend to have higher time cost and higher opportunity cost. Therefore, the search cost will be higher for higher valued properties. The initial median belief of buyer valuation, $\exp(\mu_{i1}^0)$, can be viewed as a measure of property value when comparing across properties. Thus, we model the search cost c_{it} as

$$c_{it} = \bar{c}_{it} \exp(\mu_{i1}^0), \quad \text{where} \quad \bar{c}_{it} \sim \mathcal{LN}(c_0, \sigma_c^2). \tag{16}$$

That is, the relative cost of visiting a property for different buyers is drawn from a log-normal distribution, and we make the visiting cost proportional to property values (measured by $\exp(\mu_{i1}^0)$), to account for higher search costs of higher valued properties. This proportional visiting cost is also in a similar spirit of the linear homogeneity assumption in Merlo et al. (2015). We face the same computational challenge of solving individual dynamic programming (DP) problems, and this assumption allows us to do similar normalization as described in Merlo et al. (2015). Following the prior literature, we assume that both sellers and buyers know the distribution of search costs, and each buyer additionally knows her own search cost.

We interpret the realized buyer valuation as the utility of owning the property for the buyer, and assume that the utility of buying the property at a price p_i for the buyer is $u_{it} = v_{it} - p_i$, where v_{it} is the realized buyer valuation. If the buyer does not purchase the property, she would obtain 0 utility. Thus, the buyer would purchase the property if and only if $v_{it} \geq p_i$, and from Equation (4), we know that the buyer will purchase the property at the listing price if

$$v_{it} > \frac{L_{it} - (1-\theta)R_{it}}{\theta} \equiv \bar{v}_{it}, \tag{17}$$

and purchase the property at the bargaining outcome $(\theta v_{it} + (1-\theta)R_{it})$ otherwise. From the

buyer's perspective, when she decides whether to visit the property, the utility of visiting is:

$$u(\mu_{it}, \sigma_{it}^2) = \int_{R_{it}}^{\bar{v}_{it}} (1-\theta) \cdot (\tilde{v}_{it} - R_{it}) f(\tilde{v}_{it}) \, \mathrm{d}\tilde{v}_{it} + \int_{\bar{v}_{it}}^{\infty} (\tilde{v}_{it} - L_{it}) f(\tilde{v}_{it}) \, \mathrm{d}\tilde{v}_{it} + 0 \cdot \Pr(\tilde{v}_{it} < R_{it}) - c_{it}, \tag{18}$$

where μ_{it} and σ_{it}^2 directly affect the buyer's belief about the distribution of buyer valuations (where her own valuation is drawn from), and $f(\tilde{v}_{it})$ is the probability density function (PDF) of this belief. On the right-hand-side, the first three terms correspond to the expected utility of the three possible outcomes described before, respectively; the last term is the visiting cost.

The buyer will choose to visit if and only if the utility of visiting is greater than 0. The seller does not know the visiting cost of each buyer, but knows the distribution of the visiting costs. Thus from the seller's perspective, the probability of a buyer visiting his property in period t is

$$q_{it} \equiv q(\mu_{it}, \sigma_{it}^2) = \Pr(c_{it} > \int_{R_{it}}^{\bar{v}_{it}} (1-\theta) \cdot (\tilde{v}_{it} - R_{it}) f(\tilde{v}_{it}) \, \mathrm{d}\tilde{v}_{it} + \int_{\bar{v}_{it}}^{\infty} (\tilde{v}_{it} - L_{it}) f(\tilde{v}_{it}) \, \mathrm{d}\tilde{v}_{it}) \tag{19}$$

4.3. Seller's Problem

We now move to the seller decisions. The seller decides the listing price L_{it} and the reservation price R_{it} in each period to maximize his ex ante expected profit. The seller's profit depends on the selling price, which is influenced by buyer valuations. The seller is uncertain about the distribution of buyer valuations, and like the buyers, he has a belief about the distribution of buyer valuations after learning from the previous realized buyer valuations and Zestimate signals:

$$\tilde{v}_{it} = \mathcal{LN}(\mu_{it}, \sigma_{it}^2 + (\sigma_v)^2). \tag{20}$$

With this belief, the seller makes decisions on the listing price and the reservation price. We denote the seller's ex ante expected (life-time) profit conditional on the belief about the distribution of buyer valuations $\tilde{v}_{it}$ as $\pi(\mu_{it}, \sigma_{it}^2)$, then we have

$$\begin{aligned} \pi(\mu_{it}, \sigma_{it}^2) = \max_{L_{it}, R_{it}} \{ & [(1-\delta) + \delta(1-q_{it})] \cdot \beta \cdot \pi(\mu_{it}, \sigma_{it}^2) + \delta q_{it} \int_{\bar{v}_{it}}^{\infty} L_{it} \cdot f(\tilde{v}_{it}) \, \mathrm{d}\tilde{v}_{it} \\ & + \delta q_{it} \int_{R_{it}}^{\bar{v}_{it}} [\theta \tilde{v}_{it} + (1-\theta) R_{it}] f(\tilde{v}_{it}) \, \mathrm{d}\tilde{v}_{it} + \delta q_{it} \cdot \int_{\infty}^{R_{it}} \max\{v_i^h, \beta \pi(\mu_{i(t+1)}(\tilde{v}_{it}), \sigma_{i(t+1)}^2)\} f(\tilde{v}_{it}) \, \mathrm{d}\tilde{v}_{it} \}, \end{aligned} \tag{21}$$

where β is sellers' discount factor, δ is the probability of buyer arrival in a period, q_{it} is the probability of a buyer choosing to visit (defined in Equation 19), $\bar{v}_{it}$ is the threshold for realized buyer valuations above which the property would be sold at the listing price (defined in Equation 17), and v_i^h is the holding value for the seller. $\mu_{i(t+1)}(\tilde{v}_{it})$ and $\sigma^2_{i(t+1)}$ are the mean and the variance of the updated belief about λ_i in the next period, following the updating process described in Section 4.1.

The four terms on the right-hand side of Equation 21 correspond to the four possible outcomes in a period: first, there is no buyer visit (including two sub-cases – no buyer arrives or a buyer arrives but chooses not to visit), and the seller stays on market until the next period; second, a buyer visits and the property is sold at the listing price; third, a buyer visits and the property is sold at the bargaining outcome (below the listing price); last, a buyer visits but the property is not sold, in which case the seller updates his belief about the distribution of buyer valuations and decides whether to exit the market (and obtains the holding value) or stay until the next period.

Note that the seller's holding value and the buyer valuations are closely related, as they are essentially valuations about the same property. Similar to the specification in Carrillo (2012), we assume that the seller's holding value is proportional to the initial median belief of buyer valuation:

$$v_i^h = \rho \exp(\mu_{i1}^0) \tag{22}$$

where ρ is the ratio parameter.

The seller's reservation price is the lowest price that he is willing to accept. Thus, the optimal reservation price should be the value at which he is indifferent between selling the property and moving into the next period or exiting the market. Thus, the optimal reservation price solves

$$\begin{aligned} R_{it} &= \mathbb{E}(\max\{v_i^h, \beta\pi(\mu_{i(t+1)}(\tilde{v}_{it}), \sigma^2_{i(t+1)})\} | \tilde{v}_{it} < R_{it})\} \\ &= \frac{\int_{\infty}^{R_{it}} \max\{v_i^h, \beta\pi(\mu_{i(t+1)}(\tilde{v}_{it}), \sigma^2_{i(t+1)})\} f(\tilde{v}_{it})\, \mathrm{d}\tilde{v}_{it}}{\Phi(\frac{\ln(R_{it}) - \mu_{it}}{\sigma_{it}})}. \end{aligned} \tag{23}$$

By First Order Condition, we know that that the optimal listing price satisfies

$$\frac{\partial \pi(\mu_{it}, \sigma^2_{it})}{\partial L_{it}} = 0. \tag{24}$$

After deciding on the optimal listing price and reservation price, the seller waits for buyer visits and sells the property if a buyer values the property more than the seller's reservation price. If no buyer visits or the buyer valuation is lower than the his reservation price, the seller will stay on the market if the continuation value is greater than the holding value, and exit the market otherwise. We denote the period in which the property is sold or withdrawn from the market as T_i.

4.4. Heterogeneity across Neighborhood Groups

So far, we have assumed that a common set of parameters are shared by properties in all neighborhoods. As we are interested in the potential disparate impacts of Zestimate across different neighborhoods, we extend the main model in the following ways to account for the heterogeneity in the housing market across poor, mid-range and rich neighborhoods.

First, we allow the variance of log buyer valuations (σ_v in Equation 3) to be different in different neighborhood groups to capture the potential difference in buyer heterogeneity across those groups. Specifically, we denote the variance of log buyer valuations in neighborhood group n as σ_v^n, where $n \in \{\text{P(oor)}, \text{M(id-range)}, \text{R(ich)}\}$. Note that both Zestimate and Zestimate Range are affected by heterogeneity in buyer valuations, and buyers and sellers would exclude the influence of buyer hetergeneity on Zestimate (or Zestimate Range) when they map Zestimate (or Zestimate Range) to the implied signal about λ_i (or standard deviation of Zestimate signals). If we assume all the properties share the same level of buyer heterogeneity, then the influence of buyer heterogeneity, along with many other factors, would be absorbed by the intercept terms (a_z and a_s) in the mappings (Equation 8 and Equation 9). As we allow buyer heterogeneity (σ_v) to be different in different neighborhood groups, we need to explicitly account for the exclusion of the influence of buyer heterogeneity in the two mappings. Therefore, the mapping from Zestimate (Z_{it}) to the implied signal (ζ_{it}) becomes

$$\zeta_{it} = a_z + b_z \ln(Z_{it}) - c_z \sigma_v^n, \tag{25}$$

and that from Zestimate Range (ZR_{it}) to the standard deviation of Zestimate signals (σ_{it}^z) becomes

$$\sigma_{it}^z = a_s + b_s ZR_{it} - c_s \sigma_v^n, \tag{26}$$

where c_z and c_s are two additional parameters.

Second, it is possible that different neighborhood groups have different level of market thickness. A thicker market with more participants would lead to different optimal strategies and market outcomes compared to a thinner market. To capture the potential difference in market thickness across neighborhood groups, we allow the probability of a potential buyer arriving in each period to be neighborhood-group specific, denoted as δ^n. Third, to reflect the potential differences in sellers' and buyers' knowledge and beliefs about properties' values when they enter the market and before they are exposed to Zestimate, we allow sellers and buyers in different neighborhood groups to form different initial prior beliefs about α_i. That is, $\alpha_i^n \sim \mathcal{N}(\alpha_0^n, (\sigma_0^n)^2)$. Last, we also allow ρ, the parameter that captures the ratio between the initial belief of median buyer valuation and the seller's holding value, to be neighborhood-group specific (denoted as ρ^n). This captures the potential difference in how sellers value their properties compared to potential buyers across different neighborhood groups.

5. Estimation

We estimate the model using the data described in Section 3.1. The parameters in our model include the following: the standard deviations of log buyer valuations (σ_v^n), the initial prior beliefs about α_i (α_0^n, σ_0^n); the distribution of log visiting costs (c_0, σ_c); the probabilities of a potential buyer arriving in a period (δ^n); the ratio between the initial belief about median buyer valuation and the seller's holding value (ρ^n); the coefficients of home characteristics γ; the relative bargaining power (θ); the parameters in the linear transformation of Zestimate to Zestimate signal (a_z, b_z, c_z); the parameters in the linear transformation of Zestimate range to Zestimate standard deviation (a_s, b_s, c_s); and the seller's discounting factor (β).

A few parameters cannot be identified in our data. First, we cannot identify the discounting factor β, because we do not have the natural exclusion restriction, i.e., there is no state variable that impacts the future payoffs but not the current payoffs. Therefore, we fix β to 0.995 in our estimation.[28] Second, the relative bargaining parameter θ cannot be identified. Since we directly

[28] This translates to a yearly discounting factor of 0.88, which implies higher discounting than the prevailing interest rates and makes it a conservative assumption. We have also tried other fixed values, and obtained similar estimation results.

observe neither buyer valuations nor sellers' reservation prices, we cannot distinguish between the case where the seller's bargaining power is high from the case where the buyer's valuation is high. To evaluate the relative bargaining power, we refer to the "buyer-seller index" on Zillow, which ranges from 0 to 10 and indicates the extent to which a market is a seller's market. The average buyer-seller index values in the poor, the mid-range, and the rich neighborhoods during our observational period are 9.1, 9.0 and 8.1, respectively. Thus we fix the bargaining parameter θ to be 0.91, 0.90 and 0.81 for the properties in these neighborhood groups in the estimation. Last, since we do not observe buyer arrivals and buyer visits, we cannot separately identify the probability of a potential buyer arriving in a period (δ_n) and the distribution of visiting costs (c_0, σ_c). Therefore, we fix the probability of a potential buyer arriving in the mid-range neighborhoods to be 1, and estimate the relative probabilities of buyer arrival in the other two neighborhood groups as well as the distribution of visiting costs.[29] We estimate the model using Maximum Likelihood Estimation. For the derivation of the likelihood function, we refer readers to Appendix E.

5.1. Identification

Here, we briefly explain how the parameters in our structural model can be identified. A detailed discussion on identification is provided in Appendix F.

The variance of buyer valuations (σ_v^n) is identified from the variation in selling prices across similar properties with similar listing price trajectories. The prior variance (σ_0^n) is identified from the listing price updates resulting from buyer visits. How much the listing price changes in response to the realized buyer valuations reflects the relative size of the prior variance and the variance of buyer valuations. With the variance of buyer valuation identified and the relative size of the prior variance and the variance of buyer valuations inferred from the changes in the listing prices when there is no algorithm update, the prior variance is identified.

The initial prior mean belief ($\mu_{i1}^0 = \alpha_0 + \gamma \mathbf{X} + u_i$) can be inferred from the initial listing price and the initial Zestimate for each property individually. If there are similar properties in the same

[29] The estimation result suggest that the mid-range neighborhoods have the thickest market (i.e., the largest arrival probability) compared to the other two groups.

neighborhood group, with same Zestimate and Zestimate range, but different listing prices, then the difference in listing prices can only be attributed to the difference in prior mean. With the inferred initial prior mean beliefs (μ_{it}^0) and the observed home characteristics ($\mathbf{X}$), α_0 and γ can be identified in a similar manner to ordinary least squares (OLS) in a linear regression.

The Zestimate-related parameters, i.e., the 6 parameters governing the transformation from log Zestimate to Zestimate signal (a_z, b_z, c_z), and the linear transformation from Zestimate Range to standard deviation of Zestimate signal (a_s, b_s, c_s), are jointly identified from both cross-sectional and intertemporal variation of the listing price in response to varying Zestimate values and varying Zestimate Range values. With the continuation values inferred from listing prices with other parameters, the parameter of holding values (ρ^n) is identified with the withdraw behavior of sellers of properties in neighborhood group n. Visiting costs are identified through selling time. Everything else equal, a larger visiting cost leads to fewer buyer visits and longer selling time.

5.2. The Estimation Results

Table 3 shows the estimated parameters. The median of normalized visiting cost is 0.0227, suggesting that buyers incur non-trivial cost in visiting a property.[30] The prior uncertainty is largest in poor neighborhoods (0.1535), followed by rich neighborhoods (0.1329), and is smallest in mid-range neighborhoods (0.1268). The difference in the prior uncertainty between poor neighborhoods and rich (or mid-range) neighborhoods is statistically significant. Thus, the results suggest that market participants in poor neighborhoods face greater uncertainty in the absence of Zestimate. A possible reason is that people in poor neighborhoods are generally less educated and have less access to high quality information or agents to help them make decisions.

A caveat here is that the estimates are based on the assumption of the same fixed discounting factor for all sellers. It is possible that sellers in poor neighborhoods have a lower discounting factor. If that is the case, then the properties in poor neighborhoods would have consistently lower

[30] The visiting cost includes both monetary and non-monetary cost associated with visiting a property. Also, as discussed in Appendix B, our estimate of search cost may capture the positive search value for buyers, as buyer sequential search is not explicitly modeled. Please also see Appendix B on why we do not explicitly model buyer sequential search.

listing price and reservation price, conditional on the same belief. Thus, we would overestimate the prior mean belief in poor neighborhoods (α_0^P) as compared to that in other neighborhoods (α_0^M and α_0^R). This, however, should not directly affect the relative sizes of prior uncertainty, which is identified from the listing price updates as described in Section 5.1 and Appendix F.

The estimates of a_z(-0.1480) and b_z(0.9984) suggest that the Zestimate signal about the mean of log buyer valuation (λ_i) is a fraction of the raw Zestimate value. This is intuitive as market value (expected selling price) tends to be higher than mean buyer valuation. Based on the estimates of a_s, b_s and c_s, the average information-to-noise ratio (calculated as prior variance divided by Zestimate variance) is 0.2906, which suggests that Zestimate has non-trivial effect on the beliefs about λ_i. More importantly, the average information-to-noise ratio is the largest in poor neighborhoods (0.3360) compared to other neighborhoods groups (0.2636 and 0.2845 in mid-range and rich neighborhoods, respectively).[31] Although Zestimate is less accurate in poor neighborhoods, it actually has a larger impact in poor neighborhoods because of the higher prior uncertainty in poor neighborhoods.

6. Counterfactual Analysis

In this section, we conduct counterfactual analysis to analyze the impact of Zestimate and Zestimate accuracy. We first calculate a posterior distribution of α_i (defined in Equation 5) for each individual property, and use the mean of the posterior distribution as the value of α_i in simulations. For each condition, we simulate 500 paths and calculate the average initial listing price, the average final selling price, the average TOM, the average buyer surplus, the average seller profit and the average total surplus for each property in our sample.

6.1. The (Disparate) Impact of Zestimate

In the first counterfactual analysis, we compare the scenario where Zestimate is present with the scenario where Zestimate is removed. When Zestimate is removed, buyers and sellers make decisions based on their prior belief about property values, and only update their belief on buyer valuations. In this case, the uncertainty in belief is higher and the mean belief is not shifted by an

[31] All the differences in the average information-to-noise ratio are statistically significant.

Table 3 Parameter Estimates

Parameter	Description	Estimate	Standard Errors
c_0	The mean of log visiting cost	-3.7860	0.0166
σ_c	The standard deviation (SD) of log visiting cost	1.1227	0.0014
δ^P	The (relative) probability of buyer arrival in poor neighborhoods	0.9431	0.0011
δ^R	The (relative) probability of buyer arrival in rich neighborhoods	0.9674	0.0006
σ_0^P	The SD of prior belief in poor neighborhoods	0.1535	0.0019
σ_0^M	The SD of prior belief in mid-range neighborhoods	0.1268	0.0011
σ_0^R	The SD of prior belief in rich neighborhoods	0.1329	0.0015
α_0^P	The mean of prior belief in poor neighborhoods	7.8359	0.5130
α_0^M	The mean of prior belief in mid-range neighborhoods	8.2123	0.4793
α_0^R	The mean of prior belief in rich neighborhoods	8.6341	0.4335
σ_v^P	The SD of log buyer valuations in poor neighborhoods	0.2066	0.0015
σ_v^M	The SD of log buyer valuations in mid-range neighborhoods	0.1993	0.0008
σ_v^R	The SD of log buyer valuations in rich neighborhoods	0.2104	0.0018
ρ^P	The coefficient of holding value in poor neighborhoods	1.0694	0.0033
ρ^M	The coefficient of holding value in middle neighborhoods	1.0358	0.0053
ρ^R	The coefficient of holding value in rich neighborhoods	1.0869	0.0074
a_z	The intercept in the linear transformation of Zestimate	-0.1480	0.0011
b_z	The slope of $ln(Z_{it})$ in the linear transformation of Zestimate	0.9984	0.0002
c_z	The slope of σ_v^n in the linear transformation of Zestimate	0.0933	0.0008
a_s	The intercept in the linear transformation of Zestimate range	0.0637	0.0005
b_s	The slope of ZR_{it} in the linear transformation of Zestimate range	1.7921	0.0178
c_s	The slope of σ_v^n in the linear transformation of Zestimate range	0.0737	0.0004

over-valued or under-valued Zestimate signal. The left half of Table 4 shows the changes in the initial listing price, the final selling price and the TOM brought by Zestimate. We can see that on average Zestimate leads to lower listing price, higher selling price and longer TOM.

Table 4 Average Percentage Change in Market Outcomes by Zestimate (compared to without Zestimate)

	Population	Poor	Mid-Range	Rich		Population	Poor	Mid-Range	Rich
Listing price	-1.80%	-2.10%	-1.52%	-1.93%	Total Surplus	4.05%	5.05%	3.59%	3.74%
Selling price	2.78%	3.43%	2.06%	3.29%	Buyer Surplus	5.35%	8.27%	3.16%	6.01%
TOM	4.52%	5.58%	4.20%	3.87%	Seller Profit	4.16%	4.96%	3.74%	4.02%

* All the differences in average surplus changes across neighborhoods are statistically significant.

We also calculate the expected buyer surplus and the expected seller profit for each property under the two conditions. The seller profit is calculated as the difference between the selling price and the holding value, discounted by the time it takes to sell:

$$\text{SellerProfit}_i = \beta^{T_i-1} \cdot (p_i - v_i^h) \tag{27}$$

The buyer surplus has two parts: the utility of purchasing (the difference between the final buyer valuation and the selling price) and the total visiting cost of all buyers who visit the property:

$$\text{BuyerSurplus}_i = (v_{iT_i} - p_i) - \sum_{t=1}^{T_i} c_{it} A_{it}, \tag{28}$$

where A_{it} is the binary indicator of whether a buyer visits. The total surplus is simply the sum of seller profit and buyer surplus.

The right half of Table 4 shows the average surplus change caused by Zestimate. First, Zestimate on average increases both seller profit and buyer surplus among properties in all neighborhood groups (population), suggesting that overall Zestimate benefits the market. Note that among the properties whose Zestimates are lower than the prior mean beliefs, only 42% of them have lower seller profit with Zestimate than without, suggesting that having an "undervalued" Zestimate may still be better for the seller than not having a Zestimate because of the benefit of uncertainty reduction. Moreover, although Zestimate is least accurate in poor neighborhoods, it actually leads to the greatest surplus increase in poor neighborhoods.

6.2. Understanding the Mechanism

The above counterfactual analysis suggests that 1) Zestimate overall benefits both buyers and sellers in the housing market; 2) Zestimate benefits poor neighborhoods more despite being less accurate in poor neighborhoods. In this section, we explain the mechanism behind these results.

Recall that Zestimate has two effects on the beliefs about true property value: uncertainty reduction effect and mean shift effect. In what follows, we discuss how these two effects influence sellers' and buyers' decisions as well as market outcomes (e.g., selling price and time on market). Note that all these results are consistent with the reduced-form results shown in Section 3.4.

Let us first look at Zestimate's uncertainty reduction effect on the seller's listing price. Note that the listing price is a commitment device that sellers use to attract buyers to visit the property, as a buyer knows that the final selling price is capped at the listing price. This implies that the lower the listing price, the more likely buyers will visit the property. Next, note that the uncertainty in buyer's valuations is positively related to the buyer's propensity to visit the property (which allows the buyer to learn about the property's true value), as higher uncertainty implies higher option value of search for the buyer. All else being the same, if the buyer's uncertainty decreases because of the information provided by Zestimate, then the buyer's propensity to visit the property will also decrease. To offset this decrease in visit probability, the seller will decrease the listing price in order to encourage more buyer visits. This implies that the greater the uncertainty reduction because of Zestimate, the lower will be the listing price.

We next discuss the impact of uncertainty reduction on the seller's reservation price. The seller's reservation price in a given period is directly related to the expected future profits that the seller would make if the property does not get sold — the greater the expected future profit, the greater will be the reservation price in the current period. If a property is not sold in the current period, it signals to the future buyers that the realized buyer valuation is relatively low, which would decrease their belief about the mean of log buyer valuations (i.e. the belief about property value). This decrease in valuations of future buyers will lead to lower future expected profits for the seller since the seller would have to decrease the listing price in the future. The larger the variance of the buyers' beliefs about the true property value, the greater will be decrease in the valuations of future buyers if the property is not sold, because high uncertainty in valuations implies that the true property value could come from a broad range of values. Thus if the property is not sold, then from the future buyers' perspective, there is a reasonable chance that the true value of property might be really low, which leads to a large reduction in their valuations. This will lead to lower expected future profits, and thus a lower reservation price. In other words, the greater the uncertainty reduction because of Zestimate, the higher will be the seller's reservation price.

The impact of uncertainty reduction by Zestimate on the time on market follows directly from the impact of uncertainty reduction on reservation price. Since the actual draws of buyer valuations are not affected by the beliefs about the true value, a higher reservation price will imply that the property is less likely to be sold and will be on the market for a longer time. Therefore, the greater the uncertainty reduction because of Zestimate, the longer will be the time on the market.

The impact of uncertainty reduction by Zestimate on the final selling price follows directly from the impact of uncertainty reduction on the listing price and the reservation price. Note that the final selling price will be a convex combination of the listing price and the reservation price, as determined by the relative bargaining power of the buyer and the seller. Since the decrease in uncertainty by Zestimate leads to an increase in the reservation price and a decrease in the listing price, it implies that the impact of uncertainty reduction on the selling price can go either way. In our case, we find that the uncertainty reduction has a positive effect on selling price.

The model's prediction about the impact of Zestimate uncertainty on listing price and reservation price is consistent with the reduced-form regression results in Table 2 Columns 6-8: As Zestimate uncertainty increases, the listing price would be higher, while the seller's reservation price would be lower. Since buyers' realized valuations are drawn from the true distribution of buyer valuations and not affected by Zestimate, the bargaining price would be lower with a lower reservation price.

The impact of the shift in the mean beliefs of the buyers' valuations (because of Zestimate) on the prices and the time on market is straightforward. If Zestimate increases the mean beliefs of the buyer's valuations, the seller would set a higher listing price and would have a higher reservation price, in anticipation of higher buyer valuations. Using the same logic as before, a higher reservation price will lead to longer time on market, and a higher reservation price and listing price will lead to a higher selling price. At the same time, since the realized buyer valuation is not affected by Zestimate, the property is less likely to be sold at the listing price (which requires a relatively high buyer valuation) and more likely to have listing price adjustment. Again, this is consistent with the reduced-form regression results in Table 2 Columns 1-5.

With an understanding of how Zestimate affects the decisions of sellers and buyers as well as the market outcomes, it is now easy to understand the results we show in Section 6.1. We first discuss how Zestimate benefits both sellers and buyers in the housing market. The seller profit is the time discounted difference between the selling price and the holding value. Since sellers do not heavily discount future within a relatively short time period and the holding value of the property is fixed, seller profit is largely determined by selling price. When Zestimate increases prior mean belief, both the mean shift effect and the uncertainty reduction effect leads to higher selling price and thus higher seller profit; when Zestimate decreases prior mean belief, the mean shift effect decreases selling price but the uncertainty reduction effect increases selling price. For most of the properties in our sample, the uncertainty reduction effect dominates the mean shift effect, and 58% of the sellers with Zestimate lower than prior mean belief still have increased profit.

The buyer surplus consists of two parts: the utility of purchasing, and the total visiting costs of all buyers who have visited the property. Recall that when Zestimate reduces uncertainty, buyers are less affected by the negative signals of a property staying on the market, thus sellers can be more patient and wait for a buyer with higher valuation for the properties. In other words, with lower uncertainty, the property would be sold to a buyer who values it more. This means the purchasing buyer's valuation is higher, and therefore increases the utility of purchasing. Therefore, the increased buyer surplus mainly comes from the higher valuation of the buyers who eventually purchase the houses when Zestimate reduces uncertainty.

We now come to the result that Zestimate benefits poor neighborhoods more despite of being lesser accurate in poor neighborhoods. A less accurate Zestimate is usually considered less beneficial, because it provides a noisier signal and reduces uncertainty to a less extent. However, this is true when everything else is equal. Our estimation results suggest that compared to people in other neighborhoods, people in poor neighborhoods face greater prior uncertainty, meaning that they are more uncertain about property values before they learn from Zestimate. A possible reason is that people in poor neighborhoods are generally less educated, and have less access

to high quality information or agents to help them make decisions. The larger prior uncertainty means that, in poor neighborhoods, the prior belief about property value is overall more inaccurate, and more importantly, the same signal would reduce uncertainty to a greater extent. From the previous discussion, we know that Zestimate benefits buyers and sellers mainly because of its uncertainty reduction effect. The uncertainty reduction effect depends on the relative size of the prior uncertainty and the Zestimate uncertainty. Although Zestimate is less accurate in poor neighborhoods, the prior uncertainty is also larger in poor neighborhoods, thus it is possible that Zestimate still reduces more uncertainty in poor neighborhoods and benefits buyers and sellers more in these neighborhoods, which is what we find in our empirical results.

6.3. Improving Zestimate Accuracy in Poor Neighborhoods

The previous counterfactual analysis shows that currently Zestimate leads to greater surplus increase in poor neighborhoods and therefore helps reduce housing inequality. However, more accurate Zestimate is still more desirable, and poor neighborhoods are losing some of the benefits because of the less accurate Zestimate. There are generally fewer features available on Zillow for properties in poor neighborhoods, and even the available features may be outdated or erroneous due to less well-maintained public records. Luckily, Zillow allows homeowners to provide or update home characteristics information to improve Zestimate accuracy.[32] Therefore, homeowners could provide more information about their properties and benefit from a more accurate Zestimate.

Table 5 Average Surplus Change in Poor Neighborhoods with More Accurate Zestimate

Zesetimate	Current	Improved
Total surplus	5.05%	6.62%
Seller profit	4.96%	6.05%
Buyer surplus	8.27%	13.79%

To evaluate how much poor neighborhoods are losing because of the less accurate Zestimate, in the second counterfactual analysis, we increase Zestimate accuracy in poor neighborhoods to be the same as that in rich neighborhoods. Specifically, we first adjust Zestimate Range in poor

[32] https://www.zillow.com/sellerlanding/edityourhome/

neighborhoods to have the same mean as that in rich neighborhoods. Then we adjust Zestimate values in poor neighborhoods correspondingly to be more accurate (less deviated from true property value) based on the degree of reduction in the variance of Zestimate signals (implied by the lower Zestimate Range). Table 5 shows the average surplus change in poor neighborhoods as we increase Zestimate accuracy. If on average Zestimate in poor neighborhoods were as accurate as that in rich neighborhoods, the average total surplus change would be 6.62%. Compared to the total surplus change of 5.05% with the current Zestimate accuracy, the positive impact of Zestimate on total surplus in poor neighborhoods could further increase by 31.09% with higher accuracy.

7. Conclusion

In this paper, we study the impact of Zestimate on the housing market in terms of market outcomes and surplus. Additionally, as Zestimate tends to be more accurate for rich neighborhoods compared to poor neighborhoods, raising concerns that Zestimate may widen the socio-economic inequality, we also examine how and to what extent Zestimate affects the inequality in the housing market.

We build a structural model of housing market where sellers and buyers face uncertainty about property values and Zestimate provides an unbiased signal of the property value. The estimation results reveal that people in poor neighborhoods face greater uncertainty in their prior belief compared to those in the mid-range and rich neighborhoods.

The counterfactual analysis suggests that Zestimate overall benefits both buyers and sellers in the market. Although Zestimate can overvalue or undervalue a property and may lead to incorrect belief about the property value, its uncertainty reduction effect plays a larger role. As it reduces uncertainty, buyers are less affected by the negative signals of a house staying on the market, thus sellers can be more patient to stay on the market and wait for buyers who truly value the house. This increases the match quality between buyers and sellers and increases the total surplus.

We also find that Zestimate benefits poor neighborhoods more than rich neighborhoods, despite being less accurate in poor neighborhoods. If we only look at the lower accuracy in poor neighborhoods, we may conclude that Zestimate is exacerbating social inequality with more of the benefits

going into rich people. However, our structural model allows us to find that poor neighborhoods face greater prior uncertainty without Zestimate, probably because poor people are generally less educated and lack high quality information or agents to help them make decisions. Therefore, even a noisier signal can be more helpful and benefit them to a greater extent, compared to what a precise signal can do to people in rich neighborhoods. Thus, when evaluating the disparate impact of algorithms, it is crucial to consider the status quo of different groups and the counterfactual outcomes when the algorithms are not there, rather than simply looking at the algorithm outputs.

There are several limitations of the paper. First, we do not model the case where multiple buyers come and engage in a bidding war, which usually leads to a selling price higher than the listing price. While such cases are rare in our data, it is common in a hot market. Zestimate may affect "hot" housing markets differently. Second, due to limited demand side data, we model buyers as myopic who consider one property at a time. In reality, buyers may simultaneously consider multiple listings, and in those cases, there is competition among sellers of similar properties. Last, we focus on the impact of Zestimate on on-market properties and do not consider the impact of Zestimate on market entry decisions. Despite those limitations, our work is among the first to demonstrate and explain the impact of a machine generated property value prediction on the housing market, as well as its economic and social implications. We hope our work can pave the way for future research in this important area.

Funding and Competing Interests: All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript. The authors have no funding to report.

References

Agrawal A, Gans JS, Goldfarb A (2019) Artificial intelligence: the ambiguous labor market impact of automating prediction. *Journal of Economic Perspectives* 33(2):31–50.

Calvano E, Calzolari G, Denicolo V, Pastorello S (2020) Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review* 110(10):3267–97.

Carrillo PE (2012) An empirical stationary equilibrium search model of the housing market. *International Economic Review* 53(1):203–234.

Chen Y, Rosenthal RW (1996a) Asking prices as commitment devices. *International Economic Review* 129–155.

Chen Y, Rosenthal RW (1996b) On the use of ceiling-price commitments by monopolists. *The RAND Journal of Economics* 207–220.

Cowgill B (2018) Bias and productivity in humans and algorithms: Theory and evidence from resume screening. *Columbia Business School, Columbia University* 29.

Fu R, Huang Y, Singh PV (2021) Crowds, lending, machine, and bias. *Information Systems Research* 32(1):72–92.

Han L, Strange WC (2014) Bidding wars for houses. *Real Estate Economics* 42(1):1–32.

Hansen KT, Misra K, Pai MM (2021) Frontiers: Algorithmic collusion: Supra-competitive prices via independent algorithms. *Marketing Science* 40(1):1–12.

Ihlanfeldt K, Mayock T (2009) Price discrimination in the housing market. *Journal of Urban Economics* 66(2):125–140.

King AT, Mieszkowski P (1973) Racial discrimination, segregation, and the price of housing. *Journal of Political Economy* 81(3):590–606.

Kleinberg J, Lakkaraju H, Leskovec J, Ludwig J, Mullainathan S (2018) Human decisions and machine predictions. *The quarterly journal of economics* 133(1):237–293.

Lambrecht A, Tucker C (2019) Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of stem career ads. *Management Science* 65(7):2966–2981.

Lambrecht A, Tucker CE (2020) Apparent algorithmic discrimination and real-time algorithmic learning in digital search advertising. *Available at SSRN 3570076* .

Lu G (2019) How machine learning mitigates racial bias in us housing market. *Available at SSRN 3489519* .

Merlo A, Ortalo-Magné F, Rust J (2015) The home selling problem: Theory and evidence. *International Economic Review* 56(2):457–484.

Nash JF (1950) The bargaining problem. *Econometrica* 18(2):155–162.

Piketty T, Zucman G (2014) Capital is back: Wealth-income ratios in rich countries 1700–2010. *The Quarterly Journal of Economics* 129(3):1255–1310.

Quan DC (2002) Market mechanism choice and real estate disposition: Search versus auction. *Real Estate Economics* 30(3):365–384.

Straszheim MR (1974) Housing market discrimination and black housing consumption. *The Quarterly Journal of Economics* 19–43.

Yu S (2020) Algorithmic outputs as information source: The effects of zestimates on home prices and racial bias in the housing market. *Available at SSRN 3584896* .

Yu S (2021) Algorithmic outputs as information source: The effects of zestimates on home prices and racial bias in the housing market. *Available at SSRN 3584896* .

Zhang L, Chung DJ (2020) Price bargaining and competition in online platforms: An empirical analysis of the daily deal market. *Marketing Science* 39(4):687–706.

Zhang S, Mehta N, Singh PV, Srinivasan K (2021) Can an ai algorithm mitigate racial economic inequality? an analysis in the context of airbnb. *Marketing Science* .

Appendix A Risk Aversion

In our model, we follow the previous literature that model buyer and seller behavior in housing markets and assume risk-neutral agents. However, it is possible that buyers and sellers are risk-averse, especially when dealing with such large transactions. Modeling risk aversion means that uncertainty would have a direct negative impact on buyers' and sellers' utilities. There are two types of uncertainty in our main model: The first is about the mean of log buyer valuations, and the second is about a given buyer's realized valuation (conditional on a known distribution of buyer valuations).

The second type of uncertainty (about a given buyer's realized valuation) would remain unchanged regardless of whether Zestimate is available or not. Thus, it does affect our evaluation of the effect of Zestimate. The first type of uncertainty (about the mean of log buyer valuations) will be reduced with the presence of Zestimate. When buyers and sellers are risk averse, Zestimate would directly increase their utilities because of its uncertainty reduction effect. This benefit is in addition to the indirect benefit of Zestimate's uncertainty reduction effect (through affecting the optimal decisions) that we show with the main model. Therefore, if we model risk aversion, our main results would still hold, if not strengthened by the direct benefit of uncertainty reduction. To preserve the simplicity of the model and to demonstrate that the benefit of Zestimate's uncertainty reduction effect is not driven by the assumption of risk aversion, we choose to assume risk-neutral agents. Our results suggest that Zestimate's uncertainty reduction effect would benefit buyers and sellers even when they are not assumed to be risk averse.

Appendix B Buyer Search

In our model, we assume that a buyer will exit the market and obtain 0 utility if she chooses not to visit the property or does not purchase after a visit. In other words, we do not model buyers' sequential search. This is primarily due to data limitation: we do not observe what properties a specific buyer considers or visits and in what sequence the buyer visits them. Thus, we could not model buyer sequential search without making strong assumptions.

One possible way to model buyer search is to follow (Carrillo 2012) and assume that the buyer treat all other properties equally as "potential properties" without additional property-specific information. In this case, the buyer has a value of search, which is the discounted expected utility of the maximum between visiting a property and waiting for another property in the next period. Formally, we denote the value of search as π_b, and we have

$$\pi_b = \beta \iint \max\{u(\mu_{it}, \sigma_{it}^2, \pi_b), \pi_b\} \, \mathrm{d}\Gamma(\mu_{it}, \sigma_{it}) \tag{29}$$

where β is the discounting factor, $\Gamma(\mu_{it}, \sigma_{it})$ is the joint distribution of μ_{it} and σ_{it} across all the on-market properties, and $u(\mu_{it}, \sigma_{it}^2, \pi_b)$ is the utility of visiting a property with the belief about its λ_i as $\mathcal{N}(\mu_{it}, \sigma_{it}^2)$, which we derive next.

Now with a value of search, the buyer would purchase a property if and only if the utility of purchase is greater than the value of search, i.e.,

$$u_{it} = v_{it} - p_i > \pi_b. \tag{30}$$

Also, in the bargaining process, the buyer's disagreement payoff becomes π_b (instead of 0 in our main model).This makes the bargaining outcome $\theta(v_{it} - \pi_b) + (1-\theta)R_{it}$. Since the selling price will be capped at the listing price, we know that the buyer will purchase the property at the listing price if

$$v_{it} > \frac{L_{it} - (1-\theta)R_{it}}{\theta} + \pi_b \equiv \bar{v}_{it}, \tag{31}$$

and purchase the property at the bargaining outcome otherwise.

As a result, when the buyer decides whether to visit the property, the utility of visiting becomes

$$\begin{aligned} u(\mu_{it}, \sigma_{it}^2, \pi_b) = &\int_{R_{it}+\pi_b}^{\bar{v}_{it}} [(1-\theta)\cdot(\tilde{v}_{it} - R_{it}) + \theta\pi_b] f(\tilde{v}_{it}) \, \mathrm{d}\tilde{v}_{it} \\ &+ \int_{\bar{v}_{it}}^{\infty} (\tilde{v}_{it} - L_{it}) f(\tilde{v}_{it}) \, \mathrm{d}\tilde{v}_{it} \\ &+ \pi_b \cdot \Pr(\tilde{v}_{it} < R_{it}) \\ &- c_{it}. \end{aligned} \tag{32}$$

Similar as in the main model, on the right-hand-side, the first line is the expected utility when the buyer purchases the property at a bargaining price (i.e. below the listing price); the second line is the expected utility when the buyer purchases the property at the listing price; the third line is the utility when the buyer did not purchase the property and wait for the next period; the last line is the visiting cost. The buyer would choose to visit if and only if the utility of visiting if greater than the value of search π_b.

Comparing this alternative model that captures buyer search with our main model, we can see that if buyers do search sequentially, then with our model that does not capture such searches, we will overestimate the search cost, because we would attribute all the buyer decisions of not visiting to high search costs when they may also be caused by the positive search value (value associated with not visiting the focal property). We would also underestimate buyer surplus while overestimate seller surplus. This is because we count the utility from walking away as zero, and therefore underestimate the valuation of the property from the buyer who eventually buys the property, which leads to underestimation of buyer surplus. Meanwhile, as our model does not account for the more favorable bargaining outcomes towards buyers because of their positive disagreement payoff, we would underestimate the seller's reservation price from the observed selling price. This means we would underestimate the holding value and therefore overestimate seller surplus. However, these changes would not affect our main results of how Zestimate benefits buyers and sellers by reducing their uncertainty about the mean of log buyer valuations and that Zestimate benefits poor neighborhoods more as poor neighborhoods face greater prior uncertainty. Thus, our main results would still hold even if we make the assumption of other properties being "homogeneous" and model buyer searches.

Appendix C Learning About the Variance of Log Buyer Valuations

In our main model, we assume that buyers and sellers know the variance of log buyer valuations (σ_v^2), while they are uncertain about the mean of log buyer valuations (λ_i). This assumption is based on the following reasons. First, buyers and sellers are more likely to be uncertain about the

mean than the variance of log buyer valuations, as the variance is shared across all the properties (within the same neighborhoods groups), while the mean is property specific. Thus, there would be more available information about the variance than that about the mean. Second, buyers and sellers are more likely to use Zestimate to learn about the mean, as Zestimate intends to be an estimate of a property's market value, which is reflected more in the mean rather than the variance of log buyer valuations. Last, buyers and sellers are more likely to interpret Zestimate Range as an indicator of Zestimate uncertainty instead of a signal about the variance of log buyer valuations. In the official guide about Zestimate, Zillow says "A wider range generally indicates a more uncertain Zestimate, which might be the result of unique home factors or less data available for the region or that particular home. It's important to consider the size of the Estimated Sale Range because it offers important context about the Zestimate's anticipated accuracy".[33]

Still, it is possible that buyers and sellers may learn about the variance of log buyer valuations from Zestimate and Zestimate Range. Note that from Equation 15 and Equation 20, we can see the variance of log buyer valuations (σ_v^2) and the uncertainty about the mean of log buyer valuations (σ_{it}^2) enter the buyer's utility of visiting (Equation 18) and the seller's profit function (Equation 21) in the same way. Ultimately, sellers and buyers care about the realized buyer valuations, while σ_v^2 and σ_{it}^2 affect their beliefs about the distribution of buyer valuations similarly. This means that changes in the variance of log buyer valuations would have similar effects on buyer and seller decisions (and therefore market outcomes) as changes in the uncertainty about the mean of log buyer valuations. Therefore, if buyers and sellers do learn about the variance of log buyer valuations from Zestimate, then the uncertainty reduction effect we estimate with our model is actually the combined effect of the reduced uncertainty about the mean and the learning about the variance. Nonetheless, the mechanism of how Zestimate benefits buyers and sellers by reducing the uncertainty about the mean is still valid, and the result of Zestimate benefiting poor neighborhoods more remains the same. Therefore, our main results would still hold even if we model the learning about the variance of log buyer valuations.

[33] https://www.zillow.com/z/zestimate/

Appendix D Alternative Model of Asymmetric Information

In the main model, buyers and sellers have symmetric information: they both know the home selling process, observe the same home characteristics, and share the same belief about the distribution of buyer valuations. We make this assumption for simplicity and for identification purposes, as we observe very little buyer side information. In reality, buyers and sellers may have different information sets. In this section, We present an alternative model where buyers and sellers have asymmetric information, and show in numerical simulation that this alternative model yields similar results to our main model. This is to illustrate that the assumption of symmetric information does not significant influence the main mechanisms we present in the paper.

In the alternative model, we assume that are certain home characteristics that sellers observe but buyers do not observe. Recall that in our main model, the mean value of log buyer valuations is

$$\lambda_i = \alpha_i + \gamma \mathbf{X_i} + u_i, \quad u_i \sim \mathcal{N}(0, \sigma_u), \tag{33}$$

where $\mathbf{X_i}$ is a vector of property features that buyers, sellers, and we as researchers observe, γ is a vector of parameters that reflect the contribution of those features to property value, u_i is the contribution of home characteristics that buyers and sellers observe, but we as researchers do not observe, and α_i is the part that sellers and buyers are uncertain about. In reality, there could be certain characteristics that are available to sellers but not to buyers. Thus, in this alternative model, we revise the expression of the mean of log buyer valuation to be

$$\lambda_i = \alpha_i + \gamma \mathbf{X_i} + \eta_i + \kappa_i, \quad \eta_i \sim \mathcal{N}(0, \sigma_\eta^2), \quad \kappa_i \sim \mathcal{N}(0, \sigma_\kappa^2), \tag{34}$$

where η_i is the contribution of home characteristics that both sellers and buyers observe, and κ_i is the contribution of home characteristics that only sellers observe. With loss of generality, we assume both η_i and κ_i are normally distributed with mean 0.[34]

[34] The mean of η_i cannot be separated from the mean belief of α_i; if κ_i has non-zero mean, buyers can always adjust the mean value with the knowledge of the distribution of κ_i.

It is important to note that even with this newly introduced κ_i, buyers and sellers still do not have actual asymmetric information. This is because buyers always observe sellers' listing prices, from which they could infer κ_i. In other words, even though buyers do not directly observe certain home characteristics of a property, they can infer the contribution of these home characteristics to the mean of log buyer valuations by observing the listing price as the optimal solution to the seller's problem. Therefore, a model where buyers and sellers truly have asymmetric information needs to contain another source of asymmetric information.

One possibility is to incorporate a random utility shock of selling a property. This additive utility shock captures any unobserved factors that affect the utility that a seller obtains by selling his property in the current time period. We assume that it is i.i.d. across sellers and time periods, and denote it as $\epsilon_{it} \sim \mathcal{N}(0, \sigma_\epsilon^2)$. The random utility shock is only revealed to a seller at the beginning of a period, and is unknown to the buyers. With two sources of unknown information (κ_i and ϵ), buyers could not distinguish their impacts on the optimal listing prices.

Since buyers do not have all the information that sellers have, they are unable to infer sellers' reservation prices. Instead, for each property, they would infer a distribution of possible reservation prices, based on their knowledge of the distributions of κ_i and ϵ_{it} and the observed listing price. We denote this distribution of possible reservation prices from buyers' perspective as $g(R_{it}|L_{it})$. Also, buyers now have different beliefs about the mean of log buyer valuations from the seller. We still denote the seller's belief about the distribution of buyer valuations as $\tilde{v}_{it}$:

$$\tilde{v}_{it} \sim \mathcal{LN}(\mu_{it}, \sigma_{it}^2 + \sigma_v^2), \tag{35}$$

and similarly as in the main model, we have

$$\mu_{i1}^0 = \alpha_0 + \gamma \mathbf{X_i} + \eta_i + \kappa_i. \tag{36}$$

We denote the buyers' belief about the distribution of buyer valuations as $\tilde{v}'_{it}$:

$$\tilde{v}'_{it} \sim \mathcal{LN}(\mu'_{it}, \sigma_{it}^2 + \sigma_v^2), \tag{37}$$

and it follows that

$$(\mu_{it}^0)' = \mu_{it}^0 - \kappa_i. \tag{38}$$

Now the utility of visiting the property for the buyer is:

$$\begin{aligned} u(\mu'_{it}, \sigma^2_{it}|L_{it}) = \int \{ & \int_{R_{it}}^{\bar{v}'_{it}} (1-\theta)\cdot(\tilde{v}'_{it} - R_{it}) f(\tilde{v}'_{it})\, \mathrm{d}\tilde{v}'_{it} \\ & + \int_{\bar{v}_{it}}^{\infty} (\tilde{v}'_{it} - L_{it}) f(\tilde{v}'_{it})\, \mathrm{d}\tilde{v}'_{it} \} g(R_{it}|L_{it}) \mathrm{d}R_{it} \\ & - c_{it}, \end{aligned} \tag{39}$$

and the buyer would visit the property if and only if $u(\mu'_{it}, \sigma^2_{it}|L_{it}) > 0$. The probability of buyer visit is denoted as $q(\mu'_{it}, \sigma^2_{it}, L_{it})$. The seller's problem now becomes

$$\begin{aligned} \pi(\mu_{it}, \sigma^2_{it}, \epsilon_{it}, \kappa_i) = \max_{L_{it}, R_{it}} \{ & [(1-\delta) + \delta(1 - q(\mu'_{it}, \sigma^2_{it}, L_{it}))] \cdot \beta \cdot \int \pi(\mu_{it}, \sigma^2_{it}, \epsilon_{it}, \kappa_i)\, f(\epsilon_{it}) \mathrm{d}\epsilon_{it} \\ & + \delta q(\mu'_{it}, \sigma^2_{it}, L_{it}) \int_{\bar{v}_{it}}^{\infty} (L_{it} + \epsilon_{it}) \cdot f(\tilde{v}_{it})\, \mathrm{d}\tilde{v}_{it} \\ & + \delta q(\mu'_{it}, \sigma^2_{it}, L_{it}) \int_{R_{it}}^{\bar{v}_{it}} [\theta \tilde{v}_{it} + (1-\theta) R_{it} + \epsilon_{it}] f(\tilde{v}_{it})\, \mathrm{d}\tilde{v}_{it} \\ & + \delta q(\mu'_{it}, \sigma^2_{it}, L_{it}) \cdot \int_{\infty}^{R_{it}} \max\{v_i^h, \beta \int \pi(\mu_{i(t+1)}(\tilde{v}_{it}), \sigma^2_{i(t+1)}, \epsilon_{it}, \kappa_i) f(\epsilon_{it}) \mathrm{d}\epsilon_{it}\} f(\tilde{v}_{it})\, \mathrm{d}\tilde{v}_{it}\}, \end{aligned} \tag{40}$$

We use estimated parameters reported in Table 3 to run the simulation exercise. For the two additional parameters, we fix them at $\sigma_\kappa = 0.04$ and $\sigma_\epsilon = 0.03$. Since Zestimate affects housing market outcomes by shifting the mean belief and reduce the variance in the belief, the main mechanism of our findings can be summarized by how the optimal listing price and the optimal reservation price are affected by the mean belief and the variance in the belief. The uncertainty reduction effect of Zestimate suggests that when the variance in the belief decreases, the optimal listing price reduces, while the optimal reservation price increases (as the sellers can be more patient to wait). The mean shift effect of Zestimate suggests that when the mean belief of Zestimate increases, both the optimal listing price and the optimal reservation price increase.

In the simulation exercise, we find that these patterns remain in the alternative model with asymmetric information. To illustrate the results, we calculate the optimal listing price and the

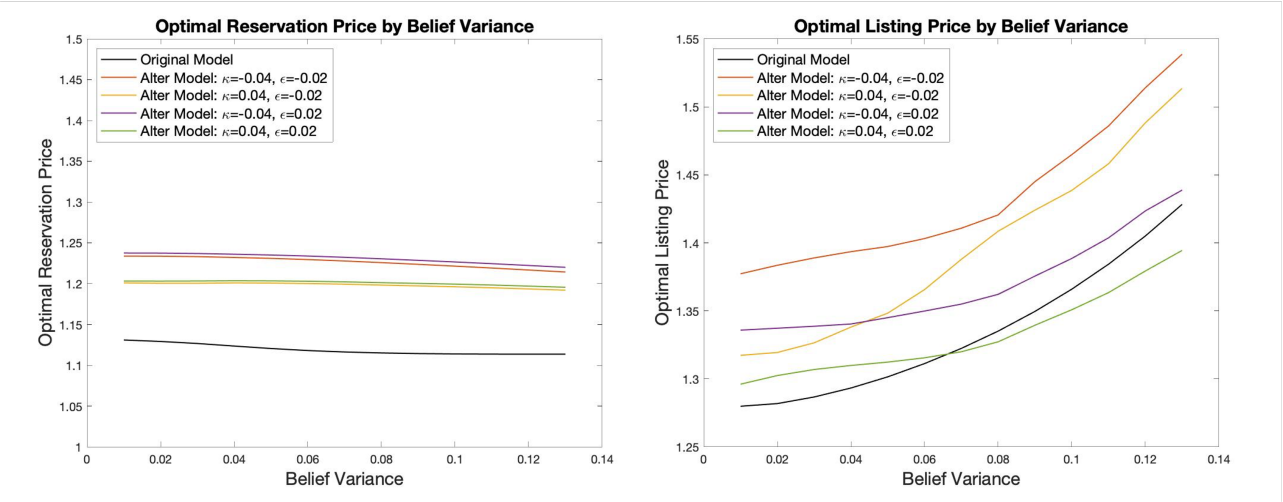

Figure 5 **Simulation results of how the optimal listing price and reservation price changes by the standard deviation of the belief in rich neighborhoods. Relative μ_{it} is fixed at 0.**

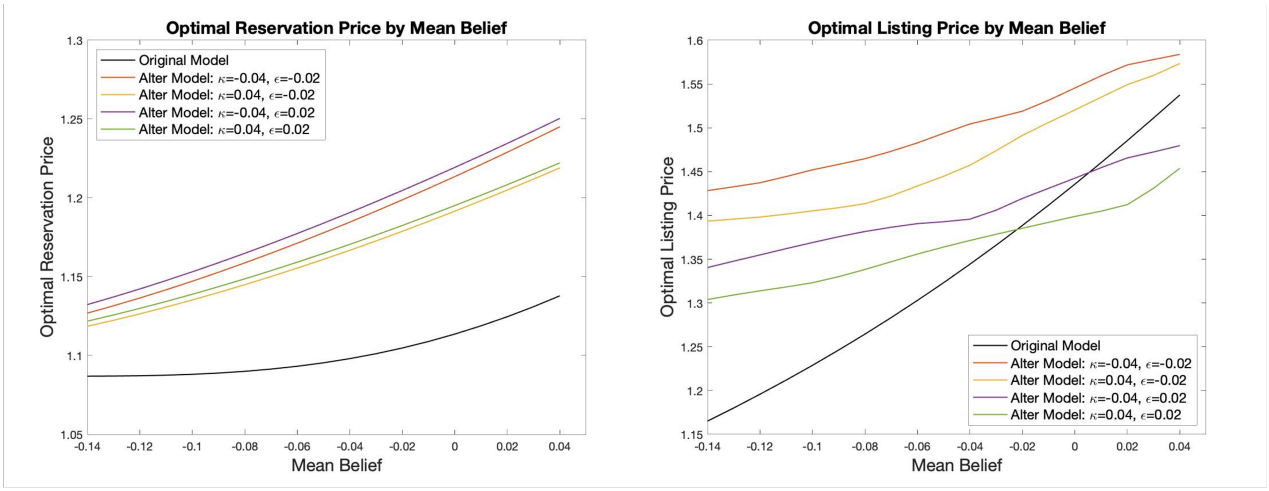

Figure 6 **Simulation results of how the optimal listing price and reservation price changes by the mean of the belief in rich neighborhoods. σ_{it} is fixed at 0.1329**

optimal reservation price in relative to a measure of property value — $\exp(\alpha_0 + \gamma \mathbf{X_i} + u_i)$ in the main model and $\exp(\alpha_0 + \gamma \mathbf{X_i} + \eta_i + \kappa_i)$ in the alternative model. We plot how the relative optimal listing prices and reservation prices vary by belief variance (σ_{it}) and mean belief (μ_{it}) in rich neighborhoods under (1) the main model and (2) the 4 cases of the alternative model with different value combinations of κ_i and ϵ_{it}. We can see from Figure 5 and Figure 6 that in all the cases, the

qualitative results remain the same.[35] Similar patterns are also found for optimal solutions in poor neighborhoods and mid-range neighborhoods. These results suggest that the main mechanisms of our findings are not significantly affected by the assumption of symmetric information between buyers and sellers.

Appendix E The Likelihood Function

In this section, we derive the likelihood function for our Maximum Likelihood Estimation. For each property, we observe its listing price (L_{it}), the Zestimate values (Z_{it}) and the Zestimate Ranges (ZR_{it}) in each period , the period in which it was sold (T_i), and the final selling price p_i.

We first solve the seller's problem characterized by Equation 23 and 24. Then with the observed listing prices, Zestimate values, and Zestimate Ranges, we infer the seller's belief about λ_i in each period (μ_{it}, σ_{it}). Let $S_{it} \in \{0,1\}$ denotes whether there is a new Zestimate signal in period t, then the likelihood of observing Zestimate Z_{it} when Zestimate range is ZR_{it} conditional on α_i is

$$l_{it}^{z}(\Theta|\alpha_i) = [\frac{1}{a_s + b_s ZR_{it} - c_s\sigma_v^n}\phi(\frac{(a_z + b_z \ln(Z_{it}) - c_z\sigma_v^n) - \lambda_i}{a_s + b_s ZR_{it} - c_s\sigma_v^n})]^{S_{it}}. \tag{41}$$

Note that with μ_{it} and σ_{it}, we can infer buyer visits (denoted as $A_{it} \in \{0,1\}$) and buyer valuations while a property stays on the market. Specifically, if $\mu_{it} \neq \mu_{i(t+1)}$ and there is no algorithm update in period t, then it implies that a buyer visits the property in period t (i.e., $A_{it} = 1$), and the buyer valuation is

$$v_{it} = \exp\{(\mu_{i(t+1)} - \mu_{it})\frac{(\sigma_{it}^n)^2 + \sigma_v^2}{\sigma_{it}^2} + \mu_{it}\}. \tag{42}$$

The probability that a buyer does not visit the property is

$$l_{it}^{n}(\Theta) = (1 - \delta_n) + \delta_n(1 - q_{it}). \tag{43}$$

where q_{it} is given by Equation 19. The likelihood of a buyer visiting the property with a realized valuation of v_{it} conditional on α_i is

$$l_{it}^{v}(\Theta|\alpha_i) = \delta_n q_{it} \cdot \frac{1}{\sigma_v^n}\phi(\frac{\ln(v_{it}) - \lambda_i}{\sigma_v^n}). \tag{44}$$

[35] The listing price curves in the alternative models are less smooth due to the lower precision in the numerical calculation with the added complexity of the model.

If a property is sold below the listing price, we know that a buyer visits and his valuation is

$$v_{iT_i} = \frac{p_i - (1-\theta)R_{iT_i}}{\theta}, \tag{45}$$

thus the likelihood of being sold below the listing price is

$$l_i^p(\Theta|\alpha_i) = \delta_n q_{it} \cdot \frac{1}{\sigma_v}\phi(\frac{\ln(v_{iT_i}) - \lambda_i}{\sigma_v}). \tag{46}$$

If a property is sold at the listing price, then we know that a buyer visits and his valuation satisfies

$$v_{iT_i} \geq \frac{L_{iT_i} - (1-\theta)R_{iT_i}}{\theta} \equiv \bar{v}_{iT_i}, \tag{47}$$

thus the likelihood of being sold at the listing price is

$$l_i^s(\Theta|\alpha_i) = \delta_n q_{it} \cdot (1 - \Phi(\frac{\ln(\bar{v}_{iT_i}) - \lambda_i}{\sigma_v})). \tag{48}$$

If a property is eventually withdrawn from the market without being sold, then we know that a buyer visits and her realized valuation leads to a belief about buyer valuations with which the seller's continuation value is less than his holding value:

$$\pi(\mu_{i(t+1)}(v_{iT_i}), \sigma^2_{i(t+1)}) < v_i^h, \tag{49}$$

and the likelihood of being withdrawn in the last period is

$$l_i^w(\Theta|\alpha_i) = \delta_n q_{it} \cdot \Phi(\frac{\ln(\hat{v}_{iT_i}) - \lambda_i}{\sigma_v}), \tag{50}$$

where $\hat{v}_{iT_i}$ solves

$$\beta\pi(\mu_{i(t+1)}(\hat{v}_{iT_i}), \sigma^2_{i(t+1)}) = v_i^h. \tag{51}$$

With the previous definitions, we construct the likelihood contribution of a property sold below the listing price T_i periods after being listed as

$$l_i^{p_i<L_{iT_i}}(\Theta) = \int [\prod_{t=1}^{T_i-1}(l_{it}^n(\Theta))^{(1-A_{it})} \cdot (l_{it}^v(\Theta|\alpha_i))^{A_{it}} \cdot (l_{it}^z(\Theta|\alpha_i))^{S_{it}}] \cdot l_i^p(\Theta|\alpha_i) \cdot (l_{iT_i}^z(\Theta|\alpha_i))^{S_{iT_i}} f(\alpha_i)\, \mathrm{d}\alpha_i, \tag{52}$$

the likelihood contribution of an observation sold at the listing price is

$$l_i^{p_i=L_{iT_i}}(\Theta) = \int [\prod_{t=1}^{T_i-1}(l_{it}^n(\Theta))^{(1-A_{it})} \cdot (l_{it}^v(\Theta|\alpha_i))^{A_{it}} \cdot (l_{it}^z(\Theta|\alpha_i))^{S_{it}}] \cdot l_i^s(\Theta|\alpha_i) \cdot (l_{iT_i}^z(\Theta|\alpha_i))^{S_{iT_i}} f(\alpha_i)\, \mathrm{d}\alpha_i, \tag{53}$$

and the likelihood contribution of an observation withdrawn from the market is

$$l_i^{p_i=NA}(\Theta) = \int [\prod_{t=1}^{T_i-1}(l_{it}^n(\Theta))^{(1-A_{it})} \cdot (l_{it}^v(\Theta|\alpha_i))^{A_{it}} \cdot (l_{it}^z(\Theta|\alpha_i))^{S_{it}}] \cdot l_i^w(\Theta|\alpha_i) \cdot (l_{iT_i}^z(\Theta|\alpha_i))^{S_{iT_i}} f(\alpha_i)\, \mathrm{d}\alpha_i, \tag{54}$$

Therefore the likelihood of observing all the properties in our sample is

$$L(\Theta) = \prod_i (l_i^{p_i<L_{iT_i}}(\Theta))^{\mathbb{1}(p_i<L_{iT_i})} \cdot (l_i^{p_i=L_{iT_i}}(\Theta))^{\mathbb{1}(p_i=L_{iT_i})} \cdot (l_i^{p_i=NA}(\Theta))^{\mathbb{1}(p_i=NA)}, \tag{55}$$

and the MLE parameter estimates are the ones that maximize the log-likelihood of observing the sample.

Appendix F Identification

We start with the identification of the variance of buyer valuations (σ_v^n). It can be identified from the variation in selling prices across properties: if multiple very similar properties with similar listing price trajectories end up with different selling prices, then the variation in selling prices among these properties reflects the variance of the realized buyer valuations (in the corresponding neighborhood groups). The prior variance (σ_0^n) is identified from the listing price updates. When sellers learn from previous buyer valuations or Zestimate, they update the listing price. For the identification of the prior beliefs, we focus on the learning from buyer valuations. How much the listing price changes in response to the realized buyer valuations reflects the weight sellers assign to the signals, and the weight is determined by the relative size of the prior variance and the variance of buyer valuations. If the prior variance is higher, then the weight given to the signal value would be higher, because sellers are less certain about their prior beliefs. With the variance of buyer valuation identified and the relative size of the prior variance and the variance of buyer

valuations inferred from the changes in the listing prices when there is no algorithm update, the prior variance is identified.

The initial prior mean belief ($\mu_{i1}^0 = \alpha_0 + \gamma \mathbf{X} + u_i$) can be inferred from the initial listing price and the initial Zestimate for each property individually. A seller decides the listing price based on his posterior belief in the first period, and the posterior belief depends on the initial prior mean belief and the initial Zestimate. We observe the listing prices and the Zestimates, so we can infer the prior mean belief for each property. If there are two similar properties in the same neighborhood group, with same Zestimate and Zestimate range, but different listing prices, then the difference in listing prices can only be attributed to the difference in prior mean. With the inferred initial prior mean beliefs (μ_{it}^0) and the observed home characteristics ($\mathbf{X}$), α_0 and γ can be identified in a similar manner to ordinary least squares (OLS) in a linear regression.

Next, we discuss the identification of Zestimate-related parameters. Note that we observe all the Zestimate values and the Zestimate Ranges, we need to estimate 6 parameters in the two linear transformations: the transformation from log Zestimate to Zestimate signal (a_z, b_z, c_z), and the linear transformation from Zestimate Range to standard deviation of Zestimate signal (a_s, b_s, c_s). We use both cross-sectional and intertemporal variation to identify the impact of Zestimate. First, we have cross sectional variation and can observe how initial listing price varies by Zestimate and Zestimate Range. Second, we have algorithm update shocks that give inter-temporal variation. With these algorithm updates, Zestimate values change suddenly and provide new signals that update sellers' beliefs.

Take the cross-sectional variation in Zestimate as an example. If multiple very similar properties in the same neighborhood have the same Zestimate Range, but different Zestimate values and different initial listing prices, then the difference in their initial listing prices can only be attributed to the difference in Zestimate values. Since for these properties, the Zestimate Ranges and therefore Zestimate variance are the same, the different impact of Zestimate on listing price across the properties will identify the b_z. That is, by comparing the Zestimate signal values and

the observed Zestimate values, we can identify b_z. Since properties in different neighborhoods have different buyer heterogeneity (σ_v^n), how the average impact of Zestimate on listing price across neighborhoods varies by buyer heterogeneity identifies c_z. Lastly, since all properties share the same intercept in the transformation of Zestimate values, the intercept a_z is identified by the average impact of Zestimate on listing prices.

Under a similar logic, we can identify the three parameters in the transformation of Zestimate Range: if multiple very similar properties in the same neighborhoods have the same Zestimate values, but different Zestimate Ranges and different initial listing prices, then the difference in their listing prices can only be attributed to the difference in Zestimate Ranges, which helps us identify b_s. How the average response in listing price to Zestimate across neighborhoods varies by buyer heterogeneity identifies c_s. And again, all properties share the same intercept in the transformation of Zestimate Ranges, therefore the intercept a_s is identified by the average response in listing price to Zestimate. Since Zestimate and Zestimate Range jointly influence listing price update, the 6 parameters governing the two linear transformations are jointly identified.

The parameter of holding values (ρ^n) is identified with the withdrawn properties in neighborhood group n. The fact that a property is withdrawn in period T_i implies that the holding value is less than the continuation value in period T_{i-1} but great than the continuation value in period T_i as a new signal(s) changes the belief about the distribution of buyer valuations in period T_i. The continuation values, inferred from listing prices with other parameters, help us identify ρ^n.

Visiting costs are identified through selling time. Everything else equal, a larger visiting cost leads to fewer buyer visits, and therefore longer selling time. Specifically, with other identified parameters, we can calculate the utility of visiting for each property in each time period. The average selling time identifies the mean of log visiting cost (c_0). If we compare two properties with different utility of visiting, then the difference in their selling time would be smaller when the variance of visiting cost is larger, and vice versa. This helps us identify σ_c. Lastly, since all the properties share the same distribution of (relative) visiting costs, the systematic difference in selling time across neighborhood groups would be attributed to the difference in market thickness, which identifies the probability of buyer arrival in poor and rich neighborhoods (δ_P, δ_R).