Algorithmic Outputs as Information Source: The Effects of Zestimates on Home Prices and Racial Bias in the Housing Market

Shuyi Yu

May 10, 2021

Abstract

This paper investigates market participants' reactions to predictive algorithms and the effects of this public information source on market outcomes. In particular, I study the extent to which buyers and sellers rely on the Zestimate, Zillow's estimate of a home's market value, as well as the interactions between the Zestimate and other information sources. Using detailed property transaction data for 120,482 properties sold between May 2017 and May 2019 in the Greater Philadelphia Area, I show that the sale price of a property does respond to exogenous shocks to its estimated home value. I develop a theoretical framework and provide empirical evidence to show how people use the Zestimate as a source of publicly available information that plays an important role in coordination and helping people reach an agreement. The results suggest that market participants tend to rely more on this public information when it is harder to reach a consensus based on private information. Moreover, I show that people's reliance on the Zestimate might mitigate racial disparities in the housing market by providing less biased information.

Electronic copy available at: https://ssrn.com/abstract=3584896

1 Introduction

There has been an increasing interest in how algorithms have reshaped the economy (Bughin et al., 2018). Breakthroughs in machine learning techniques make automated decision-making available for many giant players in the economy. For example, sharing economy companies such as Uber and Lyft dynamically adjust their prices based on the data-driven real-time pricing system that utilizes information from both supply and demand sides.[1] E-commerce sites go even further by adopting AI-powered demand forecasting tools to automate restocking of products.[2]

But more fundamentally and profoundly, algorithms may also affect economic outcomes via their influences on human decisions. One way algorithms can change human behaviors is by altering the information presented to decision-makers. For example, sophisticated ad targeting algorithms have tailored the information presented to consumers, which has proved crucial for consumers' ability to make good decisions (Payne et al., 1991). Moreover, algorithms can provide market participants with novel information sources by processing massive information and making predictions and recommendations. This use of algorithms has appeared in various domains, such as travel agencies, matchmaking service, and financial advisory service. However, it is still controversial whether people are in adherence to algorithmic forecasts (Dietvorst et al., 2015; Logg et al., 2019).

To answer this question empirically, I study real estate market the reaction of real estate market participants to the Zestimate in this paper. The Zestimate is Zillow's estimate of a home's market value based on its home valuation model, which incorporates data from multiple sources, taking into account home facts, location, tax information and market conditions.[3] It shows right below the current list price (the most recent sale price if the property is not on the market currently) on the property page (see Figure 1)[4] and it has

[1] See `https://www.forbes.com/sites/nicolemartin1/2019/03/30/uber-charges-more-if-they-think-youre-willing-to-pay-more`

[2] See `https://www.npr.org/2018/11/21/660168325/optimized-prime-how-ai-and-anticipation-power-amazons-1-hour-deliveries`

[3] See the presentation by Zillow's data scientist: `https://www.slideshare.net/NicholasMcClure1/python-datascienceatzillow/1`

[4] It has been moved to listing details after a major change came into effect in Sep 2019.

Electronic copy available at: https://ssrn.com/abstract=3584896

been displayed for 97.5 million homes out of the 110 million homes found on Zillow.com, the most popular real estate website in the United States [5]. Anecdotal evidence has shown that even though Zillow is commonly used by both home sellers and buyers in the U.S., it is not clear ex-ante whether Zillow or its home price estimates affect home-buying decisions. On one hand, it provides public and easy-to-process information that may help market participants evaluate and compare home values. On the other hand, the real estate decision is a stressful major financial decision and it is unknown whether people will still trust online information sources and algorithmic estimation techniques when making this important decision.

I combine 120,482 property transaction records with the Zestimate history collected from Zillow.com. An obvious endogeneity concern in this setting is that Zestimates may reveal the unobserved quality of a property. To address this, I turn to an instrumental variables approach, where I use the number of months since the last revaluation or reassessment as a plausibly exogenous instrument. The idea here is that the time since the last reassessment should affect the severity of covariate shifts but the differences in the frequency of revaluation and reassessment across townships are jointly decided by many forces, like laws and budget plans, which are not affected by the changes in current sale prices after controlling the current assessed value and various fixed effects. Empirical evidence proves the validity of this instrument and shows how the covariate shift problem affects the model performance. Using this approach, I find that the final transaction price tends to be higher when the estimated home value displayed on Zillow.com is higher.

Furthermore, I investigate how the algorithmic estimation changes users' information acquisition process by providing information that is available to all the market participants. The empirical results show that the reliance on Zestimates is correlated with users' costs of acquiring information from other sources but not with the perceived average accuracy of the Zestimate in the neighborhood.

Finally, I explore the heterogeneity in the effect and ask whether the algorithm helps eliminate the racial biases existing in the housing market for a long time. Due to the lingering impacts of historical "redlining", the properties located in minority or more diversified

[5] https://investors.zillowgroup.com/overview/default.aspx

Electronic copy available at: https://ssrn.com/abstract=3584896

neighborhoods are usually undervalued. I find suggestive evidence that the Zestimate doesn't fully reflect this white-premium in home values. Since the Zestimate's influence on decision making doesn't vary much across neighborhoods, this effect leads to a smaller racial gap in final sale prices compared to the gap in list prices.

2 Literature Review

This paper is related to four streams of research. The first is the literature on information search and information sources. Early work in marketing studying the prepurchase information search and acquisition (e.g. Newman and Staelin, 1972; Claxton et al., 1974; Westbrook and Fornell, 1979; Schaninger and Sciglimpaglia, 1981; Kiel and Layton, 1981; Hauser et al., 1993) focuses on how different customers determine their total information-seeking effort and the allocation of effort among information sources. More recent research by Ratchford et al. (2003) studies how the Internet as a new information source reshapes the information acquisition process by substituting other information sources, especially the dealer/manufacturer sources. Zettelmeyer et al. (2006) further extends the discussion and shows that the Internet lowers the negotiated prices in car retailing markets by providing buyers more purchase-relevant information. Kuruzovich et al. (2010) instead looks at the seller side and shows that lower search costs facilitated by the Internet also equip sellers with the ability to search for high-valuation buyers and raises the final sale price. Other studies (e.g. Brown and Goolsbee, 2002; Jensen, 2007; Ellison and Ellison, 2009) discuss the impact of IT on market structure and efficiency. This paper contributes to the literature by investigating the impact of a specific data product on the allocation of attention and offline market outcomes.

The second is a stream of research that focuses on home prices and racial differentials in housing markets. Previous research has shown that many factors affect the final transaction prices, e.g. school quality (Black, 1999), marketing platforms (Hendel et al., 2009), agent characteristics (Seagraves and Gallimore, 2013), and policy changes (Tucker et al., 2013). Most importantly, significant racial disparities have been found in the US housing market. Individual black buyers tend to pay premiums for comparable units (King and Mieszkowski,

Electronic copy available at: https://ssrn.com/abstract=3584896

1973; Myers, 2004; Ihlanfeldt and Mayock, 2009; Bayer et al., 2017) and this racial discrimination even persists in emerging online rental markets (Edelman et al., 2017; Cui et al., 2020). On the other hand, house values have been proven to decline in neighborhoods as the percentage of blacks increases (Berry, 1976; Chambers, 1992; Kiel and Zabel, 1996; Myers, 2004) as the consequence of racial prejudice. Results in this paper provide new evidence for racial differentials in home prices caused by prejudice and suggest that this gap can be mitigated by less biased home value estimates based on data-driven methods.

The last one is emerging literature on algorithms and biases. Even though evidence has shown that algorithms reproduce existing racial and gender disparities in various applications (Angwin et al., 2016; Ali et al., 2019; Lambrecht and Tucker, 2019; Obermeyer et al., 2019), it is also important to compare bias between automated algorithms and human judges or other benchmarks (for a recent survey see Cowgill and Tucker, 2020). In this study, I focus on a specific prediction model and study its effects on human prejudice in decision-making.

3 Data

I use data collected from three sources: Zillow.com, local government's property records, and social-demographic data from the Decennial Census administered by the Census Bureau.

3.1 Zillow Data

First of all, I use the property transaction data collected from Zillow.com. I collect detailed property transaction information for 209,016 properties (excluding lots and commercial buildings) sold between May 2017 and May 2019 in the Greater Philadelphia Area. 372 zip codes in 4 states (PA, NJ, DE, MD) are included in this sample. I am able to find the geographic location for 168,818 out of them using their addresses[6] and keep only these properties in the data set to guarantee the data accuracy. To avoid the extreme cases of predatory pricing and the potential threats of misrecorded information, I also drop the 45,141 cases where the list price is missing or equals zero. Finally, I exclude the observations where the sale-to-list is too large (greater than the 99 percentile) or too small (less than the

[6]See `https://geocoding.geo.census.gov`.

Electronic copy available at: https://ssrn.com/abstract=3584896

1 percentile). It excludes about 2,400 properties from the data set and reduces the total number of observations to 120,482.

Figure 1: Zillow Listing Page

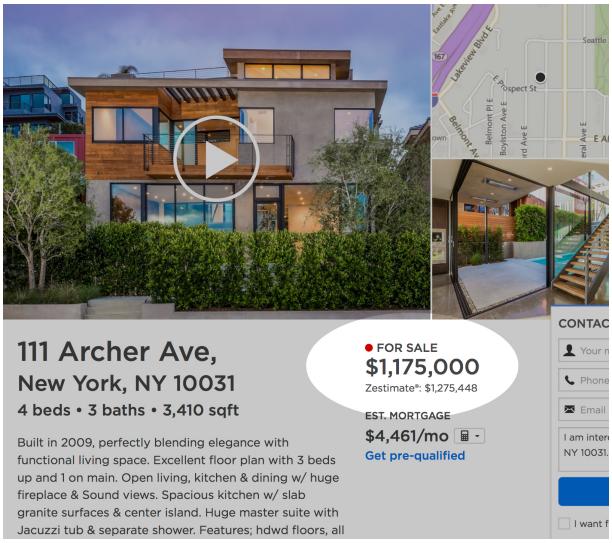


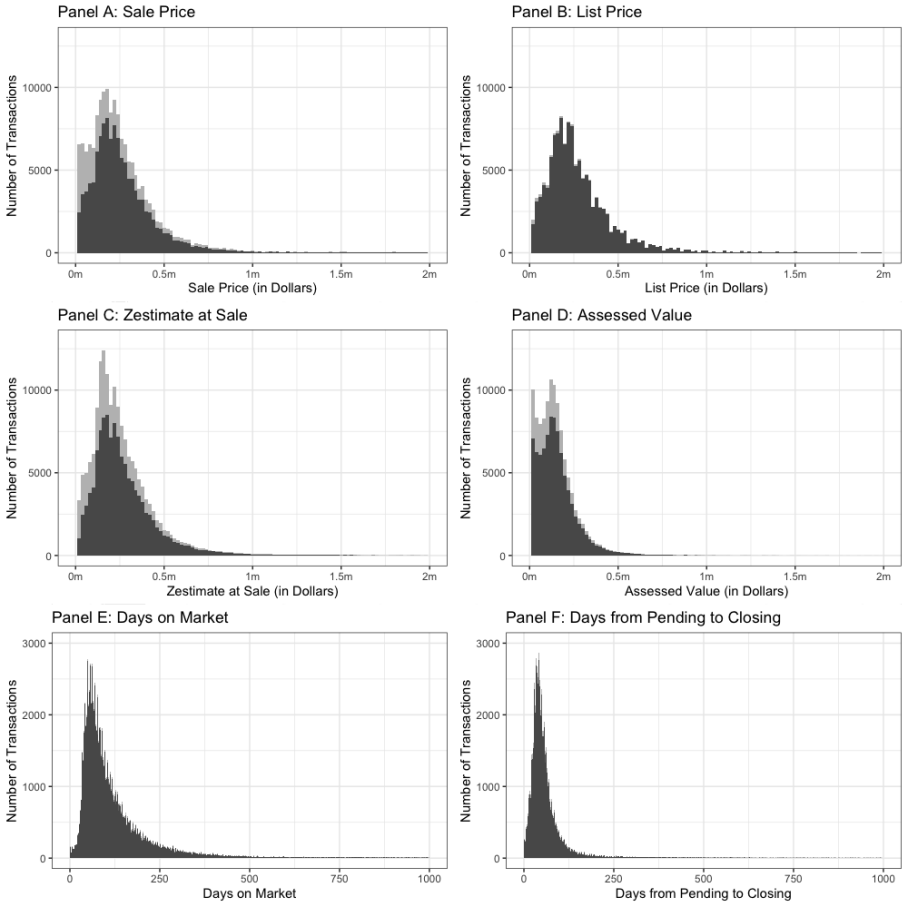
Table 1: Summary of Variables

Variable	Mean	Std. Dev.	Min.	Max.	N
Independent and Dependent Variables					
Sale Price	270,474.3	213,814.5	1,000	8,000,000	120,482
List Price	286,383.3	237,352.3	750	1,000,000	120,482
Zestimate_Sold	277,646.1	252,352.5	5,390	4,470,000	120,482
Assessed Value	154,687.5	119,269.7	0	4,108,700	88,110
Log(Days on Market)	4.637	0.905	0	8.369	120,482
Log(Days from "Pending" to "Sale")	3.905	0.817	0	8.066	68,490
Instrument					
Months_Last_Update	184.116	169.186	0	562	120,482
Moderators					
#Transactions	0.007	0.007	0.000	0.103	120,482
%Deviation	0.157	0.563	0	68.093	112,747
Important Sociodemographic Variables					
#Years of Education	13.857	1.273	8.182	17.938	120,482
Log(Median Income)	11.179	0.509	8.849	12.391	118,210
%Internet Subscription	0.826	0.136	0	1	120,482
%Computer Ownership	0.890	0.099	0.148	1	120,482
%White	0.709	0.279	0	1	120,482
%White Owners	0.740	0.274	0	1	120,482

The transaction details collected include the address, the listing date, the list price, the assessed value, the date the seller accepts the offer, the closing date, the sale price, and

Electronic copy available at: https://ssrn.com/abstract=3584896

Figure 2: Transaction Details



Notes: The distribution of sale prices, list prices, the Zestimates at sale, and assessed values are plotted in Panel A, B, C, D, respectively. These distributions are truncated at 2M. The distribution of numbers of days on market and numbers of days from pending to closing are plotted in Panel E and F, respectively. These distributions are truncated at 1000. The light grey bars represent observed frequencies in the entire sample (with 168,818 observations) and the dark grey bars represent observed frequencies in the selected sample (with 120,482 observations) The census block groups included in the sample are shown in grey.

any price changes that happened in between. In addition, I collect the historical Zestimates for the properties included in the data set. Monthly Zestimates in the last 5 years are available on Zillow.com. Based on the information provided by Zillow, the median error of Zestimates is 1.9% and more than 1.8M homes have been included in the model in the city of Philadelphia, which is very close to the national average. And there is no evidence showing

Electronic copy available at: https://ssrn.com/abstract=3584896

that Zestimates are more accurate in more active markets or more metropolitan areas [7]. In Figure 2, I plot the distribution of the sale price, the list price, the Zestimates[8], the assessed value, the number of days on market, and the number of days from pending to closing. All the distributions are right skewed and it seems that the Zestimate model is pretty accurate while assessed values don't fully reflect market values of those properties. The summary statistics are reported in Table 1.

The information collected from Zillow.com also includes many exterior and interior features of properties, including but not limited to the size, the property type, the number of bedrooms, the number of bathrooms, the exterior material, the exterior and interior amenities, the view, the cooling/heating conditions, the heating conditions, the appliances, and the flooring conditions (see Table A1 for a full list of variables).

3.2 Census Data

The second data set I use is the neighborhood-level socioeconomic characteristics collected from the Decennial Census administered by the Census Bureau. The properties are matched with the census block group level data using the street address and the social-demographic characteristics are found for the 4,108 census block groups they belong to. Those block groups are highlighted on a map in Figure A1.

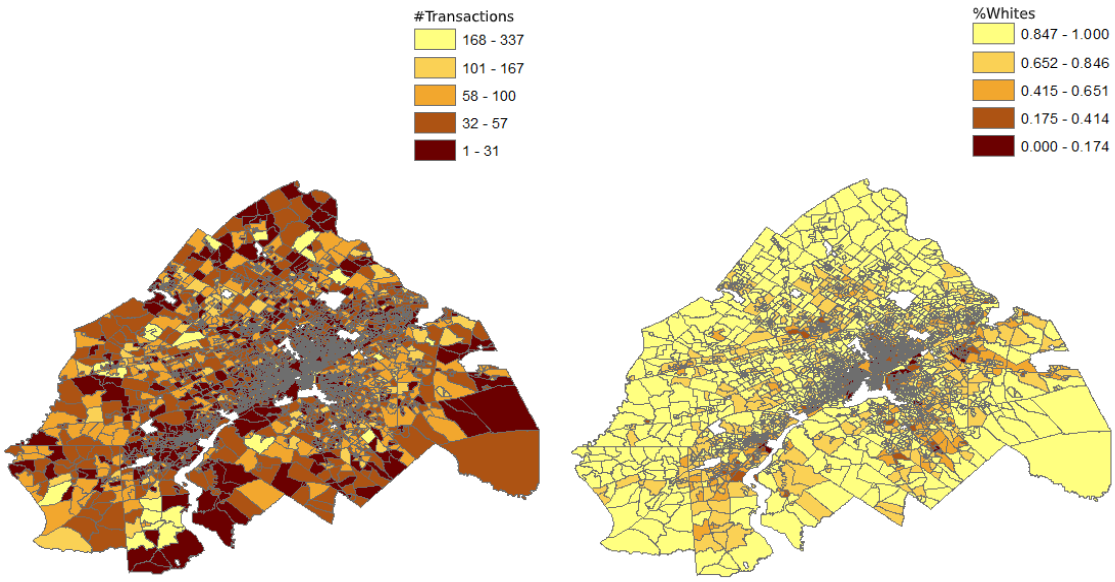
The characteristics used in this study include but not limited to the population density, the gender distribution, the age distribution, the race and ethnicity distribution, the geographical mobility, the place of work, the commute methods, the one-way commute time, the household status, the household size, the average education level, the property status, the median household income, the income sources, the number of housing units, the race and ethnicity distribution for homeowners, the distribution of the number of bedrooms in a property, and the home type distribution (see Table A2 for a full list of variables). In Figure 2, I plot the number of observations (Panel A) and the share of white residents (Panel B) in each block group. The graph shows the segmentation in the housing market: the suburbs are whiter than the city center. However, the transaction frequency is not aligned with the

[7] See `https://www.zillow.com/zestimate/`

[8] The Zestimate displayed one month before the sale is used for the plot.

Electronic copy available at: https://ssrn.com/abstract=3584896

Figure 3: Geospatial Distribution



(a) Transaction Frequency (b) Share of Whites

Notes: The number of transactions for each census block are plotted in Panel A and the share of whites is plotted for each census block in Panel B.

difference in racial markup and the market is active in some more diverse neighborhoods. The summary statistics for some important sociodemographic variables are reported in Table 1.

3.3 Assessment Information and Public Records

Finally, I collect the assessment information from the local government's website and newspapers. The time of the last revaluation is found for most towns (this information is missing for only 7 out of 529 towns (cities) included in the data set). This time varies from 0 months to 562 months as places like the city of Philadelphia and the state of Maryland conduct a regular reassessment every three years while in other places, such as Buck County, the assessed valuation of property has not been updated since 1970s.

The property records are also collected as a supplementary data set from the assessor's website for most of the properties located in New Jersey and Pennsylvania. Unfortunately, the property records are not available to the public in Delaware, Maryland, and Chester

Electronic copy available at: https://ssrn.com/abstract=3584896

County in Pennsylvania. These public records are matched with the Zillow data using the street address. The most important variable I collect from this supplementary data set is the name of the current owner of a property. I only keep the most recent transaction for a property in the data set if it was traded multiple times during our time window. So I can identify the buyers for the transactions included in the data set using the owners' information. Using a prediction model that exploits the US census data, I am able to predict a buyer's race and ethnicity based on her last name (*ethnicolr*0.2.1). Gender is predicted based on the first name as well using prediction models that utilize the Social Security data sets (*gender*). If there are two owners owning the property jointly, I collect these variables for both of them. Moreover, the properties owned by firms are identified and later removed from the analysis.

4 The Effect of Zestimates on Home Prices

4.1 Model

The analysis focuses on the effect of changes in the Zestimate on the final sale price because this market outcome is a natural measure of how market participants react to the statistic.

The sale price for home i in census block k listed in month t is modeled as:

$$\begin{aligned} Sale\_Price_{itk} =& \alpha + \beta Zestimate_{i,t-1} + \gamma List\_Price_i + \theta Assessed\_Value_{it} + \delta Log(Days \\ & \_on\_Market)_i + \lambda Log(Days\_from\_\text{“}Pending\text{”}\_to\_\text{“}Sold\text{”})_i + \mu_t + \eta L_k \\ & + \zeta S_i + \epsilon_{ikt}, \qquad (1) \\ Zestimate_{i,t-1,k} =& \alpha' + \kappa Months\_Last\_Update_{i,t-1} + \gamma' List\_Price_i + \theta' Assessed\_Value_{it} \\ & + \delta' Log(Days\_on\_Market)_i + \lambda' Log(Days\_from\_\text{“}Pending\text{”}\_to\_\text{“}Sold\text{”})_i \\ & + \mu'_t + \eta' L_k + \zeta' S_i + \epsilon'_{ikt}. \qquad (2) \end{aligned}$$

$Zestimate_{i,t-1}$ is the Zestimate in month t-1. I use the lagged Zestimate here to isolate the effect of estimated market values on home prices from final sale prices' impacts on prediction outcomes, and β is the parameter of interest that shows the magnitude of this effect. I use

Electronic copy available at: https://ssrn.com/abstract=3584896

the original list price ($List\_Price_i$), the assessed value ($Assessed\_Value_{it}$), and the number of days between listing and pending sale ($Log(Days\_on\_Market)_i$) to partially control the unobserved quality of the property and market condition. The number of days from the pending sale to closing ($Log(Days\_from\_\text{“}Pending\text{”}\_to\_\text{“}Sold\text{”})_i$) is added into the model as well to control the unobserved quality of the buyer (like whether they are making an all-cash offer or a mortgage offer and the uncertainties in eventually receiving a mortgage) and the property. The distribution of these explanatory variables is plotted in Figure 2. Moreover, μ_t is a vector of month indicators that control the month fixed effects. L_k is a vector of social-demographic controls at the block group level (see Table A2 for a detailed list of variables) and S_i is a vector of controls related to home features (see Table A1 for a detailed list of variables). ϵ_{ikt} is the idiosyncratic error term.

Given the nature of the predictive algorithm, there may be concerns over the endogeneity of Zestimates. It is because that the advanced prediction models Zillow uses may better reflect the unobserved quality of a property, even though I have controlled the important features used by Zillow in the linear model.[9]

As stated in Equation (2), I use 2SLS regressions where the lagged Zestimate is instrument with the number of months since the last revaluation or reassessment at time $t-1$ ($Months\_Last\_Update_{i,t-1}$, the summary statistics are reported in Table 1) to address those potential endogeneity issues. The idea is that the new observations a Zestimate is based on are more likely to be different from the data used by Zillow for the model training when a reassessment has been done recently. Therefore, the prediction errors caused by dataset shifts will be more severe if the reassessment frequency is higher. In particular, both the covariate shift in the assessed value itself and the shift in its relationship with the property's market value and other covariates such as property features will be larger if the assessed values are updated more frequently. Those shifts will exacerbate the estimation bias in the model (Kanamori and Shimodaira, 2009), and the first-stage results reported in Table 3 confirm this hypothesis.

[9]The variables used by Zillow can be found from the Kaggle competition hosted by them. See `https://www.kaggle.com/c/zillow-prize-1`.

Electronic copy available at: https://ssrn.com/abstract=3584896

Moreover, the difference in the frequency of revaluation and reassessment across townships is not correlated with the current sale prices (after controlling the assessed value) but jointly decided by many forces, like laws and budget plans. For example, in the city of Philadelphia and in Maryland, the assessed values have to be updated every three years by law while they have not been updated for more than 20 years in some other parts of Philadelphia and New Jersey. To illustrate the exclusion restriction assumption, I plot the geospatial distribution of the dependent variable (sale price), the independent variable of interest (the Zestimate) , and the instrumental variable (the number of months since the most recent reassessment) in Figure A2. In particular, the darker the census block is in Panel A, the longer the time since the last assessment update is in the area. The segmentation in the time since the last reassessment caused by policy differences suggests that it is unlikely to be related to the sale price besides through the Zestimate's effect.

4.2 Results

The 2SLS results are reported in Panel A of Table 2. Column (1) in the table presents the result of a regression that includes only the Zestimate at month $t-1$ as the independent variable. The original list price, the assessed value, $Log(Days\_on\_Market)_i$, $Log(Days\_from\_\text{“}Pending\text{”}\_to\_\text{“}Sold\text{”})_i$, the month fixed effects, the socio-demographic controls, and property feature controls are added into the model incrementally in Columns (2)-(6). The standard errors are clustered at city×month level to control the correlation between sales.

As we can see, the final sale price is significantly affected by the Zestimate displayed on Zillow.com – particularly, based on Column (6), on average a property is going to be sold 0.205 dollars higher if its Zestimate increases by one dollar. If we assume that the exogenous shock caused by covariate shifts is constant after controlling the variables used in the prediction model, this estimator estimates a weighted average of conditional average treatment effects. It suggests that the popular market value model does have an influence on transactions, apart from reflecting the unobserved quality and market conditions. Moreover, the final sale is more likely to be higher if the list price is higher, ceteris paribus, since the list

Electronic copy available at: https://ssrn.com/abstract=3584896

Table 2: The Effect of Zestimates on Final Sale Prices

Panel A: Full Sample						
	(1)	(2)	(3)	(4)	(5)	(6)
	Sale Price	Sale Price	Sale Price	Sale Price	Sale Price	Sale Price
Zestimate_Sold	0.998***	0.507***	0.337***	0.335***	0.270***	0.231***
	(0.008)	(0.064)	(0.048)	(0.048)	(0.049)	(0.054)
List Price		0.439***	0.599***	0.600***	0.639***	0.658***
		(0.060)	(0.045)	(0.045)	(0.044)	(0.048)
Assessed Value		-0.046***	-0.024***	-0.023***	-0.023***	-0.031***
		(0.004)	(0.003)	(0.003)	(0.003)	(0.004)
Log (Days on Market)			-12967.583***	-12778.470***	-12716.006***	-12918.462***
			(614.912)	(608.319)	(576.655)	(609.640)
Log (Days from "Pending" to "Sold")			5149.837***	5045.879***	5224.922***	5257.542***
			(364.152)	(359.637)	(362.475)	(369.564)
Month Fixed Effects	No	No	No	Yes	Yes	Yes
Socio-Demographic Controls	No	No	No	No	Yes	Yes
Property Feature Controls	No	No	No	No	No	Yes
Observations	120,482	88,110	52,981	52,981	52,387	52,164
Panel B: Subsample without Missing Information						
	(1)	(2)	(3)	(4)	(5)	(6)
	Sale Price	Sale Price	Sale Price	Sale Price	Sale Price	Sale Price
Zestimate_Sold	1.010***	0.397***	0.344***	0.342***	0.277***	0.231***
	(0.004)	(0.050)	(0.051)	(0.051)	(0.052)	(0.054)
List Price		0.545***	0.592***	0.593***	0.633***	0.658***
		(0.047)	(0.048)	(0.048)	(0.047)	(0.048)
Assessed Value		-0.036***	-0.024***	-0.023***	-0.024***	-0.031***
		(0.003)	(0.003)	(0.003)	(0.003)	(0.004)
Log (Days on Market)			-12828.849***	-12641.816***	-12617.347***	-12918.462***
			(624.945)	(618.303)	(595.072)	(609.640)
Log (Days from "Pending" to "Sold")			5119.504***	5016.118***	5225.440***	5257.542***
			(369.130)	(364.367)	(365.514)	(369.564)
Month Fixed Effects	No	No	No	Yes	Yes	Yes
Socio-Demographic Controls	No	No	No	No	Yes	Yes
Property Feature Controls	No	No	No	No	No	Yes
Observations	52,164	52,164	52,164	52,164	52,164	52,164

Robust standard errors reported in parentheses are clustered at city×month level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The dependent variable is sale price. The time between listing and pending sale and the time between pending to closing are log-transformed to control their nonlinear effects. The regressions reported in Panel A use the entire sample for estimation and the regressions reported in Panel B use only the observations without missing information.

price implies not only the quality of the property but also the seller's private information. It also makes sense the assessed value is negatively correlated to the sale price after controlling the property features and other conditions, since the property tax proportional to it is sometimes a major burden for the owner. As we can see from the table, the longer the property is on the market the lower the final sale price is: unpopular properties are more likely to be low quality and an unlucky seller who sells her house in a buyer's market has to lower the price. It is also consistent with the findings in Tucker et al. (2013). Reversely, buyers are more likely to receive a risk premium if the time between the pending sale and the final closing is longer.

Electronic copy available at: https://ssrn.com/abstract=3584896

Table 3: First Stages

Panel A: Full Sample						
	(1)	(2)	(3)	(4)	(5)	(6)
	Zestimate	Zestimate	Zestimate	Zestimate	Zestimate	Zestimate
Months_Last_Update	265.854***	163.572***	118.466***	118.341***	109.419***	96.786***
	(11.300)	(15.600)	(10.673)	(10.672)	(10.359)	(10.149)
List Price		0.748***	0.801***	0.800***	0.775***	0.765***
		(0.020)	(0.013)	(0.013)	(0.015)	(0.016)
Assessed Value		0.307***	0.223***	0.223***	0.199***	0.168***
		(0.031)	(0.022)	(0.022)	(0.021)	(0.021)
Log (Days on Market)			-10102.850***	-9828.774***	-8989.805***	-9043.551***
			(558.297)	(547.414)	(515.036)	(494.792)
Log (Days from "Pending" to "Sold")			3719.434***	3579.912***	3569.213***	3485.599***
			(370.919)	(370.504)	(358.610)	(350.474)
Month Fixed Effects	No	No	No	Yes	Yes	Yes
Socio-Demographic Controls	No	No	No	No	Yes	Yes
Property Feature Controls	No	No	No	No	No	Yes
Observations	120,482	88,110	52,981	52,981	52,387	52,164
F statistic	3939.31	980.59	2397.47	2391.47	1877.90	1398.43
R^2	0.032	0.589	0.928	0.928	0.928	0.931
Adjusted R^2	0.032	0.589	0.928	0.928	0.928	0.931
Panel B: Subsample without Missing Information						
	(1)	(2)	(3)	(4)	(5)	(6)
	Zestimate	Zestimate	Zestimate	Zestimate	Zestimate	Zestimate
Months_Last_Update	288.451***	116.164***	111.237***	111.105***	104.950***	96.786***
	(13.704)	(10.935)	(10.908)	(10.910)	(10.480)	(10.149)
List Price		0.804***	0.808***	0.807***	0.781***	0.765***
		(0.013)	(0.013)	(0.013)	(0.015)	(0.016)
Assessed Value		0.204***	0.205***	0.206***	0.190***	0.168***
		(0.022)	(0.022)	(0.022)	(0.021)	(0.021)
Log (Days on Market)			-10025.264***	-9774.361***	-9000.666***	-9043.551***
			(555.432)	(544.513)	(515.616)	(494.792)
Log (Days from "Pending" to "Sold")			3754.129***	3629.701***	3625.058***	3485.599***
			(366.551)	(365.620)	(354.691)	(350.474)
Month Fixed Effects	No	No	No	Yes	Yes	Yes
Socio-Demographic Controls	No	No	No	No	Yes	Yes
Property Feature Controls	No	No	No	No	No	Yes
Observations	52,164	52,164	52,164	52,164	52,164	52,164
F statistic	2680.57	2293.42	2107.03	2100.93	1732.58	1398.43
R^2	0.049	0.927	0.928	0.928	0.930	0.931
Adjusted R^2	0.049	0.927	0.928	0.928	0.930	0.931

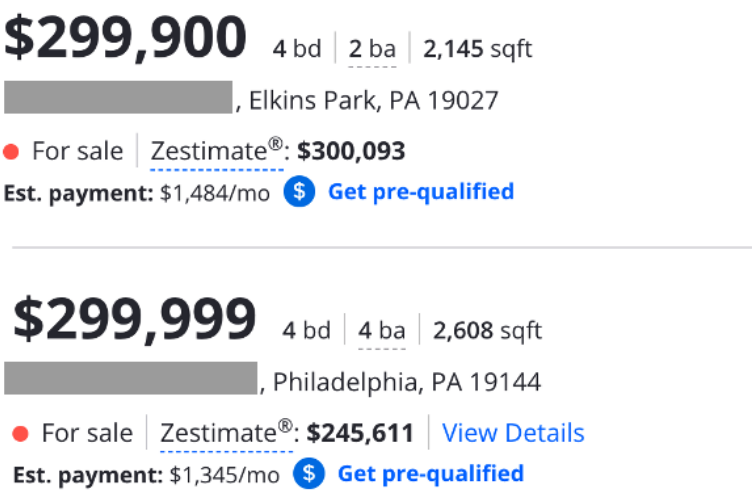
Robust standard errors reported in parentheses are clustered at city×month level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The dependent variable is sale price. The time between listing and pending sale and the time between pending to closing are log-transformed to control their nonlinear effects. The regressions reported in Panel A use the entire sample for estimation and the regressions reported in Panel B use only the observations without missing information.

The first stages are reported in Panel A of Table 3. The Wald F statistic always suggests a strong first stage and the number of months since the last update has a significant impact on the Zestimate displayed. Based on the results reported in Column (6), the Zestimate will be more than 92 dollars higher if the assessed value was updated one month earlier. In other words, Zestimates are more biased (i.e., underestimated, since the Zestimate is on average smaller than the final sale price) in those areas where assessed values are updated more frequently when the model used allows heterogeneity. It can be explained by the additional

Electronic copy available at: https://ssrn.com/abstract=3584896

biases caused by covariate shifts as discussed before. To further illustrate it, in Figure 4 I show the estimated market values of two similar properties located next to each other. The first one is located in Elkins Park, PA, and the last time its assessed value got updated was in 1996 (the current assessed value is $118,860). Even though the second home is only ten-minute away from the first one, it is located in the city of Philadelphia, which means that its assessed value is updated every three years (the current assessed value is $148,200). Despite the second home is larger and have more bathrooms, its Zestimate is significantly lower than that of the first home. This example shows how the impact of reassessment frequency on the Zestimate.

Figure 4: First Stage: Illustrative Examples



The overall explanatory power of the first stage model represented by the R-squared is also reasonably high – the full model can explain 93% of the variability. Consistent with how Zestimates are computed, the original list price, the assessed value, the neighborhood socio-demographic features, and the property features play a significant role in explaining the variation in Zestimates. $Log(Days\_on\_Market)$ and $Log(Days\_from\_\text{“}Pending\text{”}\_to\_\text{“}Sold\text{”})$ also have significant effects because the effect of list price decays and the pending price which is recorded as the usually higher original list price is more likely to enter the model when it takes a longer time to finalize the deal.

Electronic copy available at: https://ssrn.com/abstract=3584896

4.3 Robustness Checks

To check the robustness of the results, I first replicate those results using a sub-sample where all the observations with missing values are excluded from the analysis, which is the same sub-sample as the one used in Column (6) of Panel A in both Table 2 and 3. The replication results reported in Panel B of the corresponding table are very close to the ones estimated using the full sample. It suggests that the missing information doesn't reflect anything fundamental nor leads to biased results.

Another concern regarding the identification is that Zestimates displayed at the time the data was collected are different from the Zestimates displayed when buyers were making their purchase decisions. Though Zillow makes a major improvement in how they calculate their Zestimates every year or two[10], I find no evidence suggesting that they update the historical Zestimates as well. However, if it is true, the fact that the Zestimate is predicted by a model trained using the final sale price will undermine the identification results here. So I check the robustness of the model by replicating the results using only recent sales. The idea here is that the more recent the sale was the less likely it will be included in the training data set. I present the results for properties sold less than one month before the data collection in Panel A of Table A3 and for those properties sold less than three months before the data collection in Panel B of Table A3. Despite the data sparsity, we can still observe a positive effect and the effect size is not different from the one reported in Table 2 at a significance level of 0.05. The first stages reported in A4 are also similar to the ones for the main model.

Finally, I check the model specifications by replacing the social demographic controls with city fixed effects. So the variation (in both reassessment frequency and market prices) across cities is fully captured by the fixed effects. The results reported in Appendix (Tables A5 an A6) are close to those from the main model. It also further validates the exclusion restriction assumption.

[10]See `https://www.forbes.com/sites/johnwake/2019/06/30/new-zillow-zestimate-accuracy/\#5548c2a28a07`

Electronic copy available at: https://ssrn.com/abstract=3584896

5 Mechanism: Zestimates as a Public Source of Information

The Zestimate home valuation, as a summary statistic from a popular online real estate database, affects how real estate market participants acquire and process information. The sellers (buyers) often need to search market information before making the sell (purchase) decision and there are various information sources available to them, like advertising, word-of-mouth, internet, and even a trial sale (purchase). Particularly, there is an enormous need for information when making a major financial decision such as home sale and purchase and people spend tremendous time and effort acquiring and processing necessary information. For example, sellers and buyers usually hire professional agents for advice and register in the MLS system to receive notifications about new listings. In addition, buyers also go to open houses to see properties in person, search-related information (like crime maps and recently sold properties) online for their reference, and ask friends and colleagues for suggestions. In this section, I extend the basic model presented in the previous section and study how the market value estimated using algorithms differ from other information sources.

I find that people are more likely to rely on the Zestimate home valuation when there is a larger set of private signals. I suggest that it is because users view the estimated market value as a public source of information that other parties can also observe, and relying on this summary is more cost-efficient when it becomes harder to reach a consensus with private information sources.[11] A theoretical model that explains the underlying information acquisition process is presented in Appendix.

5.1 Subgroup Analysis

Comparable sales are one of most important private signals in the housing market. Usually agents will go through all the comparable sales and compare them with the focal property before they make price suggestions for their clients. Those signals are private because different people have different opinions on the same set of properties and it may lead to different

[11] Because people are less likely to choose to observe the same set of private signals when the choice set becomes larger if there is a constraint on the number of signals they can observe.

Electronic copy available at: https://ssrn.com/abstract=3584896

Table 4: Mechanism: Interaction with Number of Comparable Sales

Panel A: Interaction with Frequency of Similar Transactions Nearby ($t-6$)								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Area	Block Group				City			
Measurement	per Capita		per Housing Unit		per Capita		per Housing Unit	
Zestimate_Sold	0.345***	0.240***	0.349***	0.239***	0.364***	0.242***	0.351***	0.240***
	(0.049)	(0.054)	(0.049)	(0.055)	(0.053)	(0.055)	(0.048)	(0.054)
# Transactions	-1275060.638***	-436734.618*	-548333.146***	-164017.904	-1365560.718*	-370443.009***	-386168.813***	-129545.705***
	(291592.872)	(244943.577)	(151556.395)	(111914.450)	(716143.083)	(140184.685)	(112902.358)	(44622.119)
Zestimate_Sold	3.755***	2.424***	1.702***	0.948**	3.499*	1.076***	0.996***	0.390***
× # Transactions	(1.018)	(0.895)	(0.486)	(0.375)	(1.823)	(0.345)	(0.283)	(0.110)
List Price	0.585***	0.645***	0.580***	0.645***	0.565***	0.646***	0.579***	0.648***
	(0.046)	(0.048)	(0.046)	(0.048)	(0.051)	(0.048)	(0.045)	(0.048)
Assessed Value	-0.024***	-0.030***	-0.024***	-0.030***	-0.029***	-0.033***	-0.027***	-0.033***
	(0.003)	(0.004)	(0.004)	(0.004)	(0.005)	(0.004)	(0.004)	(0.004)
Log (Days on Market)	-12825.156***	-12813.274***	-12809.992***	-12823.744***	-12320.859***	-12751.286***	-12565.889***	-12780.412***
	(626.918)	(610.874)	(625.494)	(610.132)	(771.537)	(619.890)	(645.662)	(614.987)
Log (Days from "Pending" to "Sold")	5118.994***	5229.373***	5138.132***	5239.319***	5035.742***	5219.117***	5088.805***	5229.582***
	(371.941)	(372.352)	(373.685)	(372.502)	(378.147)	(369.121)	(364.850)	(368.241)
Month Fixed Effects	No	Yes	No	Yes	No	Yes	No	Yes
Socio-Demographic Controls	No	Yes	No	Yes	No	Yes	No	Yes
Property Feature Controls	No	Yes	No	Yes	No	Yes	No	Yes
Observations	52,981	52,164	52,981	52,164	52,981	52,164	52,981	52,164
Panel B: Interaction with Frequency of Similar Transactions Nearby ($t-12$)								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Area	Block Group				City			
Measurement	per Capita		per Housing Unit		per Capita		per Housing Unit	
Zestimate_Sold	0.347***	0.241***	0.349***	0.240***	0.365***	0.246***	0.354***	0.243***
	(0.049)	(0.054)	(0.049)	(0.054)	(0.052)	(0.055)	(0.048)	(0.054)
# Transactions	-796409.968***	-298056.195*	-328162.582***	-107087.095	-745732.650**	-252819.535***	-221249.713***	-88025.446***
	(182748.518)	(157134.731)	(94141.464)	(73426.032)	(365159.708)	(80605.890)	(62574.799)	(25756.795)
Zestimate_Sold	2.399***	1.629***	1.039***	0.629***	1.904**	0.724***	0.573***	0.262***
× # Transactions	(0.629)	(0.565)	(0.298)	(0.241)	(0.916)	(0.194)	(0.156)	(0.062)
List Price	0.584***	0.644***	0.580***	0.645***	0.566***	0.643***	0.577***	0.645***
	(0.045)	(0.047)	(0.045)	(0.047)	(0.050)	(0.048)	(0.045)	(0.048)
Assessed Value	-0.023***	-0.030***	-0.024***	-0.030***	-0.028***	-0.033***	-0.027***	-0.033***
	(0.003)	(0.004)	(0.003)	(0.004)	(0.005)	(0.004)	(0.004)	(0.004)
Log (Days on Market)	-12847.774***	-12825.006***	-12837.488***	-12835.724***	-12386.913***	-12717.164***	-12577.214***	-12748.203***
	(622.250)	(608.236)	(622.280)	(608.759)	(726.864)	(617.871)	(638.630)	(613.655)
Log (Days from "Pending" to "Sold")	5125.016***	5218.766***	5144.814***	5229.976***	5060.833***	5205.761***	5100.364***	5217.095***
	(369.134)	(370.178)	(371.661)	(371.068)	(371.365)	(368.280)	(363.435)	(367.640)
Month Fixed Effects	No	Yes	No	Yes	No	Yes	No	Yes
Socio-Demographic Controls	No	Yes	No	Yes	No	Yes	No	Yes
Property Feature Controls	No	Yes	No	Yes	No	Yes	No	Yes
Observations	52,981	52,164	52,981	52,164	52,981	52,164	52,981	52,164

Robust standard errors reported in parentheses are clustered at city×month level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The dependent variable is sale price. The time between listing and pending sale and the time between pending to closing are log-transformed to control their nonlinear effects. The number of transactions (adjusted for the neighborhood size) is mean centered to allow easy interpretation of the main effects.

Electronic copy available at: https://ssrn.com/abstract=3584896

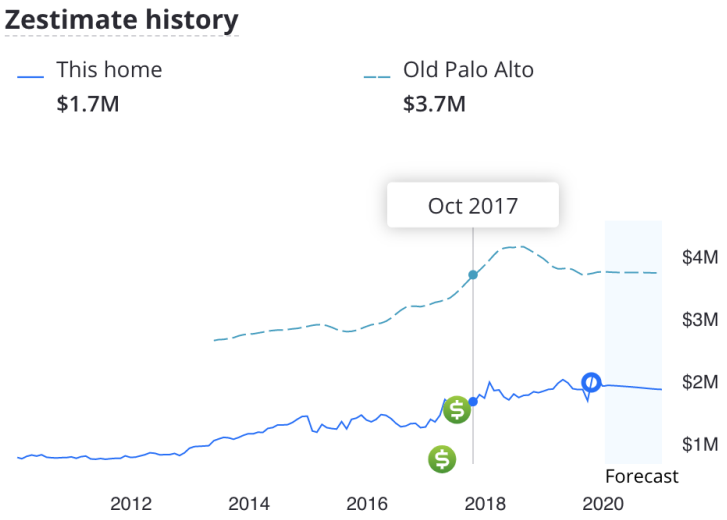
estimated values for the focal property.

In Table 4, for each property. its Zestimate is interacted with the transaction frequency of comparable properties sold within 6 months prior to the sale (in Panel A, or within 12 months prior to the sale in Panel B). I divide the properties into 5 segments based on their final sale prices: the properties sold at a price lower than 135k, the properties sold at a price between 135k and 196k, the properties sold at a price between 265k and 375k, the properties sold at a price higher than 375k. I only count the number of sales in the same segment because users usually only consider comparable homes when they are evaluating properties. To control the size of the neighborhood and thus that of the buyer consideration set, I standardize the number of properties sold in the same segments and divide it by either the population of the area (Columns (1), (2), (5), and (6) in each panel) or the total number of housing units in the area (Columns (3), (4), (7), and (8) in each panel). The summary statistics for the moderator used in Column (1) of Panel A are reported in Table 1. The Zestimate is interacted with the housing market activeness at both block group (Columns (1)-(4) in each panel) and city (Columns (5)-(8) in each panel) levels. As we can see, the number of recent sales has a negative effect on the final sale price, which is consistent with the hypothesis that a market with higher supply is less likely to be a seller market where buyers usually overbid for the properties. However, the effect of the Zestimate on the sale price increases with the number of recent sales in the neighborhood: based on the results reported in Column (2), the effect size will become about 10 times larger if there is one more recent transaction per Capita. It suggests that users rely more on the statistics that everyone can easily observe rather than focusing on analyzing the comparable sales and getting the private signals when the private signals they can from the comparable sales are noisier.

Robustness checks reported in Appendix (Table A7) replicate the results but exclude those sales with a logged sale-to-list ratio that is too high or too low, which may not be used as a reliable reference for the home's market value. The results are consistent with the conclusion drawn from Table 4.

Electronic copy available at: https://ssrn.com/abstract=3584896

Figure 5: Zestimate Details



5.2 Alternative Explanation

Another potential explanation for the increasing usage of Zestimates when there are more comparable sales is that real estate market participants may use the richness of data as a proxy for the summary statistic's accuracy. Here I provide evidence to eliminate this possibility by interacting the Zestimate with the average accuracy of Zestimates for the comparable properties sold in the surrounding area recently (sold within 6 months prior to the sale for the results reported in Panel A of Table 5 and sold within 12 months prior to the sale for the results reported in Panel B of the same table). The idea here is that if users attempt to infer the reliability of the Zestimate of their focal interest from the recent sales, they will be very likely to also look at more direct evidence, for example, the realized difference between the Zestimate and the final sale price for recent sales.[12]

Similar to Table 4, I interact the Zestimate with the average difference for properties sold in the same price segment at both block group (Columns (1)-(4) in each panel) and city (Columns (5)-(8) in each panel) levels. I also measure the average difference in terms of both the percent deviation of the final sale price from the Zestimate reported during the month

[12] As shown in Figure 5, this difference has been visualized on Zillow.com.

Electronic copy available at: https://ssrn.com/abstract=3584896

Table 5: Mechanism: Interaction with Information Accuracy

Panel A: Interaction with Average Deviation for Similar Transactions Nearby ($t-6$)								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Area	Block Group				City			
Measurement	$Zestimate_t$		$Zestimate_{t-1}$		$Zestimate_t$		$Zestimate_{t-1}$	
Zestimate_Sold	0.345***	0.244***	0.344***	0.244***	0.340***	0.240***	0.338***	0.240***
	(0.050)	(0.058)	(0.050)	(0.059)	(0.048)	(0.053)	(0.048)	(0.053)
% Deviation	-2511.465	-3280.874	-2834.807	-3447.803	-17722.572	-9872.767	-17722.429	-8375.687
	(4186.535)	(2948.192)	(4052.144)	(2857.283)	(11046.479)	(6830.298)	(12888.196)	(7716.131)
Zestimate_Sold	0.002	0.006	0.003	0.007	0.039	0.023	0.040	0.020
× % Deviation	(0.009)	(0.006)	(0.009)	(0.006)	(0.025)	(0.016)	(0.030)	(0.018)
List Price	0.592***	0.645***	0.592***	0.645***	0.598***	0.651***	0.599***	0.651***
	(0.047)	(0.051)	(0.047)	(0.051)	(0.045)	(0.047)	(0.045)	(0.047)
Assessed Value	-0.028***	-0.034***	-0.028***	-0.034***	-0.023***	-0.032***	-0.024***	-0.032***
	(0.004)	(0.005)	(0.004)	(0.005)	(0.003)	(0.004)	(0.003)	(0.004)
Log (Days on Market)	-12991.715***	-12763.784***	-12953.288***	-12726.558***	-12716.353***	-12713.136***	-12696.927***	-12686.735***
	(646.046)	(660.328)	(645.423)	(660.545)	(617.036)	(603.497)	(615.141)	(602.462)
Log (Days from "Pending" to "Sold")	5381.650***	5461.936***	5369.044***	5449.588***	5142.910***	5246.551***	5140.017***	5234.109***
	(379.645)	(389.125)	(379.913)	(389.371)	(363.891)	(367.774)	(363.869)	(367.439)
Month Fixed Effects	No	Yes	No	Yes	No	Yes	No	Yes
Socio-Demographic Controls	No	Yes	No	Yes	No	Yes	No	Yes
Property Feature Controls	No	Yes	No	Yes	No	Yes	No	Yes
Observations	48,193	47,501	48,195	47,503	52,508	51,711	52,510	51,713
Panel B: Interaction with Average Deviation for Similar Transactions Nearby ($t-12$)								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Area	Block Group				City			
Measurement	$Zestimate_t$		$Zestimate_{t-1}$		$Zestimate_t$		$Zestimate_{t-1}$	
Zestimate_Sold	0.335***	0.237***	0.335***	0.237***	0.324***	0.224**	0.287	0.197
	(0.049)	(0.056)	(0.049)	(0.056)	(0.102)	(0.100)	(0.463)	(0.421)
% Deviation	-9327.655	-6291.070	-8940.396	-5986.132	-43440.374	-30206.358	-92556.569	-58276.898
	(6417.173)	(4345.996)	(5874.783)	(3970.178)	(166395.047)	(122284.624)	(704970.950)	(504676.677)
Zestimate_Sold	0.015	0.011	0.015	0.011	0.093	0.066	0.204	0.129
×% Deviation	(0.012)	(0.008)	(0.011)	(0.008)	(0.356)	(0.261)	(1.548)	(1.107)
List Price	0.600***	0.651***	0.600***	0.651***	0.613***	0.665***	0.650	0.689*
	(0.046)	(0.049)	(0.046)	(0.049)	(0.101)	(0.088)	(0.463)	(0.371)
] Assessed Value	-0.026***	-0.033***	-0.026***	-0.033***	-0.023***	-0.032***	-0.023***	-0.033*
	(0.003)	(0.004)	(0.003)	(0.004)	(0.004)	(0.005)	(0.008)	(0.018)
Log (Days on Market)	-13065.105***	-12890.977***	-13030.870***	-12854.729***	-12689.190***	-12772.371***	-12619.408***	-12849.709***
	(631.133)	(633.741)	(631.390)	(634.396)	(720.845)	(646.015)	(1205.348)	(1581.054)
Log (Days from "Pending" to "Sold")	5298.621***	5368.578***	5287.444***	5356.610***	5247.413***	5350.865***	5441.910**	5495.004**
	(368.005)	(377.875)	(368.708)	(378.272)	(632.851)	(648.783)	(2639.005)	(2515.343)
Month Fixed Effects	No	Yes	No	Yes	No	Yes	No	Yes
Socio-Demographic Controls	No	Yes	No	Yes	No	Yes	No	Yes
Property Feature Controls	No	Yes	No	Yes	No	Yes	No	Yes
Observations	49,305	48,587	49,307	48,589	52,585	51,782	52,587	51,784

Robust standard errors reported in parentheses are clustered at city×month level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The dependent variable is sale price. The time between listing and pending sale and the time between pending to closing are log-transformed to control their nonlinear effects. The accuracy of the Zestimate for a recent sale is calculated as $|Zestimate - Sale\_Price|/Zestimate$ and the average accuracy is mean centered to allow easy interpretation of the main effects.

Electronic copy available at: https://ssrn.com/abstract=3584896

of sale (Columns (1), (2), (5), and (6) in each panel) and that from the Zestimate reported one month before the sale (Columns (3), (4), (7), and (8) in each panel). The former one is a natural comparison because both numbers have the same y-value on Zillow's trend graph. The later one further entrenches the conclusion by considering the possibility that users are sophisticated so that they are able to realize that the potential changes in the Zestimate between the time they are looking at it and the closing. The summary statistics for the moderator used in Column (1) of Panel A are reported in Table 1. The results reported in Table 5 suggest that the average deviation level doesn't have any material impact on how much customers rely on the Zestimate to make their purchase decision. It implies that either users don't update their belief about the Zestimate's accuracy using recent sales or the inferred accuracy is not the main factor that determines their reliance on the Zestimate. Unlike how users construct their reference sets, it is possible that customers will infer the accuracy of the summary statistic from how it performed for properties in other price segments because they may still browse them just out of curiosity. So I include the properties from other segments and replicate Table 5. The results are reported in A8 and they lead to confirm the finding that it is unlikely that users increase their usage of Zestimates because of the alleviation of accuracy concerns.

6 Heterogeneity in Effect Size and How it Moderates Racial Biases in Housing Market

In this section, I further investigate how the influence of the Zestimate varies across neighborhoods and how it potentially affects the racial biases in the housing market. I first show how the effect changes with demographic variables in Section 6.1. Then I examine the racial biases in the housing market caused by the redlining policy in section 6.2. Finally, in Section 6.3 I provide evidence of a smaller bias in Zestimate and combine the results from the previous sections I discuss how it leads to a moderated racial bias in final sale price.

Electronic copy available at: https://ssrn.com/abstract=3584896

6.1 Influence of Demographics on Zestimate's Effect

In this subsection, I inspect what socioeconomic variables affect the market participants' reliance on the Zestimate. First, I focus on how the dependence on the summary statistic changes with education and income levels. The answer to this question is ambiguous. On one hand, previous studies (Newman and Staelin, 1972; Claxton et al., 1974; Schaninger and Sciglimpaglia, 1981) have shown that the depth and breadth of information search before a purchase are usually higher among buyers with higher education and higher income. So those privileged people are likely to search for more information and be less reliant on a single information source. Also, as shown in Table 4, they are also well-trained to process the raw data. On the other hand, due to the well-known digital divide, low-income populations may not have access to the Internet and thus Zillow.com.

Table 6: Heterogeneous Effect: Influence of Education and Income Levels

	(1)	(2)	(3)	(4)	(5)	(6)
Moderator	#Years of Education			Log(Median Income)		
	Sale Price	Sale Price	Sale Price	Sale Price	Sale Price	Sale Price
Zestimate_Sold	0.281***	0.239***	0.238***	0.304***	0.263***	0.236***
	(0.046)	(0.052)	(0.054)	(0.047)	(0.053)	(0.054)
Moderator	8116.035***	7649.521***	7833.780***	16851.218***	13647.872***	11603.691***
	(957.874)	(939.070)	(1277.983)	(2057.227)	(1922.090)	(3510.055)
Zestimate_Sold	-0.015***	-0.014***	-0.021***	-0.020**	-0.019**	-0.032***
× Moderator	(0.004)	(0.004)	(0.005)	(0.009)	(0.008)	(0.011)
List Price	0.654***	0.678***	0.682***	0.627***	0.649***	0.670***
	(0.040)	(0.044)	(0.046)	(0.042)	(0.045)	(0.045)
Assessed Value	-0.017***	-0.025***	-0.028***	-0.026***	-0.032***	-0.028***
	(0.004)	(0.005)	(0.005)	(0.004)	(0.005)	(0.004)
Log (Days on Market)	-13055.079***	-13084.741***	-13089.605***	-12891.582***	-12927.252***	-12991.159***
	(570.259)	(606.991)	(591.699)	(573.050)	(601.435)	(594.689)
Log (Days from "Pending" to "Sold")	5391.538***	5326.206***	5307.059***	5201.606***	5181.476***	5290.396***
	(356.011)	(358.672)	(362.335)	(354.627)	(357.244)	(365.609)
Month Fixed Effects	No	Yes	Yes	No	Yes	Yes
Socio-Demographic Controls	No	No	Yes	No	No	Yes
Property Feature Controls	No	Yes	Yes	No	Yes	Yes
Observations	52,981	52,757	52,164	52,515	52,292	52,164

Robust standard errors reported in parentheses are clustered at city×month level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The dependent variable is sale price. The time between listing and pending sale and the time between pending to closing are log-transformed to control their nonlinear effects. The average number of years of education (the moderator used in Columns (1)-(3)) and the logged median income (the moderator used in Columns (4)-(6)) in the census block are mean centered to allow easy interpretation of the main effects.

To study it, I interact the Zestimate with the average education level and the median income separately. I take a log transformation of the median income since its distribution is highly right-skewed and we use the average number of years of education received to measure the average education level. The results are reported in Table 6: The Zestimate is interacted with the average number of years of education for results reported in Columns

Electronic copy available at: https://ssrn.com/abstract=3584896

(1)-(3) while it is interacted with the logged median income for results reported in Columns (4)-(6). Control variables are added into the model incrementally. The results show that the areas with a better-educated population and higher median income are more likely to be a "seller market", in which the final sale prices are higher due to the excess demand. However, users participating in those markets are less likely to rely on the summary statistics provided on Zillow.com to make their purchase decisions. This finding is consistent with our previous conclusion that the Zestimate serves as the summary of one of the information sources and its effect will diminish when 1) people's ability to process the raw data is higher 2) other information sources are more accessible. I have shown that more educated people are better at processing detailed transaction information. Moreover, other information sources are more accessible to privileged people, for example, they can pay a premium and hire experienced buyer agents or they are more likely to have friends or colleagues who have participated in local markets before and know the local market very well.

Table 7: Heterogeneous Effect: Influence of Internet and Computer Access

	(1)	(2)	(3)	(4)	(5)	(6)
Moderator	%Internet Subscription			%Computer Ownership		
	Sale Price	Sale Price	Sale Price	Sale Price	Sale Price	Sale Price
Zestimate_Sold	0.324***	0.272***	0.237***	0.330***	0.276***	0.223***
	(0.048)	(0.053)	(0.055)	(0.047)	(0.051)	(0.054)
Moderator	38017.777***	29607.734***	18377.405*	55582.731***	46708.265***	28989.568*
	(7609.364)	(6984.501)	(9649.062)	(10542.538)	(9915.323)	(14801.575)
Zestimate_Sold	-0.035	-0.036	-0.079*	-0.114**	-0.117***	-0.116*
× Moderator	(0.037)	(0.033)	(0.044)	(0.049)	(0.045)	(0.063)
List Price	0.609***	0.639***	0.659***	0.608***	0.640***	0.673***
	(0.044)	(0.046)	(0.047)	(0.043)	(0.046)	(0.046)
Assessed Value	-0.024***	-0.032***	-0.031***	-0.024***	-0.032***	-0.031***
	(0.003)	(0.004)	(0.004)	(0.003)	(0.004)	(0.005)
Log (Days on Market)	-12857.756***	-12953.386***	-12908.936***	-12876.829***	-12972.860***	-12952.265***
	(607.520)	(632.734)	(607.453)	(604.060)	(628.033)	(605.820)
Log (Days from "Pending" to "Sold")	5143.732***	5129.542***	5252.362***	5139.963***	5133.594***	5209.047***
	(360.727)	(363.317)	(368.700)	(358.690)	(361.118)	(363.482)
Month Fixed Effects	No	Yes	Yes	No	Yes	Yes
Socio-Demographic Controls	No	No	Yes	No	No	Yes
Property Feature Controls	No	Yes	Yes	No	Yes	Yes
Observations	52,981	52,757	52,164	52,981	52,757	52,164

Robust standard errors reported in parentheses are clustered at city×month level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The dependent variable is sale price. The time between listing and pending sale and the time between pending to closing are log-transformed to control their nonlinear effects. The percentage of households that have an Internet subscription (the moderator used in Columns (1)-(3)) and the percentage of households that have one or more types of computing devices (the moderator used in Columns (4)-(6)) in the census block are mean centered to allow easy interpretation of the main effects.

To further investigate this difference, I replace the average education level and the median income with measures of information access and search cost. The two measures I use here are the percentage of households that have an Internet subscription and the percentage of

Electronic copy available at: https://ssrn.com/abstract=3584896

households that have one or more types of computing devices. The results are reported in Table 7: Zestimate is interacted with the percentage of households that have an Internet subscription for results reported in Columns (1)-(3) while it is interacted with the percentage of households that have one or more types of computing devices for results reported in Columns (4)-(6). Control variables are added into the model incrementally. I find that even though that people who don't have stable access to the Internet or computing devices are less likely to utilize the Zestimate, the Internet subscription and computing device ownership do slightly reduce buyers' reliance on Zestimate. It suggests that people are more likely to rely on accessible and simple information sources such as Zestimate and other summary statistics when it is hard for them to acquire and search for other information.

Table 8: Heterogeneous Effect: Influence of Racial Makeup

	(1)	(2)	(3)	(4)	(5)	(6)
Moderator	%White			%White Owners		
	Sale Price	Sale Price	Sale Price	Sale Price	Sale Price	Sale Price
Zestimate_Sold	0.323***	0.264***	0.229***	0.323***	0.264***	0.232***
	(0.048)	(0.052)	(0.054)	(0.048)	(0.053)	(0.054)
Moderator	7308.393*	4495.520	3259.813	9986.045**	6684.631*	-3589.206
	(4109.398)	(3740.966)	(4259.760)	(4436.761)	(3930.080)	(6772.646)
Zestimate_Sold	0.012	0.019	0.010	-0.001	0.009	-0.004
× Moderator	(0.022)	(0.019)	(0.020)	(0.025)	(0.022)	(0.026)
List Price	0.607***	0.640***	0.658***	0.608***	0.641***	0.658***
	(0.045)	(0.047)	(0.048)	(0.044)	(0.047)	(0.048)
Assessed Value	-0.022***	-0.031***	-0.031***	-0.022***	-0.031***	-0.031***
	(0.003)	(0.004)	(0.004)	(0.003)	(0.004)	(0.004)
Log (Days on Market)	-12983.688***	-13083.290***	-12928.026***	-12983.307***	-13083.137***	-12917.784***
	(611.579)	(636.319)	(611.637)	(611.800)	(637.583)	(611.125)
Log (Days from "Pending" to "Sold")	5243.095***	5222.780***	5263.553***	5253.606***	5231.249***	5256.706***
	(368.246)	(372.687)	(372.891)	(367.849)	(372.907)	(372.971)
Month Fixed Effects	No	Yes	Yes	No	Yes	Yes
Socio-Demographic Controls	No	No	Yes	No	No	Yes
Property Feature Controls	No	Yes	Yes	No	Yes	Yes
Observations	52,981	52,757	52,164	52,981	52,757	52,164

Robust standard errors reported in parentheses are clustered at city×month level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The dependent variable is sale price. The time between listing and pending sale and the time between pending to closing are log-transformed to control their nonlinear effects. The share of white residents (the moderator used in Columns (1)-(3)) and the share of white home owners (the moderator used in Columns (4)-(6)) in the census block are mean centered to allow easy interpretation of the main effects.

As I have already shown in Table 6, users' reliance on the summary statistic depends on some socio-demographic characteristics of the neighborhood. Here I investigate whether there is a difference between "whiter" neighborhoods and more diverse neighborhoods after controlling the socio-demographic status. In Table 8, I interact the Zestimate with two measures of racial makeup – the percentage of white residents (reported in Columns (1)-(3)) and the percentage of white homeowners (reported in Columns (4)-(6)). The share of white

Electronic copy available at: https://ssrn.com/abstract=3584896

residents is the most common measure of neighborhood racial diversity and I introduce the percentage of white homeowners here as well to better approximate the race of the average market participant in the neighborhood. Those two measures are close to each other: the average neighborhood is about 79 percent white in the data set while the share of white is 2 percent higher in terms of home-ownership. The heavy tails show that the residential segregation between whites and minorities. If minorities are inferior in terms of information acquisition and processing even after controlling the education and wealth inequality, it is likely that we will see heavier use of Zestimates in those less-white neighborhoods. This interaction effect exists in our data but it is not significant: In Table 8, the sale price is less likely to be influenced by the Zestimate when the percentage of white homeowners in the neighborhood increases, even though the interaction term is not significant.

To double-check this difference, I split the observations into two subgroups based on the share of white residents in the neighborhoods and run the 2SLS regressions separately for each group. The results for those properties that sold in a neighborhood where the white population share is greater than the median (0.852) is reported in Columns (1)-(3) of Panel A in Table 9 and the results for the properties that sold in a neighborhood that the white population share is less or equal to the median is reported in Columns (4)-(6) in the same panel.[13] Similar to the results reported with interactions, the coefficient of the Zestimate is more significant and greater in magnitude for less white neighborhoods. A similar subgroup analysis has been done based on the share of white homeowners in the neighborhoods (where the median is 0.885) and the results reported in Panel B in Table 9 show a similar pattern.[14] All of these findings show that people living in more diverse neighborhoods are relying more on the online summary statistic when making the important financial decision but this disparity driven by the racial makeup is much less significant than those differences among neighborhoods with various education or income level. Finally, I dig into the individual data obtained from the public land records where the buyers' race is

[13]The average white population share is 0.927 for the first group and 0.653 for the second group.
[14]The average share of white homeowners is 0.949 for the first group and 0.693 for the second group.

Electronic copy available at: https://ssrn.com/abstract=3584896

Table 9: Subgroup Analysis: Influence of Racial Makeup

Panel A: %White Residents in Census Block Group						
	(1)	(2)	(3)	(4)	(5)	(6)
	%White > Median			%White ≤ Median		
	Sale Price	Sale Price	Sale Price	Sale Price	Sale Price	Sale Price
Zestimate_Sold	0.300***	0.269***	0.222***	0.374***	0.276**	0.235**
	(0.049)	(0.061)	(0.067)	(0.105)	(0.111)	(0.110)
List Price	0.627***	0.640***	0.664***	0.568***	0.637***	0.657***
	(0.045)	(0.051)	(0.054)	(0.099)	(0.102)	(0.099)
Assessed Value	-0.024***	-0.028***	-0.028***	-0.021***	-0.034***	-0.036***
	(0.004)	(0.006)	(0.006)	(0.005)	(0.006)	(0.008)
Log (Days on Market)	-14667.518***	-14480.005***	-14506.841***	-11161.042***	-11535.069***	-11406.667***
	(813.330)	(942.634)	(927.317)	(1078.286)	(1040.761)	(969.328)
Log (Days from "Pending" to "Sold")	5996.536***	5819.319***	5948.616***	4390.895***	4484.136***	4563.159***
	(561.078)	(595.009)	(599.379)	(479.313)	(468.427)	(467.066)
Month Fixed Effects	No	Yes	Yes	No	Yes	Yes
Socio-Demographic Controls	No	No	Yes	No	No	Yes
Property Feature Controls	No	Yes	Yes	No	Yes	Yes
Observations	26,513	26,418	26,142	26,468	26,339	26,022
Panel B: %White Owners in Census Block Group						
	(1)	(2)	(3)	(4)	(5)	(6)
	%White Owners > Median			%White Owners ≤ Median		
	Sale Price	Sale Price	Sale Price	Sale Price	Sale Price	Sale Price
Zestimate_Sold	0.292***	0.262***	0.219***	0.409***	0.301**	0.255**
	(0.051)	(0.063)	(0.071)	(0.122)	(0.129)	(0.127)
List Price	0.636***	0.647***	0.669***	0.535***	0.611***	0.636***
	(0.047)	(0.053)	(0.057)	(0.115)	(0.118)	(0.114)
Assessed Value	-0.024***	-0.028***	-0.028***	-0.020***	-0.033***	-0.036***
	(0.004)	(0.005)	(0.005)	(0.005)	(0.008)	(0.009)
Log (Days on Market)	-15050.206***	-14853.645***	-14719.009***	-10570.085***	-11040.001***	-11104.400***
	(830.547)	(949.506)	(923.403)	(1194.037)	(1178.574)	(1100.818)
Log (Days from "Pending" to "Sold")	6218.804***	5988.047***	6066.440***	4070.516***	4264.311***	4396.335***
	(520.761)	(536.843)	(532.046)	(538.006)	(545.629)	(550.096)
Month Fixed Effects	No	Yes	Yes	No	Yes	Yes
Socio-Demographic Controls	No	No	Yes	No	No	Yes
Property Feature Controls	No	Yes	Yes	No	Yes	Yes
Observations	26,498	26,419	26,138	26,483	26,338	26,026
Panel C: Race of Individual Buyer						
	(1)	(2)	(3)	(4)	(5)	(6)
	White Buyers			Non-white Buyers		
	Sale Price	Sale Price	Sale Price	Sale Price	Sale Price	Sale Price
Zestimate_Sold	0.354***	0.349***	0.299***	0.387*	0.419*	0.302
	(0.067)	(0.069)	(0.081)	(0.226)	(0.214)	(0.191)
List Price	0.590***	0.591***	0.623***	0.571***	0.541***	0.631***
	(0.062)	(0.061)	(0.070)	(0.212)	(0.199)	(0.175)
Assessed Value	-0.018***	-0.035***	-0.035***	-0.013	-0.025	-0.023
	(0.006)	(0.008)	(0.010)	(0.017)	(0.022)	(0.024)
Log (Days on Market)	-12343.447***	-11716.425***	-11842.400***	-10366.669***	-10322.607***	-12527.589***
	(933.093)	(920.590)	(975.997)	(3082.661)	(2871.303)	(2211.687)
Log (Days from "Pending" to "Sold")	5202.912***	5096.307***	5008.896***	2394.762*	2365.489*	3006.329***
	(548.518)	(498.374)	(505.707)	(1382.115)	(1232.515)	(1040.552)
Month Fixed Effects	No	Yes	Yes	No	Yes	Yes
Socio-Demographic Controls	No	No	Yes	No	No	Yes
Property Feature Controls	No	Yes	Yes	No	Yes	Yes
Observations	11,161	11,135	10,974	2,339	2,333	2,281

Robust standard errors reported in parentheses are clustered at city×month level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The dependent variable is sale price. The time between listing and pending sale and the time between pending to closing are log-transformed to control their nonlinear effects. In Panel A, the regressions reported in Columns (1)-(3) use the purchases made in block groups where the share of white residents is greater than the median (0.858) for estimation while the regressions reported in Columns (4)-(6) use the purchases made in block groups where the share of white residents is less than the median. In Panel B, the regressions reported in Columns (1)-(3) use the purchases made in block groups where the share of white home owners is greater than the median (0.889) for estimation while the regressions reported in Columns (4)-(6) use the purchases made in block groups where the share of white home owners is less than the median. The regressions reported in Columns (1)-(3) use the purchases made by whites for estimation while the regressions reported in Columns (4)-(6) use the purchases made by non-whites. Buyers' races are identified from the names shown on public records.

identified from the public land records and check the racial difference on the individual level. I run regressions similar to Panel A of Table 9 but this time separately for white buyers

Electronic copy available at: https://ssrn.com/abstract=3584896

and nonwhite buyers. The 2SLS results are reported in Panel C of Table 9. Consistent with previous results, nonwhites are more responsive to the changes in the Zestimate even though the difference is not significant.

6.2 Lingering Impact of Federal "Redlining": Home Value Gap between Whites and Nonwhites

I then investigate the lingering effects of "redlining", in other words, whether the minority neighborhoods are suffered from an unfair housing market. "Redlining" refers to the practice that the Federal Home Loan Bank Board (FHLBB)'s practices to create a map to indicate the level of security for real-estate investments in each neighborhood. The neighborhoods that were considered the riskiest for mortgages were outlined in red on this map and these neighborhoods are often minority neighborhoods. The existence of the map led to mortgage loan denials (Jackson, 1987) and a persistent decline in home values (Rutan and Glass, 2018) in these minority communities.

To quantify this effect, I run a set of OLS regressions where the independent variable of interest is the share of whites in the neighborhood. Two different measures are used here for the home value: the list price and the final sale price. Those variables are used to inspect the level of Redlining's lingering impacts on different aspects of the housing market. The basic model for the prices is:

$$P_{itk} = \alpha^D + \beta^D Share\_Whites_k + \mu_t^D + \eta^D L_k + \zeta S_i^D + \epsilon_{ikt}^D,$$

where all the variables are the same as the ones described on section 4.1 except the independent variable of interest is now $Share\_Whites_k$ and β^D shows the level of Redlining's lingering impact on the housing market.

The results are reported in Column (1) of Table 10 for the list price and in Column (3) of the same table for the sale price. As we can see, both measures increase when the neighborhood is whiter: The average list price will increase by 964.8 dollars and the average sale price will increase by 900.1 dollars when the neighborhood is one percent whiter. The

Electronic copy available at: https://ssrn.com/abstract=3584896

Table 10: Lingering Impact of Federal "Redlining" on Housing Market

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable	List Price		Sale Price			Δ
% White	100933.142***	101434.944***	94258.580***	95896.516***	101149.820***	-8374.646***
	(3151.928)	(4133.697)	(3029.549)	(3908.028)	(4687.799)	(1089.829)
Assessed Value		0.577***		0.454***	0.498***	-0.113***
		(0.034)		(0.026)	(0.031)	(0.012)
Log (Days on Market)					-8377.772***	-16609.433***
					(1105.969)	(546.676)
Log (Days from "Pending" to "Sold")					91.014	7160.211***
					(760.541)	(399.763)
Month Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Socio-Demographic Controls	Yes	Yes	Yes	Yes	Yes	Yes
Property Feature Controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	116,587	86396	116,587	87,254	52,585	52,585
R-squared	0.562	0.637	0.594	0.652	0.662	0.215
Adjusted R-squared	0.562	0.637	0.593	0.651	0.661	0.213

Robust standard errors reported in parentheses are clustered at city×month level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The dependent variable is list price in Columns (1) and (2), sale price in Columns (3)-(5) and their difference (Δ = sale price) - list price in Column (6).

difference between the two gaps is significant. To test the robustness of the results, I add the assessed value into the model and rerun the regressions again. The results reported in Columns (2) and (4) show that the estimated racial biases are slightly greater and still very significant. A similar result for the sale price can be drawn from the robustness check where the number of days and the number of days from the pending sale to closing are also added to control the unobserved house/seller quality. Those findings are consistent with the findings in the previous studies (Aaronson et al., 2017; Perry et al., 2018). Another thing that is noticeable here is that the white-premium is higher for the list prices and it might have been moderated by some factors during the sale process, which leads to a lower premium in the final sale prices. To confirm this finding, I regress the logged sale-to-list ratios on the share of whites. The results reported in Column (6) suggest that the sale-to-list ratio is indeed significantly lower in those predominately white neighborhoods.

6.3 Racial Biases in Zestimate and its Moderation Effect on "Redlining"

One burning issue regarding algorithms is whether the outcomes they produced are biased (Cowgill and Tucker, 2020). So here I focus on investigating how biased the Zestimate is compared to the existing racial biases in the housing market – whether the Zestimate reflects this gap between white and nonwhites or the gap will attenuate through how the Zestimate is calculated. To study this question, I run regressions similar to those in Table 10 but use the log-transformed Zestimate-to-List ratio as the dependent variable.

Electronic copy available at: https://ssrn.com/abstract=3584896

Table 11: Racial Biases in Zestimate

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable	Log(Zestimate/List Price)			Log(Zestimate/Sale Price)		
% White	-0.034***	-0.033***	-0.020***	-0.052***	-0.052***	-0.015*
	(0.005)	(0.006)	(0.007)	(0.006)	(0.008)	(0.008)
Assessed Value		0.000**	0.000***		0.000***	0.000***
		(0.000)	(0.000)		(0.000)	(0.000)
Log (Days on Market)			-0.037***			0.025***
			(0.002)			(0.002)
Log (Days from "Pending" to "Sold")			0.013***			-0.008***
			(0.001)			(0.002)
Month Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Socio-Demographic Controls	Yes	Yes	Yes	Yes	Yes	Yes
Property Feature Controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	116,423	86,306	52,869	116,423	87,164	52,533
R-squared	0.051	0.067	0.074	0.077	0.094	0.087
Adjusted R-squared	0.049	0.065	0.071	0.076	0.093	0.085

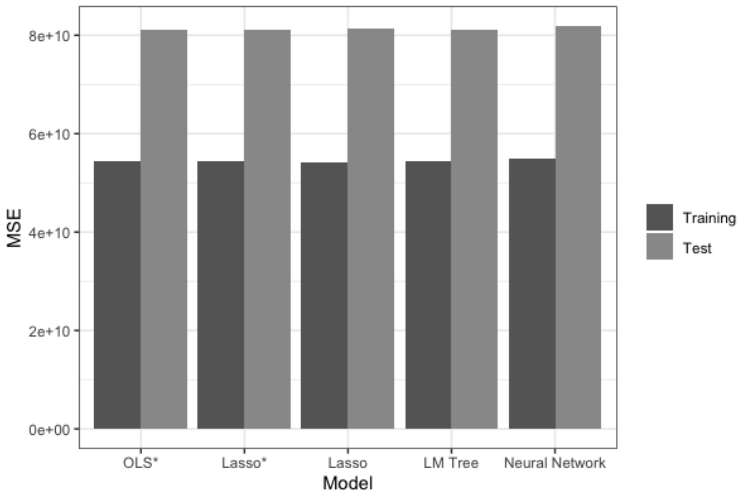
Robust standard errors reported in parentheses are clustered at city×month level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The dependent variable is logged Zesimate-to-List ratio for Column (1) -(3) and logged Zesimate-to-Sale ratio for Columns (4)-(6).

The OLS results are reported in Table 11. I compare the Zestimate with both the list price (Columns (1)-(3)) and the sale price (Columns (4)-(6)). The assessed home value, the time on the market, and the time between the pending sale and the final deal are controlled in various models. As we can see, the Zestimate home valuation predicted by algorithms doesn't fully reflect the white-premium in home owner's valuation. Based on the results in Column (3), on average, the logged Zestimate-to-list ratio is 0.02 lower for properties located in nonwhite neighborhoods (where $\%white = 0$) compared to similar properties sold in white neighborhoods (where $\%white = 1$). It indicates that even though the Zestimate is heavily influenced by the list price after the information becomes available [15], it doesn't fully incorporate the racial biases existing in the listing price. Similarly, the logged Zestimate-to-sale ratio is 0.01 lower for properties located in nonwhite neighborhoods (where $\%white = 0$) compared to similar properties sold in white neighborhoods (where $\%white = 1$). The difference is not as significant as the one observed for the logged Zestimate-to-list ratio. This finding is consistent with the previous conclusion drawn from Table 10 that the sale-to-list ratio is lower in those predominately white neighborhoods. It might be due to the attenuated racial biases in the Zestimate does affect the level of biases in the sale price, since buyers do use the Zestimate as a reference when making the final price decision.

[15] See `https://www.inman.com/2018/08/08/whats-the-deal-with-zillow-changing-its-`

Electronic copy available at: https://ssrn.com/abstract=3584896

Figure 6: Reverse Engineering: Model Comparison



Notes: In-sample and out-of-sample mean squared errors are reported here. 20% of the observations in each city are randomly selected into the holdout dataset for model evaluation.

To further investigate the racial biases in Zestimates, I made a few attempts at reverse-engineering the Zestimate by predicting the sale price using some of most popular prediction models. Most of the models are constructed with the property features used by Zillow.com and the city indicators.[16] Figure 6 presents the in-sample and out-of-sample goodness-of-fit for different models. First I fit a simple OLS linear regression model using the property features as predictors and report the in-sample and out-of-sample mean squared errors. Then variable selection and regularization are performed in a LASSO regression model with the same set of predictors. The LASSO model is further improved by adding the city indicators. Finally, I allow the heterogeneity in model parameters and fit a tree-based regression model with recursive partitioning. The data is stratified according to the city indicators and then separate regression models are fit to each stratum. As we can see from the figure, the goodness-of-fit measure doesn't improve significantly when the model becomes more complex. In addition, I train a neural network model which allows complex nonlinearities with the full set of predictors (property features and city indicators). It again doesn't show a better performance than the simple regression model. So the comparison here suggests that a

zestimates/

[16]See the Kaggle competition: `https://www.kaggle.com/c/zillow-prize-1`.

Electronic copy available at: https://ssrn.com/abstract=3584896

simple regression model might be good enough to at least shed a light on how Zillow's prediction model works.

Table 12: Reverse Engineering: Racial Differences in Residuals

	(1)	(2)	(3)	(4)	(5)
Model	OLS	Lasso	Lasso	LM Tree	Neural Network
%White	9660.034**	9660.034**	10498.351***	10511.615***	9683.719**
	(3821.234)	(3821.234)	(3830.025)	(3820.765)	(3817.075)
Observations	53,176	53,176	53,176	53,176	53,176
R^2	0.023	0.024	0.031	0.025	0.034
Adjusted R^2	0.021	0.022	0.029	0.022	0.031

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The dependent variable is residuals.

I then regress the residuals ($P_{Sale} - \hat{P}_{Sale}$) from those models on the set of predictors including the list price, the assessed value, the time on the market, the time from pending to close, the month fixed effects, the property features, and the neighborhood socio-demographic controls including the white share of the population. The coefficients of the white share are reported in Table 12. Based on the results reported here, the gap between the predicted outcome based on the OLS model and the final sale price is $\$9,660$ smaller when the white share increases by 1. This difference significantly adjusts the existing racial biases in the housing market by systematically underestimating the price markup on homes located in whiter neighborhoods.

Therefore thanks to the algorithmic prediction model it is calculated from, Zestimate is less in favor of those white-dominated neighborhoods than the actual list price. Combined with the fact that people who live in the diverse neighborhoods do follow the Zestimate, at least at the same level as their peers living in the white-dominated neighborhood, it is plausible that the Zestimate is one of the factors that moderate the lingering effect of "Redlining" in the home buying process.

7 Conclusion

In the previous analyses, I investigate how predictive algorithms change housing markets. Using data collected from Zillow.com and public records, I show that the Zestimate influences market participants' decisions as a public source of market information. I also show that this estimate doesn't fully reflect the racial biases in the housing market and thus it might

Electronic copy available at: https://ssrn.com/abstract=3584896

have mitigated the home value gap between whites and nonwhites.

These results matter because sometimes policymakers and researchers might fear that the predictive algorithms may augment the inequality by reinforcing the privileged people's advantages. However, the preliminary results here show that at least in our setting, the algorithm-powered home value estimates can actually mitigate the existing racial biases in the housing market by providing more neutral information. It can be generalized to other types of estimates which aim to provide summarized information to customers. Those statistics that summarize the unprejudiced market information may help customers understand the market better and make more objective decisions without putting more effects.

There are some limitations to the study. First, I do not know the browsing history of buyers and whether they did check the Zestimate, as well as the negotiation between buyers and sellers. Second, I do not have individual customer data on race and ethnicity for all the properties and instead focus on the neighborhood properties. Last, I have not considered the profitability of the information provider and whether they have incentives to provide less biased information.

Electronic copy available at: https://ssrn.com/abstract=3584896

References

Aaronson, D., D. A. Hartley, and B. Mazumder (2017). The effects of the 1930s HOLC "redlining" maps. FRB of Chicago Working Paper No. WP-2017-12.

Ali, M., P. Sapiezynski, M. Bogen, A. Korolova, A. Mislove, and A. Rieke (2019). Discrimination through optimization: How Facebook's ad delivery can lead to skewed outcomes. arXiv preprint arXiv:1904.02095.

Angwin, J., J. Larson, S. Mattu, and L. Kirchner (2016). Machine bias. *ProPublica 23*, 2016.

Bayer, P., M. Casey, F. Ferreira, and R. McMillan (2017). Racial and ethnic price differentials in the housing market. *Journal of Urban Economics 102*, 91–105.

Berry, B. J. (1976). Ghetto expansion and single-family housing prices: Chicago, 1968–1972. *Journal of Urban Economics 3*(4), 397–423.

Black, S. E. (1999). Do better schools matter? parental valuation of elementary education. *Quarterly Journal of Economics 114*(2), 577–599.

Brown, J. R. and A. Goolsbee (2002). Does the internet make markets more competitive? evidence from the life insurance industry. *Journal of Political Economy 110*(3), 481–507.

Bughin, J., J. Seong, J. Manyika, M. Chui, and R. Joshi (2018). Notes from the AI frontier: Modeling the impact of AI on the world economy. McKinsey Global Institute.

Bulow, J. I., J. D. Geanakoplos, and P. D. Klemperer (1985). Multimarket oligopoly: Strategic substitutes and complements. *Journal of Political Economy 93*(3), 488–511.

Chambers, D. N. (1992). The racial housing price differential and racially transitional neighborhoods. *Journal of Urban Economics 32*(2), 214–232.

Claxton, J. D., J. N. Fry, and B. Portis (1974). A taxonomy of prepurchase information gathering patterns. *Journal of Consumer Research 1*(3), 35–42.

Cowgill, B. and C. E. Tucker (2020). Algorithmic fairness and economics. *The Journal of Economic Perspectives*.

Cui, R., J. Li, and D. J. Zhang (2020). Reducing discrimination with reviews in the sharing economy: Evidence from field experiments on Airbnb. *Management Science 66*(3), 1071–1094.

Dietvorst, B. J., J. P. Simmons, and C. Massey (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General 144*(1), 114.

Edelman, B., M. Luca, and D. Svirsky (2017). Racial discrimination in the sharing economy: Evidence from a field experiment. *American Economic Journal: Applied Economics 9*(2), 1–22.

Electronic copy available at: https://ssrn.com/abstract=3584896

Ellison, G. and S. F. Ellison (2009). Search, obfuscation, and price elasticities on the Internet. *Econometrica 77*(2), 427–452.

Hauser, J. R., G. L. Urban, and B. D. Weinberg (1993). How consumers allocate their time when searching for information. *Journal of Marketing Research 30*(4), 452–466.

Hellwig, C. and L. Veldkamp (2009). Knowing what others know: Coordination motives in information acquisition. *The Review of Economic Studies 76*(1), 223–251.

Hendel, I., A. Nevo, and F. Ortalo-Magné (2009). The relative performance of real estate marketing platforms: MLS versus FSBOMadison. com. *American Economic Review 99*(5), 1878–98.

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

Jackson, K. T. (1987). *Crabgrass frontier: The suburbanization of the United States*. Oxford University Press.

Jensen, R. (2007). The digital provide: Information (technology), market performance, and welfare in the south indian fisheries sector. *Quarterly Journal of Economics 122*(3), 879–924.

Kanamori, T. and H. Shimodaira (2009). Geometry of covariate shift with applications to active learning. *Dataset Shift in Machine Learning*, 87–105.

Kiel, G. C. and R. A. Layton (1981). Dimensions of consumer information seeking behavior. *Journal of Marketing Research 18*(2), 233–239.

Kiel, K. A. and J. E. Zabel (1996). House price differentials in US cities: Household and neighborhood racial effects. *Journal of housing economics 5*(2), 143–165.

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

Kuruzovich, J., S. Viswanathan, and R. Agarwal (2010). Seller search and market outcomes in online auctions. *Management Science 56*(10), 1702–1717.

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

Logg, J. M., J. A. Minson, and D. A. Moore (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes 151*, 90–103.

Myers, C. K. (2004). Discrimination and neighborhood effects: Understanding racial differentials in US housing prices. *Journal of Urban economics 56*(2), 279–302.

Electronic copy available at: https://ssrn.com/abstract=3584896

Newman, J. W. and R. Staelin (1972). Prepurchase information seeking for new cars and major household appliances. *Journal of Marketing Research 9*(3), 249–257.

Obermeyer, Z., B. Powers, C. Vogeli, and S. Mullainathan (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science 366*(6464), 447–453.

Payne, J., J. Bettman, and E. Johnson (1991). Consumer decision making. *Handbook of Consumer Behaviour*, 50–84.

Perry, A., J. Rothwell, and D. Harshbarger (2018). The devaluation of assets in black neighborhoods: The case of residential property. *Metropolitan Policy Program at Brookings*.

Ratchford, B. T., M.-S. Lee, and D. Talukdar (2003). The impact of the internet on information search for automobiles. *Journal of Marketing Research 40*(2), 193–209.

Rutan, D. Q. and M. R. Glass (2018). The lingering effects of neighborhood appraisal: Evaluating redlining's legacy in Pittsburgh. *Professional Geographer 70*(3), 339–349.

Schaninger, C. M. and D. Sciglimpaglia (1981). The influence of cognitive personality traits and demographics on consumer information acquisition. *Journal of Consumer Research 8*(2), 208–216.

Seagraves, P. and P. Gallimore (2013). The gender gap in real estate sales: negotiation skill or agent selection? *Real Estate Economics 41*(3), 600–631.

Tucker, C., J. Zhang, and T. Zhu (2013). Days on market and home sales. *RAND Journal of Economics 44*(2), 337–360.

Westbrook, R. A. and C. Fornell (1979). Patterns of information source usage among durable goods buyers. *Journal of Marketing Research 16*(3), 303–312.

Zettelmeyer, F., F. S. Morton, and J. Silva-Risso (2006). How the internet lowers prices: Evidence from matched survey and automobile transaction data. *Journal of Marketing Research 43*(2), 168–181.

Electronic copy available at: https://ssrn.com/abstract=3584896

A Appendix

A.1 Pricing Game

This section consists of three parts. In Section A.1.1, I examine a strategic pricing game between two sellers and one buyer, and demonstrates the complementarity of the sellers' prices. In Section A.1.2, I incorporate sellers' information choices into the strategic pricing game, and use existing results in information economics to establish sellers' coordination motives in acquiring information about market demand. In Section A.1.3, I map the model into the housing markets and explain my empirical findings, that people rely more on the Zestimates when there are more private signals.

A.1.1 Strategic Pricing

Here I consider a two-stage game between two sellers and one buyer.[17] Each seller has a good for sale. Each seller's valuation for his good is normalized to 0. In stage 1, the two sellers simultaneously choose prices p_1 and p_2. In stage 2, the buyer observes p_1 and p_2, as well as his valuation for the two goods η_1 and η_2, drawn from some independent uniform distributions,[18] and then chooses which seller to buy from.

To highlight the key factors in the model, the buyer's choice is assumed to be binary: either she buys from seller 1 or she buys from seller 2. As a result, the buyer buys from seller 1 if

$$\eta_1 - p_1 \geq \eta_2 - p_2,$$

and vice versa. Seller i's utility is p_i if the buyer buys his good, and is 0 if the buyer buys from the other seller. Therefore, seller 1's utility can be written as the following function of p_1 and p_2:

$$U_1(p_1, p_2) \equiv p_1 \Pr_{\eta_1, \eta_2}(\eta_1 - p_1 \geq \eta_2 - p_2).$$

[17] It can be extended to model with two buyers and one seller and other market settings.

[18] The random variables η_1 and η_2 are interpreted as the difference in buyer's taste between the two products, which captures the differentiation across products. When the buyer's valuation is perfectly revealed to sellers, the two sellers will engage in a Bertrand competition.

Electronic copy available at: https://ssrn.com/abstract=3584896

and similarly,

$$U_2(p_1, p_2) \equiv p_2 \Pr_{\eta_1,\eta_2}(\eta_2 - p_2 \geq \eta_1 - p_1).$$

Following Bulow et al. (1985), I establish the strategic complementarity in sellers' price setting decisions by showing that $U_1(p_1, p_2)$ and $U_2(p_1, p_2)$ are both supermodular functions:

Proposition 1. *$U_1(p_1, p_2)$ and $U_2(p_1, p_2)$ are supermodular.*

Proof. Since U_1 and U_2 are symmetric, it is sufficient to show that U_1 is supermodular. For every $p_1^* > p_1'$ and $p_2^* > p_2'$, I show that

$$U_1(p_1^*, p_2^*) + U_1(p_1', p_2') - U_1(p_1^*, p_2') - U_1(p_1', p_2^*) > 0. \tag{1}$$

This is equivalent to:

$$p_1^* \int_{\eta_1} G_2(\eta_1 - p_1^* + p_2^*) - G_2(\eta_1 - p_1^* + p_2') dG_1(\eta_1) - p_1' \int_{\eta_1} G_2(\eta_1 - p_1' + p_2^*) - G_2(\eta_1 - p_1' + p_2') dG_1(\eta_1) \tag{2}$$

Given that $(\eta_1 - p_1^* + p_2^*) - (\eta_1 - p_1^* + p_2') = (\eta_1 - p_1' + p_2^*) - (\eta_1 - p_1' + p_2')$, and η_2 follows a uniform distribution,

$$G_2(\eta_1 - p_1^* + p_2^*) - G_2(\eta_1 - p_1^* + p_2') = G_2(\eta_1 - p_1' + p_2^*) - G_2(\eta_1 - p_1' + p_2') > 0.$$

Since $p_1^* > p_1^{'}$, we have (2) being strictly positive, which in turn implies that (1) is strictly positive. □

This result implies that the marginal benefit for seller 1 to increase his price increases with the price set by seller 2, in another word, sellers have incentives to coordinate when setting their prices. Such complementarities are well-known in monopolist competition models with sticky prices.

A.1.2 Strategic Pricing with Endogenous Information Acquisition

I expand the strategic price setting game in Section A.1.1 by introducing an information acquisition stage before sellers choosing their prices. Suppose there are n informative signals

Electronic copy available at: https://ssrn.com/abstract=3584896

$(s_1, s_2, ..., s_n)$ about the buyers' valuations $\overrightarrow{\eta} \equiv (\eta_1, \eta_2)$. I assume $s_1, s_2, ..., s_n$ are random variables that follow the same distribution and are independent conditional on $\overrightarrow{\eta}$.

The game proceeds in three stages. In stage 1, the two sellers simultaneously choose a subset of signals to observe, with $S_i \subset \{s_1, s_2, ..., s_n\}$ the set of signals observed by seller i. I assume that each seller faces a capacity constraint when acquiring information, in the sense that there exists $m \in \{1, 2, ..., n-1\}$ such that $|S_i| \leq m$ for every $i \in \{1, 2\}$. In stage 2, the two sellers simultaneously choose prices p_1 and p_2, after observing the realizations of the signals they choose to observe. Importantly, each seller cannot observe the other seller's informational choice, i.e., seller i cannot observe S_j. In stage 3, the buyer observes p_1, p_2, η_1, and η_2, and chooses between buying the item sold by seller 1 or from seller 2.

According to a well-known result in Hellwig and Veldkamp (2009) that establishes the strategic complementarity in sellers' informational choices, if players' actions in the price-setting stage are strategic complements, then players' informational choices are also strategic complements. In my setting, their result implies that when seller 2 observes signal s_i, the increase in seller 1's expected payoff by observing s_i strictly increases. It also suggests that seller i's pricing decision in the second stage becomes more responsive to signal s_i relative to the other signals he observe.

A.1.3 Application to Housing Market

Consider a market with two sellers and one buyer. There are $n+1$ signals $\{s, s_1, s_2, ..., s_n\}$ available for the sellers to acquire, which are informative about $\overrightarrow{\eta}$ and are conditionally independent. s is the Zestimate (the public signal), and $s_1, s_2, ..., s_n$ are the prices (and conditions) of comparable sales (private signals). I assume that $s_1, s_2, ..., s_n$ are drawn from identical distributions.

Based on a modified three-stage game studied in Section A.1.2. In stage 1, each seller observes the public signal s for free, and choose to observe a subset of private signals $\{s_1, s_2, ..., s_n\}$, while facing the constraint that the number of signals she can observe is no more than m. Stage 2 and stage 3 of the game remains the same as in Section A.1.2. Consider a symmetric equilibrium of the game in which each signal in the set $\{s_1, s_2, ..., s_n\}$

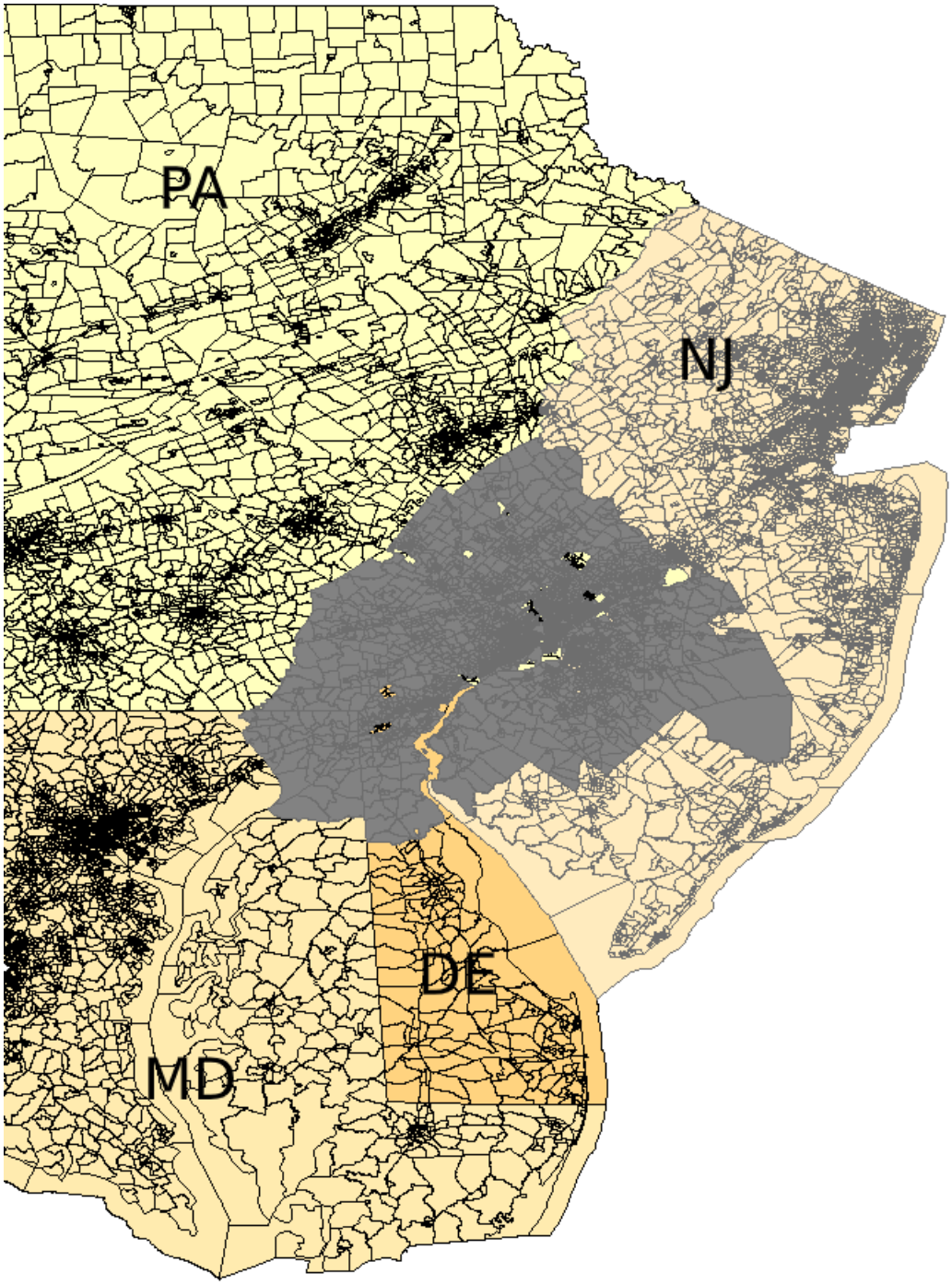
Electronic copy available at: https://ssrn.com/abstract=3584896

is observed by each seller with equal (ex ante) probability. Fixing each seller's capacity to process information m while increasing the number of available signals n, we know that for every $s_i \in \{s_1, s_2, ..., s_n\}$, the probability with which seller 1 observing s_i conditional on seller 2 observing s_i decreases. The conclusion in Section A.1.2 then implies that as n increases, the seller's pricing decisions rely more on the public signal s compared to the other signals he can observe within the set $\{s_1, s_2, ..., s_n\}$.

Electronic copy available at: https://ssrn.com/abstract=3584896

A.2 Figures

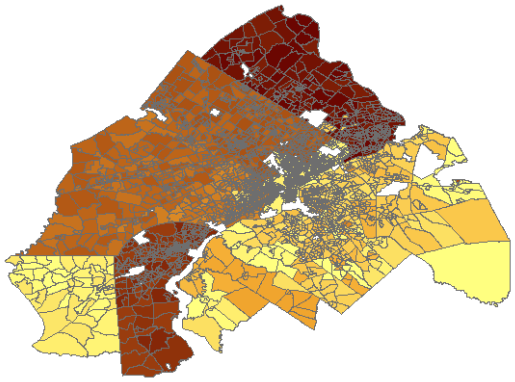
Figure A1: Data Coverage



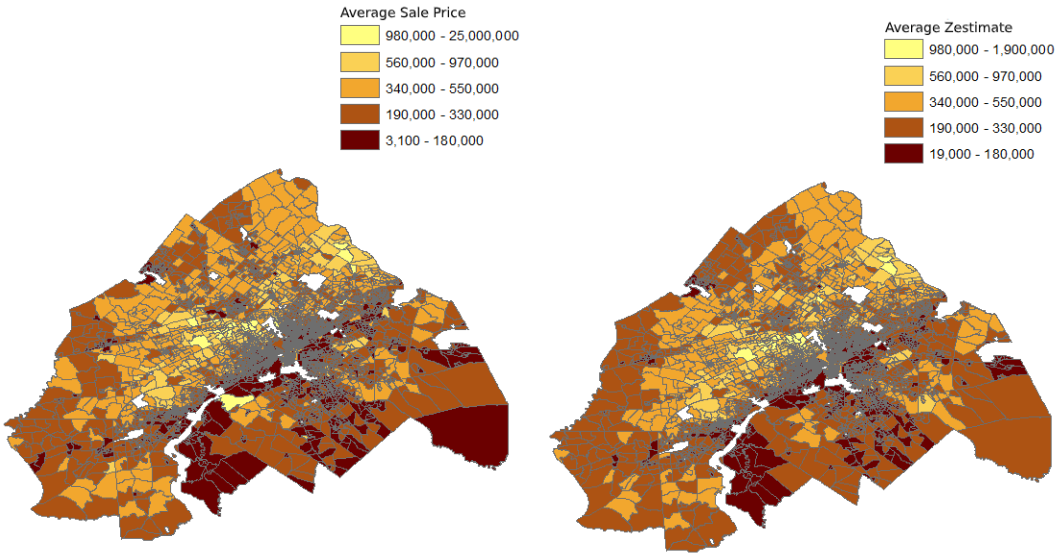
Notes: The census block groups included in the sample are shown in grey. It covers 4,108 census block groups in the Greater Philadelphia Area.

Electronic copy available at: https://ssrn.com/abstract=3584896

Figure A2: Instrumental Variables: Exclusion Restriction



(a) IV: Months since Last Reassessment



(b) DV: Sale Price

(c) IDV: Zestimate

Notes: The average number of months since the last reassessment/revaluation are plotted for each census block in Panel A. The average sale price is plotted for each census block in Panel B and the average Zestimate is plotted for each census block in Panel C.

Electronic copy available at: https://ssrn.com/abstract=3584896

A.3 Tables

Table A1: Summary of Covariates: Property Features

	Category	Covariates
S_i	Size	size
	Type	type
	Number of Bedrooms	number_of_bedrooms
	Number of Bathrooms	number_of_bathrooms
	Cooling	cooling_central, cooling_wall
	Heating	heating_electric, heating_gas, heating_wood, heating_air, heating_radiator, heating_baseboard
	Parking	parking_number, parking_carport, parking_detached, parking_attached
	Exterior Material	exterior_material_wood, exterior_material_vinyl, exterior_material_metal, exterior_material_brick, exterior_material_stone, exterior_material_cement, exterior_material_stucco
	Exterior Features	exterior_feature_deck, exterior_feature_porch, exterior_feature_garden, exterior_feature_lawn, exterior_feature_patio, exterior_feature_pool, exterior_feature_yard, exterior_feature_waterfront
	Views	view_park, view_mountain, view_water, view_territorial, view_city
	Water Sources	water_well, water_private
	Interior Flooring	interior_flooring_carpet, interior_flooring_hardwood
	Interior Heating	interior_heating_electric, interior_heating_gas, interior_heating_wood, interior_heating_air, interior_heating_radiator, interior_heating_baseboard
	Interior Appliances	interior_appliances_cleaning, interior_appliances_efficient, interior_appliances_stainless, interior_appliances_disposal, interior_appliances_efficiency, interior_appliances_hookups, interior_appliances_wall, interior_appliances_builtin, interior_appliances_island, interior_appliances_dishwasher, interior_appliances_washer

Electronic copy available at: https://ssrn.com/abstract=3584896

Table A2: Summary of Covariates: Neighborhood Socio-demographic Characteristics

	Category	Covariates
L_k	Population	population
	Gender	male
	Age	age1, age2, age3, age4
	Race	white, black, asian, othersingle
	Mobility	samehouse, greatbostonarea, abroad
	Working Places	principalcity, workinarea, workincity, workoutcity
	Commute Methods	car, publictransportation,othermethods
	Time Leaving Home	before7, between7and9
	Commute Times	workers, time1, time2, time3, time4
	Children	householdwithchild, marriedparents, singlemothers
	Household Status	nonfamilyhousehold, familyhousehold_married, familyhousehold_female
	Household Size	size1, size2
	Education	highschooldiploma, somecollege, bachelor, graduateschool,averageyear
	Poverty	povertyratio_hou
	Household Income	medianincome
	Income Sources	withearning, withsalary, withselfemployment, withpublicassis
	Housing: Units	housingunit, occupiedhousingunits
	Housing: Rooms	tworooms, threerooms, fourrooms, fiverooms, sixrooms, sevenrooms, eightrooms, ninerooms
	Housing: Types	singlefamily, townhouse, mutiplefamily

Electronic copy available at: https://ssrn.com/abstract=3584896

Table A3: The Effect of Zestimates on Final Sale Prices (Recent Sales)

Panel A: Sales in Last Month						
	(1)	(2)	(3)	(4)	(5)	(6)
	Sale Price	Sale Price	Sale Price	Sale Price	Sale Price	Sale Price
Zestimate_Sold	0.955***	0.281	0.419	0.419	0.513*	0.598**
	(0.031)	(0.268)	(0.296)	(0.296)	(0.284)	(0.257)
List Price		0.670**	0.527*	0.527*	0.438	0.393*
		(0.267)	(0.298)	(0.298)	(0.277)	(0.232)
Assessed Value		-0.057**	-0.004	-0.004	-0.016	0.008
		(0.025)	(0.026)	(0.026)	(0.024)	(0.029)
Log (Days on Market)			-18268.811***	-18268.811***	-16708.126***	-13676.983***
			(3872.564)	(3872.564)	(3875.786)	(4187.314)
Log (Days from "Pending" to "Sold")			6705.032***	6705.032***	6833.305**	7489.551***
			(2487.337)	(2487.337)	(2747.382)	(2624.267)
Month Fixed Effects	No	No	No	Yes	Yes	Yes
Socio-Demographic Controls	No	No	No	No	Yes	Yes
Property Feature Controls	No	No	No	No	No	Yes
Observations	826	561	408	408	402	402
Panel B: Sales in Last Three Months						
	(1)	(2)	(3)	(4)	(5)	(6)
	Sale Price	Sale Price	Sale Price	Sale Price	Sale Price	Sale Price
Zestimate_Sold	1.009***	0.685***	0.226**	0.226**	0.209*	0.208
	(0.014)	(0.123)	(0.104)	(0.103)	(0.112)	(0.155)
List Price		0.272**	0.712***	0.712***	0.719***	0.712***
		(0.118)	(0.098)	(0.098)	(0.101)	(0.135)
Assessed Value		-0.050***	-0.016*	-0.016*	-0.015*	-0.028**
		(0.009)	(0.009)	(0.009)	(0.009)	(0.013)
Log (Days on Market)			-17954.930***	-17896.563***	-17577.336***	-16699.422***
			(1715.162)	(1714.378)	(1712.232)	(2170.714)
Log (Days from "Pending" to "Sold")			7051.685***	7049.720***	6957.326***	6399.865***
			(1151.418)	(1151.035)	(1160.103)	(1271.241)
Month Fixed Effects	No	No	No	Yes	Yes	Yes
Socio-Demographic Controls	No	No	No	No	Yes	Yes
Property Feature Controls	No	No	No	No	No	Yes
Observations	10,238	6,483	4,225	4,225	4,188	4,173

Robust standard errors reported in parentheses are clustered at city×month level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The dependent variable is sale price. The time between listing and pending sale and the time between pending to closing are log-transformed to control their nonlinear effects. The regressions reported in Panel A use the purchases made in May 2019 for estimation while the regressions reported in Panel B use the purchases made from March to May 2019.

Electronic copy available at: https://ssrn.com/abstract=3584896

Table A4: First Stages (Recent Sales)

Panel A: Full Sample						
	(1)	(2)	(3)	(4)	(5)	(6)
	Zestimate	Zestimate	Zestimate	Zestimate	Zestimate	Zestimate
Months_Last_Update	228.281***	171.833***	171.089***	171.089***	178.617***	162.060**
	(75.280)	(36.251)	(58.787)	(58.787)	(64.663)	(63.281)
List Price		0.786***	0.798***	0.798***	0.762***	0.730***
		(0.041)	(0.068)	(0.068)	(0.078)	(0.087)
Assessed Value		0.246***	0.243**	0.243**	0.258**	0.202*
		(0.062)	(0.113)	(0.113)	(0.122)	(0.109)
Log (Days on Market)			-14261.522**	-14261.522**	-14640.705**	-14487.552***
			(6481.182)	(6481.182)	(5823.710)	(4841.351)
Log (Days from "Pending" to "Sold")			3399.035	3399.035	5605.201	2884.266
			(4379.311)	(4379.311)	(5119.440)	(4612.766)
Month Fixed Effects	No	No	No	Yes	Yes	Yes
Socio-Demographic Controls	No	No	No	No	Yes	Yes
Property Feature Controls	No	No	No	No	No	Yes
Observations	826	561	408	408	402	402
F statistic	15.04	72.90	48.68	48.68	44.99	32.11
R^2	0.018	0.971	0.975	0.975	0.978	0.983
Adjusted R^2	0.017	0.971	0.975	0.975	0.974	0.976
Panel B: Subsample without Missing Information						
	(1)	(2)	(3)	(4)	(5)	(6)
	Zestimate	Zestimate	Zestimate	Zestimate	Zestimate	Zestimate
Months_Last_Update	221.339***	256.893***	102.498***	102.445***	92.046***	78.325***
	(36.837)	(96.481)	(17.705)	(17.722)	(15.770)	(13.981)
List Price		0.633***	0.835***	0.835***	0.816***	0.817***
		(0.126)	(0.022)	(0.022)	(0.023)	(0.024)
Assessed Value		0.430***	0.165***	0.165***	0.140***	0.102***
		(0.162)	(0.034)	(0.034)	(0.031)	(0.027)
Log (Days on Market)			-11947.337***	-11862.975***	-10970.462***	-10688.369***
			(1249.633)	(1251.189)	(1155.346)	(1198.952)
Log (Days from "Pending" to "Sold")			3138.087***	3137.065***	3257.768***	3247.184***
			(915.980)	(916.247)	(880.256)	(866.687)
Month Fixed Effects	No	No	No	Yes	Yes	Yes
Socio-Demographic Controls	No	No	No	No	Yes	Yes
Property Feature Controls	No	No	No	No	No	Yes
Observations	10,238	6,483	4,225	4,225	4,188	4,173
F statistic	372.74	1225.97	317.68	317.19	249.43	177.03
R^2	0.035	0.898	0.967	0.967	0.970	0.973
Adjusted R^2	0.035	0.898	0.967	0.967	0.970	0.972

Robust standard errors reported in parentheses are clustered at city×month level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The dependent variable is sale price. The time between listing and pending sale and the time between pending to closing are log-transformed to control their nonlinear effects. The regressions reported in Panel A use the purchases made in May 2019 for estimation while the regressions reported in Panel B use the purchases made from March to May 2019.

Electronic copy available at: https://ssrn.com/abstract=3584896

Table A5: The Effect of Zestimates on Final Sale Prices (Alternative Specification)

Panel A: Full Sample						
	(1)	(2)	(3)	(4)	(5)	(6)
	Sale Price	Sale Price	Sale Price	Sale Price	Sale Price	Sale Price
Zestimate_Sold	0.998***	0.507***	0.337***	0.335***	0.170**	0.094
	(0.008)	(0.064)	(0.048)	(0.048)	(0.075)	(0.083)
List Price		0.439***	0.599***	0.600***	0.715***	0.756***
		(0.060)	(0.045)	(0.045)	(0.061)	(0.066)
Assessed Value		-0.046***	-0.024***	-0.023***	0.017	0.005
		(0.004)	(0.003)	(0.003)	(0.018)	(0.017)
Log (Days on Market)			-12967.583***	-12778.470***	-13345.063***	-13641.824***
			(614.912)	(608.319)	(768.189)	(816.893)
Log (Days from "Pending" to "Sold")			5149.837***	5045.879***	5515.926***	5589.280***
			(364.152)	(359.637)	(426.878)	(435.927)
Month Fixed Effects	No	No	No	Yes	Yes	Yes
City Fixed Effects	No	No	No	No	Yes	Yes
Property Feature Controls	No	No	No	No	No	Yes
Observations	120,482	88,110	52,981	52,981	52,981	52,757
Panel B: Subsample without Missing Information						
	(1)	(2)	(3)	(4)	(5)	(6)
	Sale Price	Sale Price	Sale Price	Sale Price	Sale Price	Sale Price
Zestimate_Sold	1.009***	0.391***	0.339***	0.337***	0.159**	0.094
	(0.004)	(0.049)	(0.050)	(0.050)	(0.077)	(0.083)
List Price		0.551***	0.597***	0.598***	0.725***	0.756***
		(0.046)	(0.047)	(0.047)	(0.063)	(0.066)
= Assessed Value		-0.037***	-0.024***	-0.024***	0.018	0.005
		(0.004)	(0.003)	(0.003)	(0.018)	(0.017)
Log (Days on Market)			-12924.107***	-12731.735***	-13411.288***	-13641.824***
			(631.588)	(624.879)	(782.676)	(816.893)
Log (Days from "Pending" to "Sold")			5146.118***	5042.462***	5552.183***	5589.280***
			(367.350)	(362.473)	(430.062)	(435.927)
Month Fixed Effects	No	No	No	Yes	Yes	Yes
City Fixed Effects	No	No	No	No	Yes	Yes
Property Feature Controls	No	No	No	No	No	Yes
Observations	52,757	52,757	52,757	52,757	52,757	52,757

Robust standard errors reported in parentheses are clustered at city×month level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The dependent variable is sale price. The time between listing and pending sale and the time between pending to closing are log-transformed to control their nonlinear effects.

Electronic copy available at: https://ssrn.com/abstract=3584896

Table A6: First Stages (Alternative Specification)

Panel A: Full Sample						
	(1)	(2)	(3)	(4)	(5)	(6)
	Zestimate	Zestimate	Zestimate	Zestimate	Zestimate	Zestimate
Months_Last_Update	265.854***	163.572***	118.466***	118.341***	99.387***	87.254***
	(11.300)	(15.600)	(10.673)	(10.672)	(16.012)	(15.253)
List Price		0.748***	0.801***	0.800***	0.747***	0.733***
		(0.020)	(0.013)	(0.013)	(0.016)	(0.017)
Assessed Value		0.307***	0.223***	0.223***	0.285***	0.242***
		(0.031)	(0.022)	(0.022)	(0.026)	(0.026)
Log (Days on Market)			-10102.850***	-9828.774***	-8337.747***	-8217.160***
			(558.297)	(547.414)	(496.131)	(476.036)
Log (Days from "Pending" to "Sold")			3719.434***	3415.088***	3569.213***	3300.225***
			(370.919)	(370.504)	(354.101)	(343.639)
Month Fixed Effects	No	No	No	Yes	Yes	Yes
City Fixed Effects	No	No	No	No	Yes	Yes
Property Feature Controls	No	No	No	No	No	Yes
Observations	120,482	88,110	52,981	52,981	52,981	52,757
F statistic	3939.31	980.59	2397.47	2391.47	5088.59	1398.43
R^2	0.032	0.589	0.928	0.928	0.932	0.931
Adjusted R^2	0.032	0.589	0.928	0.928	0.931	0.931
Panel B: Subsample without Missing Information						
	(1)	(2)	(3)	(4)	(5)	(6)
	Zestimate	Zestimate	Zestimate	Zestimate	Zestimate	Zestimate
Months_Last_Update	288.451***	116.164***	111.237***	111.105***	104.950***	87.254***
	(13.704)	(10.935)	(10.908)	(10.910)	(10.480)	(15.253)
List Price		0.804***	0.808***	0.807***	0.781***	0.733***
		(0.013)	(0.013)	(0.013)	(0.015)	(0.017)
Assessed Value		0.204***	0.205***	0.206***	0.190***	0.242***
		(0.022)	(0.022)	(0.022)	(0.021)	(0.026)
Log (Days on Market)			-10025.264***	-9774.361***	-9000.666***	-8217.160***
			(555.432)	(544.513)	(515.616)	(476.036)
Log (Days from "Pending" to "Sold")			3754.129***	3629.701***	3625.058***	3300.225***
			(366.551)	(365.620)	(354.691)	(343.639)
Month Fixed Effects	No	No	No	Yes	Yes	Yes
City Fixed Effects	No	No	No	No	Yes	Yes
Property Feature Controls	No	No	No	No	No	Yes
Observations	52,757	52,757	52,757	52,757	52,757	52,757
F statistic	2680.57	2293.42	2107.03	2100.93	1732.58	734.16
R^2	0.049	0.927	0.928	0.928	0.930	0.935
Adjusted R^2	0.049	0.927	0.928	0.928	0.930	0.934

Robust standard errors reported in parentheses are clustered at city×month level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The dependent variable is sale price. The time between listing and pending sale and the time between pending to closing are log-transformed to control their nonlinear effects.

Electronic copy available at: https://ssrn.com/abstract=3584896

Table A7: Robustness Check: Excluding Irregular Transactions

Panel A: Interaction with Frequency of Similar Transactions Nearby ($t-6$)								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Area	Block Group				City			
Measurement	per Capita		per Housing Unit		per Capita		per Housing Unit	
Zestimate_Sold	0.346***	0.241***	0.347***	0.239***	0.353***	0.238***	0.346***	0.238***
	(0.049)	(0.054)	(0.049)	(0.055)	(0.048)	(0.054)	(0.047)	(0.054)
# Transactions	-1283465.204***	-431214.423	-491146.326**	-144981.184	-928884.192**	-139023.185	-266809.429***	-59254.066
	(411187.553)	(381471.852)	(198683.449)	(161015.638)	(415417.446)	(178413.526)	(77349.750)	(54369.531)
Zestimate_Sold	4.234***	3.151**	1.776***	1.196**	2.441**	0.588	0.724***	0.251*
× # Transactions	(1.443)	(1.394)	(0.634)	(0.540)	(1.030)	(0.434)	(0.190)	(0.133)
List Price	0.584***	0.643***	0.581***	0.645***	0.579***	0.651***	0.586***	0.651***
	(0.045)	(0.047)	(0.045)	(0.047)	(0.046)	(0.048)	(0.044)	(0.048)
Assessed Value	-0.024***	-0.031***	-0.025***	-0.031***	-0.027***	-0.032***	-0.026***	-0.032***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Log (Days on Market)	-12846.030***	-12797.553***	-12839.669***	-12813.182***	-12627.432***	-12818.694***	-12743.050***	-12815.553***
	(622.769)	(608.088)	(619.970)	(607.690)	(644.643)	(612.469)	(618.371)	(609.689)
Log (Days from "Pending" to "Sold")	5120.181***	5223.773***	5131.294***	5234.423***	5088.570***	5241.073***	5117.284***	5241.056***
	(371.069)	(372.286)	(371.826)	(372.316)	(364.660)	(369.314)	(362.322)	(368.388)
Month Fixed Effects	No	Yes	No	Yes	No	Yes	No	Yes
Socio-Demographic Controls	No	Yes	No	Yes	No	Yes	No	Yes
Property Feature Controls	No	Yes	No	Yes	No	Yes	No	Yes
Observations	52,981	52,164	52,981	52,164	52,981	52,164	52,981	52,164
Panel B: Interaction with Frequency of Similar Transactions Nearby ($t-12$)								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Area	Block Group				City			
Measurement	per Capita		per Housing Unit		per Capita		per Housing Unit	
Zestimate_Sold	0.346***	0.241***	0.347***	0.239***	0.353***	0.240***	0.347***	0.240***
	(0.048)	(0.054)	(0.048)	(0.055)	(0.048)	(0.054)	(0.047)	(0.054)
# Transactions	-794187.629***	-279127.294	-280325.184**	-83079.017	-497698.857**	-117310.936	-150053.503***	-46428.193
	(252223.437)	(229384.500)	(121488.689)	(100591.578)	(200713.777)	(100783.925)	(41717.756)	(31593.745)
Zestimate_Sold	2.689***	2.111**	1.053***	0.774**	1.322***	0.447*	0.415***	0.182**
× # Transactions	(0.868)	(0.834)	(0.382)	(0.332)	(0.490)	(0.240)	(0.102)	(0.076)
List Price	0.584***	0.643***	0.583***	0.645***	0.580***	0.649***	0.586***	0.649***
	(0.045)	(0.047)	(0.044)	(0.047)	(0.045)	(0.048)	(0.044)	(0.047)
Assessed Value	-0.024***	-0.031***	-0.024***	-0.030***	-0.026***	-0.032***	-0.025***	-0.032***
	(0.003)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.003)	(0.004)
Log (Days on Market)	-12873.928***	-12817.599***	-12871.119***	-12835.454***	-12668.480***	-12796.219***	-12751.256***	-12795.818***
	(621.320)	(607.848)	(620.390)	(608.736)	(629.564)	(609.994)	(615.125)	(608.208)
Log (Days from "Pending" to "Sold")	5129.426***	5211.745***	5140.891***	5224.837***	5107.555***	5232.039***	5126.767***	5232.833***
	(369.488)	(370.710)	(371.289)	(371.523)	(362.468)	(368.152)	(361.702)	(367.574)
Month Fixed Effects	No	Yes	No	Yes	No	Yes	No	Yes
Socio-Demographic Controls	No	Yes	No	Yes	No	Yes	No	Yes
Property Feature Controls	No	Yes	No	Yes	No	Yes	No	Yes
Observations	52,981	52,164	52,981	52,164	52,981	52,164	52,981	52,164

Robust standard errors reported in parentheses are clustered at city×month level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The dependent variable is sale price. The time between listing and pending sale and the time between pending to closing are log-transformed to control their nonlinear effects. The number of transactions (adjusted for the neighborhood size) is mean centered to allow easy interpretation of the main effects.

Electronic copy available at: https://ssrn.com/abstract=3584896

Table A8: Robustness Check: Including Transactions in Other Segments

Panel A: Interaction with Average Deviation for Similar Transactions Nearby ($t-6$)								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Area	Block Group				City			
Measurement	$Zestimate_t$		$Zestimate_{t-1}$		$Zestimate_t$		$Zestimate_{t-1}$	
Zestimate_Sold	0.327***	0.228***	0.327***	0.228***	0.335***	0.233***	0.335***	0.233***
	(0.048)	(0.055)	(0.048)	(0.055)	(0.048)	(0.054)	(0.048)	(0.054)
% Deviation	607.840	105.046	625.326	13.921	-3224.611**	-2328.539**	-3809.026**	-2818.507**
	(783.025)	(430.648)	(545.037)	(252.289)	(1321.064)	(997.843)	(1561.759)	(1181.979)
Zestimate_Sold	-0.006	0.001	-0.006*	0.001	0.005	0.012**	0.007	0.013**
× % Deviation	(0.004)	(0.002)	(0.003)	(0.001)	(0.007)	(0.005)	(0.007)	(0.005)
List Price	0.606***	0.660***	0.606***	0.660***	0.600***	0.657***	0.601***	0.657***
	(0.045)	(0.048)	(0.045)	(0.048)	(0.045)	(0.048)	(0.045)	(0.048)
Assessed Value	-0.024***	-0.031***	-0.024***	-0.031***	-0.023***	-0.032***	-0.023***	-0.032***
	(0.003)	(0.004)	(0.003)	(0.004)	(0.003)	(0.004)	(0.003)	(0.004)
Log (Days on Market)	-13193.512***	-13026.155***	-13158.554***	-12990.231***	-12973.777***	-12907.775***	-12937.282***	-12874.245***
	(610.837)	(616.487)	(610.593)	(616.226)	(619.673)	(608.580)	(618.863)	(608.071)
Log (Days from "Pending" to "Sold")	5253.598***	5316.434***	5241.672***	5303.866***	5147.748***	5250.566***	5135.670***	5239.306***
	(363.703)	(372.565)	(363.616)	(372.434)	(365.348)	(369.104)	(365.138)	(368.970)
Month Fixed Effects	No	Yes	No	Yes	No	Yes	No	Yes
Socio-Demographic Controls	No	Yes	No	Yes	No	Yes	No	Yes
Property Feature Controls	No	Yes	No	Yes	No	Yes	No	Yes
Observations	52,309	51,524	52,311	51,526	52,889	52,079	52,891	52,081
Panel B: Interaction with Average Deviation for Similar Transactions Nearby ($t-12$)								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Area	Block Group				City			
Measurement	$Zestimate_t$		$Zestimate_{t-1}$		$Zestimate_t$		$Zestimate_{t-1}$	
Zestimate_Sold	0.327***	0.228***	0.327***	0.229***	0.329***	0.231***	0.329***	0.231***
	(0.048)	(0.055)	(0.048)	(0.055)	(0.048)	(0.054)	(0.048)	(0.054)
% Deviation	961.481	127.186	819.373	56.314	-4636.818**	-2581.407*	-5974.797**	-3684.205**
	(931.458)	(452.163)	(575.523)	(265.360)	(2055.943)	(1406.958)	(2645.534)	(1826.377)
Zestimate_Sold	-0.008*	0.000	-0.008**	0.000	0.001	0.008	0.006	0.012*
×% Deviation	(0.004)	(0.002)	(0.004)	(0.002)	(0.009)	(0.006)	(0.009)	(0.007)
List Price	0.606***	0.660***	0.606***	0.660***	0.605***	0.659***	0.605***	0.659***
	(0.045)	(0.048)	(0.045)	(0.048)	(0.045)	(0.048)	(0.045)	(0.048)
Assessed Value	-0.024***	-0.031***	-0.024***	-0.031***	-0.023***	-0.031***	-0.023***	-0.031***
	(0.003)	(0.004)	(0.003)	(0.004)	(0.003)	(0.004)	(0.003)	(0.004)
Log (Days on Market)	-13180.586***	-13018.715***	-13146.123***	-12982.399***	-13017.191***	-12931.334***	-12977.848***	-12898.397***
	(610.969)	(616.652)	(610.295)	(616.285)	(615.602)	(608.576)	(614.731)	(608.073)
Log (Days from "Pending" to "Sold")	5244.152***	5310.026***	5233.125***	5297.344***	5177.715***	5266.077***	5169.186***	5256.957***
	(363.594)	(372.422)	(363.334)	(372.255)	(364.037)	(369.471)	(363.827)	(369.377)
Month Fixed Effects	No	Yes	No	Yes	No	Yes	No	Yes
Socio-Demographic Controls	No	Yes	No	Yes	No	Yes	No	Yes
Property Feature Controls	No	Yes	No	Yes	No	Yes	No	Yes
Observations	52,351	51,561	52,353	51,563	52,901	52,090	52,903	52,092

Robust standard errors reported in parentheses are clustered at city×month level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The dependent variable is sale price. The time between listing and pending sale and the time between pending to closing are log-transformed to control their nonlinear effects. The accuracy of the Zestimate for a recent sale is calculated as $|Zestimate - Sale\_Price|/Zestimate$ and the average accuracy is mean centered to allow easy interpretation of the main effects.

Electronic copy available at: https://ssrn.com/abstract=3584896