HARVARD | BUSINESS | SCHOOL

Nowcasting Gentrification: Using Yelp Data to Quantify Neighborhood Change

Edward L. Glaeser
Michael Luca

Hyunjin Kim

Working Paper 18-077

Nowcasting Gentrification: Using Yelp Data to Quantify Neighborhood Change

Edward L. Glaeser
Harvard University

Hyunjin Kim
Harvard Business School

Michael Luca
Harvard Business School

Working Paper 18-077

NOWCASTING GENTRIFICATION: USING YELP DATA TO QUANTIFY NEIGHBORHOOD CHANGE

Edward L. Glaeser, Hyunjin Kim and Michael Luca*

Session Title: Economic Applications of Machine Learning
Session Chair: Daniel Björkegren

Abstract

Data from digital platforms have the potential to improve our understanding of gentrification and enable new measures of how neighborhoods change in close to real time. Combining data on businesses from Yelp with data on gentrification from the Census, Federal Housing Finance Agency, and Streetscore (an algorithm using Google Streetview), we find that gentrifying neighborhoods tend to have growing numbers of local groceries, cafes, restaurants, and bars, with little evidence of crowd-out of other types of businesses. For example, the entry of a new coffee shop into a zip code in a given year is associated with a 0.5 percent increase in housing prices. Moreover, Yelp measures of local business activity provide leading indicators for housing price changes and help to forecast which neighborhoods are gentrifying.

I. Introduction

"Gentrification: New Yorkers can sense it immediately. It plumes out of Darling Coffee, on Broadway and 207th Street, and mingles with the live jazz coming from the Garden Café next door" – New York Magazine (2014)

Gentrification has emerged as a central policy issue in cities from New York to Edinburgh to Seoul, but measuring neighborhood change can be difficult. While data from agencies such as the Census Bureau have provided important insight into gentrification, these datasets often arrive only after a long lag and provide limited information about, for example, the types of businesses operating in a neighborhood. Data from digital platforms have the potential

* Glaeser, Harvard Economics Department, 1805 Cambridge St., Cambridge, MA 02138; eglaeser@harvard.edu; Kim, Harvard Business School, Soldiers Field Rd., Boston, MA 02163; hkim@hbs.edu, Luca, Harvard Business School, Soldiers Field Rd., Boston, MA 02163; mluca@hbs.edu. We thank Byron Perpetua for excellent research assistance. We thank Susan Athey, Shane Greenstein, and Luther Lowe for valuable feedback, and Yelp for providing data for this analysis. All remaining errors are our own.

to improve our understanding of gentrification, by providing unique insight into the local economy, helping to predict which neighborhoods are gentrifying, and more generally shedding light on the evolution of neighborhoods. In this paper, we explore the potential for Yelp data to provide real-time information on how neighborhoods change.

How might we expect local business activity to change with gentrification? As a neighborhood's residential demographics change, so might the business landscape, since it is a function of local demand (Waldfogel 2008). For example, richer neighborhoods might attract more businesses if wealthier residents spend more at local businesses or are willing to pay more to cut travel time – leading to business densification. Gentrifying neighborhoods might also attract more upscale establishments, shifting the composition of businesses from downscale to upscale establishments. Alternatively, the local economy might provide insight into where gentrification might occur. For example, houses near a Starbucks have seen increased prices in recent years (Rascoff and Humphries 2015). While their claim is not meant to be causal, it does suggest that information about the local economy might help to understand, and even predict, gentrification.

Combining Yelp and American Community Survey (ACS) data, we explore this potential and find that entry of Starbucks (and cafés more generally) into a neighborhood is in fact indicative of housing price growth across the U.S. The number of reviews of Starbucks increases predictive power, suggesting that gentrifying neighborhoods might also attract more reviewers. We then turn to three demographic measures of neighborhood change in New York City based on ACS at the Census Zip Code Tabulation Area (ZCTA) level: percent with a college degree, percent aged between 25 and 34, and percent white. We find that growth in groceries, laundromats, bars, and cafes are particularly good predictors of increases in the share of college

educated within an area. However, changes in the business ecosystem are less indicative of changes in the age and racial composition of a neighorbood. Finally, we examine how local business activity varies with changes in the block-level Streetscore, a computer-generated measure of how humans perceive the safety of a Google Streetview image, which proxies for the change in the physical quality of the neighborhood (Naik et al. (2017)). Changes to neighborhood perceived quality (based on images of the neighborhood) can also be predicted by changes in the local economy. For example, increases in the number of Starbucks and cafes, vegetarian restaurants, and wine bars and bars more generally (drawing on Yelp's business classifications) are all markers of improving neighborhood quality, as measured by Streetscore.

In our working paper, we expand our analysis to more cities and also explore whether gentrification precedes or follows the changes in business categories (Glaeser, Kim, and Luca (2018)). We find that business changes predict future as well as contemporaneous gentrification. While we do not focus on causal links in this paper, our results highlight potential for new data sources (in this case, Yelp) to provide new insight into where gentrification is occurring and what changes transpire in the business landscape.

II. Data Description

Our first measure of gentrification is housing price data provided by the Federal Housing Finance Agency (FHFA). This data is an annual repeat sales index for over 18,000 five-digit ZIP codes in the U.S., described in Bogin, Doerner and Larson (2016). We use data from 2012 to 2016, and the average real growth of this index over this period is 3.1 percentage points.

While ZIP code level pricing data is available annually for a large number of ZIP codes, our demographic data is available only for five-year windows, as smaller geographic units are only surveyed sporadically in the ACS. Considering a long difference between the 2007-2011

period and the 2012-2016 period, we use three measures of neighborhood demographics: percent college educated, percent aged between 25 and 34, and share of the population that is white. Because education tends to be reliably correlated with both income and housing costs, the percent of people with college education in an area provides a reasonable metric for gentrification. In our sample, the average ZIP code in New York saw the share of adults with college degrees increase by 2.6 percent.

Our final measure of neighborhood change is change in Streetscore, drawn from Naik et al. (2017). This measure starts with a crowdsourced data set in which respondents rated images from Google Streetview on perceived safety. These ratings were used as training data for computer vision techniques, which generated Streetscores for more neighborhoods. We interpret Streetscore measure as a proxy for the overall physical quality of the neighborhood, rather than safety per se. In related work, Naik et al. (2017) find that changes in Streetscore correlate with density and education. We have Streetscore data from 2007 and 2014.

For measures of changes in business categories, we use data from Yelp, an online platform with business listings that are sourced through user submissions, business owner reports, partner acquisitions, and internal data quality checks. The data begin in 2004 when Yelp was founded, which enables business listings to be aggregated at the ZIP code, city, state, and country level for any given time period post-2004. To predict housing prices, we aggregate Yelp data annually. To examine correlations with demographic indicators, we average over two corresponding five-year periods to the ACS.

Despite its granularity and availability, Yelp data has limitations, discussed in further detail in Glaeser, Kim and Luca (2017). Yelp's business classification is assigned through user and business owner reports, which results in unsystematic industry categorization that does not

correspond to government data sets. Furthermore, the quality of Yelp data depends on the degree of Yelp adoption, which has grown over time. Given these issues, we only count businesses as open if they have received at least one recommended Yelp review.

III. Results on Local Housing Prices

We first explore the ability of Yelp data to predict contemporaneous changes in housing price growth at the ZIP code level, looking at the period from 2012 to 2016. We start by following Rascoff and Humphries (2015) who link proximity to Starbucks and price growth on Zillow. In our version, we examine whether price growth is correlated with contemporaneous growth in the number of Starbucks cafes – which allows us to understand whether the entry of Starbucks is an indicator of gentrification.

Table 1 Correlations between Annual Percent Change in HPI and Annual Absolute Change in the Number of Starbucks and Cafes across ZIP codes (2012-2016)

	(1)	(2)	(3)	(4)	(5)	(6)
	Percent Change in HPI	Percent Change in HPI	Percent Change in HPI	Percent Change in HPI	Percent Change in HPI	Percent Change in HPI
Yelp Starbucks/Cafes Growth	0.536***	0.171*	0.206*	0.535***	0.020	0.250***
	(0.082)	(0.075)	(0.087)	(0.023)	(0.023)	(0.024)
Yelp Starbucks/Cafes Growth (lag1)			0.261**			0.277***
			(0.086)			(0.024)
Yelp Starbucks/Cafes Growth (lag2)			0.195**			0.292***
			(0.070)			(0.024)
Yelp Growth in Closed Starbucks/Cafes			-0.042			-0.077*
			(0.149)			(0.033)
Yelp Starbucks/Cafes			0.136***			0.009***

Reviews Growth						
			(0.007)			(0.001)
Constant	-0.858***	-0.826***	-0.952***	-1.523***	-1.231***	-1.679***
	(0.057)	(0.056)	(0.061)	(0.044)	(0.043)	(0.048)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
ZIP FE	No	Yes	No	No	Yes	No
Observations	24865	24865	24865	59180	59180	59180
Adjusted R^2	0.240	0.372	0.256	0.157	0.211	0.172

All regressions include a full set of calendar year dummies and cluster standard errors at the ZIP Code level. Models (1) through (3) show correlations between HPI and Starbucks, and models (4) through (6) show correlations between HPI and cafes. * p<0.10, ** p<0.05, *** p<0.01.

Our first specification in Table 1 regresses percent growth in home prices on the absolute increase in the number of Starbucks in the ZIP code during that same year. We include year dummies and cluster our standard errors by ZIP code. A one-unit increase in the number of Starbucks in a given year is associated with a 0.5 percent increase in housing prices. This effect is large, both economically and statistically, but the explanatory power of the Starbucks control is modest.

The direction of causality in this relationship is a priori unclear. For example, Starbucks may target its cafes in places that are on the upswing, so the correlation may reflect Starbucks' strategy. On the other hand, businesses may contribute to gentrification. To partially distinguish between these hypotheses, our second regression includes a ZIP code fixed effect. Our time period is short, so if Starbucks is targeting growing areas, then the fixed effect should eliminate much of the correlation. Including these fixed effects causes the estimated coefficient to fall to 0.17 and the r-squared to rise to 0.37.

Because our time period is too short to simultaneously estimate ZIP code level fixed effects and the timing of the relationship between Starbucks and price growth, we drop the ZIP code fixed effects for the remainder of the table. In the third regression, we include both the

current and lagged Starbucks growth, as well as the change in the number of Starbucks that are closed and the growth in the number of Starbucks reviews. The growth in the number of Starbucks reviews is also predictive of neighborhood change: a 10-unit increase in the number of reviews is associated with a 1.4 percent increase in housing prices in the ZIP code. Including this variable reduces the significance of the other Starbucks variables and increases the r-squared of the regression from 0.24 to 0.26. Since the presence of a Starbucks is less important than whether the community reviews Starbucks, this finding pushes against the interpretation that people are paying for proximity to Starbucks.

While Starbucks may be a particularly prominent coffee shop, it is not the only possible retail establishment that may signal gentrification at the local level. In the next three regressions of Table 1, we expand our analysis to include all of the cafes listed in Yelp over the same time period. This change increases the number of ZIP codes, because more ZIP codes have at least one café in our time period. We find similar results, though the magnitude of the power of café reviews is somewhat weaker than for Starbucks. The difference between Starbucks and café results lends some support to the upscaling hypothesis. In the working paper version of this paper, we expand our analysis to other industries that Yelp has classified (Glaeser, Kim, and Luca (2018)). In many cases, as in Starbucks, the number of Yelp reviews provided additional predictive power beyond the entry of a business, suggesting that both changes in the local economy and changes in the use of Yelp are related to gentrification.

IV. Results on Demographic Change

We now explore whether the local business ecosystem shifts with demographic changes in a neighborhood. We focus on New York City and examine demographic changes between periods

2007-2011 and 2012-2016, as well as Streetscore change between 2007-2014. We list only business categories from Yelp that appear in more than 100 ZIP codes, except for Starbucks.

Table 2 Correlations between Changes in Demographics, Street Score, and Yelp Number of Establishments between 2007-2011 and 2012-2016 across New York City ZIP Codes

	(1)	(2)	(3)		(4)	
	Change in percent of college educated	Change in percent of ages 25 to 34	Change in percent white	*Obs. (1)-(3)*	Change in street score 2007-2014	*Obs. (4)*
Change in the number of groceries	0.352***	0.178*	0.189*	173	0.103	146
	(0.000002)	(0.019)	(0.013)		(0.218)	
Change in the number of laundromats	0.338***	0.200*	0.120	122	0.034	109
	(0.0001)	(0.027)	(0.187)		(0.729)	
Change in the number of cafes	0.319***	0.093	0.084	179	0.318***	150
	(0.00001)	(0.216)	(0.264)		(0.00007)	
Change in the number of bars	0.313***	0.140	0.114	176	0.327***	147
	(0.00002)	(0.064)	(0.132)		(0.00005)	
Change in the number of restaurants	0.270***	0.152*	0.098	180	0.275***	150
	(0.0003)	(0.041)	(0.191)		(0.001)	
Change in the number of barbers	0.237**	0.197*	0.084	160	0.316***	140
	(0.003)	(0.012)	(0.291)		(0.0001)	
Change in the number of wine bars	0.232**	0.143	0.144	136	0.339***	119
	(0.007)	(0.097)	(0.094)		(0.0002)	
Change in the number of convenience stores	0.222**	0.079	0.128	162	0.208*	141
	(0.004)	(0.320)	(0.104)		(0.014)	
Change in the number of fast food restaurants	0.200**	0.024	0.046	173	0.270***	148
	(0.008)	(0.758)	(0.544)		(0.001)	
Change in the number of $$$$ restaurants	0.193**	0.125	0.066	180	0.148	150
	(0.009)	(0.094)	(0.378)		(0.070)	
Change in the number of vegetarian restaurants	0.175	0.067	0.054	108	0.372***	100
	(0.069)	(0.490)	(0.580)		(0.0001)	
Change in the number of florists	0.173*	0.185*	0.053	142	0.290***	127
	(0.039)	(0.028)	(0.534)		(0.001)	
Change in the number of Starbucks	0.067	-0.099	-0.010	95	0.355***	88
	(0.522)	(0.338)	(0.923)		(0.001)	

* p<0.10, ** p<0.05, *** p<0.01.

Each row in Table 2 shows the pairwise correlation between the growth in the number of businesses in each category and the change in the demographic variable or the Streetscore. We use the absolute change in the number of establishments, which eliminates the need to worry about cases where there are zero establishments in the pre-period. Beneath each correlation coefficient, we report the p-value or the estimated probability that the correlation is actually zero. We also show the number of observations, which differ across rows, because not all ZIP codes have at least one example of the business category in the relevant period.

We order the results by the strength of the correlation with change in the share of the population in the ZIP code that is college-educated. Our first row shows that the change in the number of groceries is significantly correlated (0.35) with the change in the share of adults with college degrees. Its correlations with the age and racial composition of the ZIP code are also significant at the five percent level, but they are about one-half the size of the correlation with change in percent college educated. These results seem compatible with the literature on "food deserts" that documents how poorer people live in areas with fewer options for healthy food.

The second row shows the 0.338 correlation between growth in laundromats and the share of the population with college degrees. The number of laundromats also has a 0.2 correlation with the share of the population that is young, which is perhaps less surprising. As laundromats are rarely "upscale," this result seems more compatible with business densification.

The table also shows significant correlations between the change in the share of the population that is college educated and changes in the number of cafes, bars, restaurants, barbers, wine bars, convenience stores, fast food restaurants, florists, and restaurants categorized by Yelp as being pricey. Restaurants, barbers, and florists also correlate with the number of people who are young. Correlations with the racial composition were almost uniformly weaker.

In our working paper, we reproduce these results for Boston, Chicago, Los Angeles and San Francisco, and also examine correlations with the number of Yelp reviews by category. Many of the patterns are broadly similar, with two significant differences. The number of laundromats is no longer a strong correlation of gentrification. In the other four cities, unlike New York, several of the review counts correlate strongly with the number of younger people in the ZIP code – potentially due to geographic variation in the age of Yelp reviewers.

Our final outcome is the physical change in the neighborhood as measured by Streetscore. As before, we begin with New York City and then turn to other large urban areas in our working paper. To keep results comparable, we continue to look at ZIP code level data, although there is no reason why we could not look at the block itself. At the ZIP code level, the strongest correlation (0.37) is with the number of vegetarian restaurants, which had a much weaker correlation with the change in the share of college educated. The second strongest correlation (0.36) is with the change in the number of Starbucks restaurants, and the third strongest (0.34) is with wine bars. This mirrors our results with demographic change.

V. Conclusion

In Glaeser, Kim, and Luca (2017), we highlight the potential for Yelp data to forecast local economic activity. Here, we highlight the potential for data from digital platforms to improve our understanding of gentrification: it can provide data in close to real time, and enable new measures, such as ways to categorize businesses to understand which parts of the economy are growing. While our focus is on measurement and prediction, our results also suggest that businesses respond to exogenous changes in neighborhood composition. In our working paper, we find that Yelp establishments from 2007-2011 predict changes in education levels over the

next five years, but not the reverse. Consequently, it is possible that Yelp is also measuring neighborhood amenities that help drive neighborhood change.

Our results also relate to a growing body of research exploring the ways in which digital platforms contain valuable data that can be used to enhance our understanding of the economy. While these platforms are not a substitute for traditional government statistical data, they provide an important complement – offering novel insights into the economy, often in close to real time.

References

Bogin, Alexander, William Doerner, and William D. Larson (2016). "Local house price growth acceleration" (2016). FHFA Staff Working Paper Series, No. 16-02, Federal Housing Finance Agency, Washington, DC.

Davidson, Justin, "Is Gentrification all that bad?" New York Magazine. Feb 2, 2014. <http://nymag.com/news/features/gentrification-2014-2/>. Accessed Jan. 4, 2018.

Glaeser, Edward L. and Kim, Hyunjin and Luca, Michael (2017). "Nowcasting the Local Economy: Using Yelp Data to Measure Economic Activity." NBER Working Paper No. w24010.

Glaeser, Edward L. and Kim, Hyunjin and Luca, Michael (2018). "Nowcasting Gentrification: Using Yelp Data to Predict Neighborhood Change" NBER Working Paper.

Naik, Nikhil, Scott Duke Kominers, Ramesh Raskar, Edward L. Glaeser (2017). "Computer vision uncovers predictors of physical urban change." Proceedings of the National Academy of Sciences 114(29): 7571-7576.

Rascoff, Spencer, and Stan Humphries (2015). "Confirmed: Starbucks knows the next hot neighborhood before everybody else does." Quartz. Jan. 28, 2015. < https://qz.com/334269/what-starbucks-has-done-to-american-home-values/>. Accessed Jan. 4, 2018.

Waldfogel, Joel (2008). "The median voter and the median consumer: Local private goods and population composition." *Journal of Urban Economics* 63(2): 567-582.