Nowcasting the Local Economy:

Using Yelp Data to Measure Economic Activity[1]

Edward L. Glaeser[†], Hyunjin Kim[‡], and Michael Luca[§]

June 2019

Abstract

Can new data sources from online platforms help to measure local economic activity? Government datasets from agencies such as the U.S. Census Bureau provide the standard measures of local economic activity at the local level. However, these statistics typically appear only after multi-year lags, and the public-facing versions are aggregated to the county or ZIP code level. In contrast, crowdsourced data from online platforms such as Yelp are often contemporaneous and geographically finer than official government statistics. In this paper, we present evidence that Yelp data can complement government surveys by measuring economic activity in close to real time, at a granular level, and at almost any geographic scale. Changes in the number of businesses and restaurants reviewed on Yelp can predict changes in the number of overall establishments and restaurants in County Business Patterns. An algorithm using contemporaneous and lagged Yelp data can explain 29.2 percent of the residual variance after accounting for lagged CBP data, in a testing sample not used to generate the algorithm. The algorithm is more accurate for denser, wealthier, and more educated ZIP codes.

[1] Byron Perpetua and Louis Maiden provided excellent research assistance. Glaeser thanks the Taubman Center for financial support. Kim and Luca have done consulting for tech companies including Yelp, but their compensation and ability to publish are not tied to the results of this paper. All remaining errors are our own.

[†] Harvard University, eglaeser@harvard.edu

[‡] Harvard Business School, hkim@hbs.edu

[§]Harvard Business School, mluca@hbs.edu

1. Introduction

Public statistics on local economic activity, provided by the Census Bureau's County Business Patterns, the Bureau of Economic Analysis, the Federal Reserve System, and state agencies, provide invaluable guidance to local and national policy-makers. Whereas national statistics, such as Bureau of Labor Statistics' monthly job report, are reported in a timely manner, local data sets are often published only after long lags. They are also aggregated to coarse geographic areas, which impose practical limitations on their value. For example, as of August 2017, the latest available County Business Patterns data was from 2015, aggregated to the ZIP code level, and much of the ZIP code data is suppressed for confidentiality reasons. Similarly, the Bureau of Economic Analysis' metropolitan area statistics have limited value to the leaders of smaller communities within a large metropolitan area.

Data from online platforms such as Yelp, Google, and LinkedIn raise the possibility of enabling researchers and policy-makers to supplement official government statistics with crowd-sourced data at the granular level, provided years before statistics become available. A growing body of research has demonstrated the potential of digital exhaust to predict economic outcomes of interest (e.g. Choi and Varian 2012, Cavallo 2012, Einav and Levin 2014, Kang et al. 2013, Wu and Brynjolfsson 2015, Goel et al 2010, Guzman and Stern 2016). Online data sources also make it possible to measure new outcomes that were never included in traditional data sources (Glaeser et al. 2017).

In this paper, we explore the potential for crowdsourced data from Yelp to measure the local economy. Relative to the existing literature on various forecasting activities, our key contribution is to evaluate whether online data can forecast government statistics that provide traditional measures of economic activity, at geographic scale. Previous related work has been less focused on how predictions perform relative to traditional data sources, especially for core local data sets, like County Business Patterns (Goel et al 2010). We particularly focus on whether Yelp data predicts more accurately in some places than in others.

By the end of 2016, Yelp listed over 3.7 million businesses with 65.4 million recommended reviews.[2] This data is available on a daily basis and with addresses for each business, raising the

[2] Yelp algorithmically classifies reviews, flagging reviews that appear to be fake, biased, unhelpful, or posted by less established users as "not recommended." Recommended reviews represent about three quarters of all reviews, and

possibility of measuring economic activity day-by-day and block-by-block. At the same time, it is a priori unclear whether crowdsourced data will accurately measure the local economy at scale, since changes in the number of businesses reflect both changes in the economy and changes in the popularity of a given platform. Moreover, to the extent that Yelp does have predictive power, it is important to understand the conditions under which Yelp is an accurate guide to the local economy.

To shed light on these questions, we test the ability of Yelp data to predict changes in the number of active businesses as measured by the County Business Patterns. We find that changes in the number of businesses and restaurants reviewed on Yelp can help to predict changes in the number of overall establishments and restaurants in County Business Patterns, and that predictive power increases with ZIP-code level population density, wealth, and education level.

In Section II, we discuss the data. We use the entire set of businesses and reviews on Yelp, which we merged with CBP data on the number of businesses open in a given ZIP code and year. We first assess the completeness of Yelp data relative to County Business Patterns, beginning with the restaurant industry where Yelp has significant coverage. In 2015, CBP listed 542,029 restaurants in 24,790 ZIP codes, and Yelp listed 576,233 restaurants in 22,719 ZIP codes. Yelp includes restaurants without paid employees that may be overlooked by the Census' Business Register. There are 4,355 ZIP codes with restaurants in County Business Patterns that do not have any Yelp restaurants. Similarly, there are 2,284 ZIP codes with Yelp restaurants and no CBP restaurants.

We find that regional variation in Yelp coverage is strongly associated with the underlying variation in population density. There are more Yelp restaurants than CBP restaurants in New York City. Rural areas like New Madison, Ohio have limited Yelp coverage. In 2015, 95% of the U.S. population lived in ZIP codes in which Yelp counted at least 50% of the number of restaurants that CBP recorded. This cross-sectional analysis suggests that Yelp data is likely to be more useful to policy analyses in areas with higher population density.

In Section III, we turn to the predictive power of Yelp for overall ZIP code-level economies across all industries, across all geographies. We look both at restaurants and, more importantly,

the remaining reviews are accessible from a link at the bottom of each business's page but do not factor into a business's overall star rating or review count.

establishments across all industries. Lagged and contemporaneous Yelp measures appear to predict annual changes in CBP's number of establishments, even when controlling for prior CBP measures. We find similar results when restricting the analysis to the restaurant sector.

To assess the overall predictive power of Yelp, we use a random forest algorithm to predict the growth in CBP establishments. We start by predicting the change in CBP establishments with the two lags of changes in CBP establishments, as well as ZIP code and year fixed effects. We then work with the residual quantity. We find that contemporaneous and lagged Yelp data can generate an algorithm that is able to explain 21.4 percent of the variance of residual quantity using an out-of-bag estimate in the training sample, which represents 75 percent of the data. In a testing sample not used to generate the algorithm, our prediction is able to explain 29.2 percent of the variance of this residual quantity.

We repeat this exercise using Yelp and CBP data at the restaurant level. In this case, the basic Yelp prediction is able to explain 21.2 percent of variance out of the training sample, using an out-of-bag estimate. The augmented Yelp prediction can explain 26.4 percent of the variance in the testing sample.

In Section IV, we look at the conditions under which Yelp is most effective at predicting local economic change. First, we examine the interaction between growth in Yelp and characteristics of the locale, including population density and income. We find that Yelp has more predictive power in denser, wealthier, and more educated areas. Second, we examine whether Yelp is more predictive in some industries than others using a regression framework. We find that Yelp is more predictive in retail, leisure, and hospitality industries, as well as professional and business services industries. We then reproduce our random forest approach using geographic and industry sub-groups. Overall, this suggests that Yelp can help to complement more traditional data sources, especially in more urban areas and in industries with better coverage.

Our results highlight the potential for using Yelp data to complement CBP by nowcasting – in other words, by shedding light on recent changes in the local economy that have not yet appeared in official statistics due to long reporting lags. A second potential use of crowdsourced data is to measure the economy at a more granular level than can be done in public facing government statistics. For example, it has the potential to shed light on variation in economic growth within a metropolitan area.

Section V concludes that Yelp data can provide a useful complement to government surveys by measuring economic activity in close to real time, at a granular level, and with data such as prices and reputation that are not contained in government surveys. Yelp's value for nowcasting is greatest in higher density, income, and education areas and in the retail and professional services industry. Data from online platforms such as Yelp are not substitutes for official government statistics. To truly understand the local economy, it would be better to have timelier and geographically fine official data, but as long as that data does not exist, Yelp data can complement government statistics by providing data that are more up to date, granular, and broader in metrics than would otherwise be available.

2. Data

County Business Patterns (CBP) is a program of the Census Bureau that publishes annual statistics for businesses with paid employees within the United States, Puerto Rico, and Island Areas. Statistics include the number of businesses, employment during the week of March 12, first quarter payroll, and annual payroll, and are available by state, county, metropolitan area, ZIP code, and congressional district levels. It has been published annually since 1964, and covers most North American Industry Classification System (NAICS) industries excluding a few categories.[3] CBP's data are extracted from the Business Register, a database of all known single and multi-establishment employer companies maintained by the U.S. Census Bureau; the annual Company Organization Survey; and various Census Bureau Programs including the Economic Census, Annual Survey of Manufactures, and Current Business Surveys. County-level statistics for a given year are available approximately 18 months later, and slightly later for ZIP code-level data.

As an online platform that publishes crowdsourced reviews about local businesses, Yelp provides a quasi-real-time snapshot of retail businesses that are open (see **Figure 1** for a screenshot of the Yelp website). As of spring 2017, Yelp was operating in over 30 countries, with over 127 million reviews written and 84 million unique desktop visitors on a monthly average basis (Yelp 2017). Business listings on Yelp are continually sourced from Yelp's internal team, user

[3] Excluded categories include crop and animal production; rail transportation; National Postal Service; pension, health, welfare, and vacation funds; trusts, estates, and agency accounts; private households; and public administration. CBP also excludes most establishments reporting government employees.

submissions, business owner reports of their own business, and partner acquisitions, and then checked by an internal data quality team. Businesses on Yelp span many categories beyond restaurants, including shopping, home services, beauty, and fitness. Each business listing reports various attributes to the extent that they are available, including location, business category, price level, opening and closure dates, hours, and user ratings and reviews. The data begin in 2004 when Yelp was founded, which enables U.S. business listings to be aggregated at the ZIP code, city, county, state, and country level for any given time period post-2004.

For our analysis, we merge these two sources of data at the ZIP code level from 2004 to 2015. We create two data sets: one on the total number of businesses listed in a given ZIP code and year, and another focusing on the total number of restaurants listed in a given ZIP code and year. For the latter, we use the following NAICS codes to construct the CBP number of restaurants, in order to pull as close a match as possible to Yelp's restaurant category: 722511 (full-service restaurants), 722513 (limited-service restaurants), 722514 (cafeterias, grill buffets, and buffets), and 722515 (snack and nonalcoholic beverage bars).[4]

The resulting data set shows that in 2015, Yelp listed a total number of 1,436,442 U.S. businesses across 25,820 unique ZIP codes, representing approximately 18.7% of CBP's 7,663,938 listings across 38,748 ZIP codes.[5] In terms of restaurants, CBP listed 542,029 restaurants in 24,790 ZIP codes, and Yelp listed 576,233 restaurants in 22,719 ZIP codes, for an overall Yelp coverage of 106.3%. Across the U.S., 33,120 ZIP Code Tabulation Areas (ZCTAs) were reported by the 2010 Census, and over 42,000 ZIP codes are currently reported to exist, some of which encompass non-populated areas.

Yelp data also has limitations that may reduce its ability to provide a meaningful signal of CBP measures. First, while CBP covers nearly all NAICS industries, Yelp focuses on local businesses. Since retail is a small piece of the business landscape, the extent to which Yelp data relates to the overall numbers of CBP businesses or growth rates in other industries depends on the broader relationship between retail and the overall economy. Even a comparison to the restaurant-only CBP data has challenges, as CBP's industry classification is derived from the Economic Census or other Census surveys. In contrast, Yelp's classification is assigned through

[4] Some notable exclusions are 722330 (mobile food services), 722410 (drinking places), and all markets and convenience stores.

[5] These numbers exclude any businesses in Yelp that are missing a ZIP code, price range, or any recommended reviews.

user and business owner reports, as well as Yelp's internal quality check. As a result, some businesses may not be categorized equivalently across the two data sets (e.g. a bar that serves snack food may be classified as a "drinking place" in CBP, while Yelp may classify it as both a bar and a restaurant). Furthermore, Yelp includes restaurants with no employees, while CBP does not count them. Second, the extent of Yelp coverage also depends on the number of Yelp users, which has grown over time as the company has become more popular. In areas with thicker user bases, one might expect business openings and closings to be more quickly reported by users, allowing Yelp to maintain a fairly real-time snapshot of the local economy. However, in areas with low adoption, businesses may take longer to be flagged as closed or open, adding noise to the true number of businesses currently open in the economy. As Section 3 will further discuss, a snapshot of Yelp coverage of CBP businesses and restaurants across the U.S. in 2015 shows that more highly populated areas are more likely to have reliable Yelp data. Third, businesses with no reviews may receive less attention from users – and therefore may be less likely to be flagged as open or marked as closed even after they close, since this relies on user contributions.

To account for these limitations, we only count businesses as open if they have received at least one recommended Yelp review. In the ZIP codes covered by both CBP and Yelp, Yelp's mean and median number of restaurants has steadily increased over the past ten years (see **Figure 2**). This increase reflects steadily increasing Yelp usage. We limit our sample to after 2009, because the mean number of restaurants per ZIP code between CBP and Yelp becomes comparable around 2009. The mean number of restaurants in Yelp actually surpassed the mean number of restaurants in CBP in 2013, which may be explained by differences in accounting such as industry category designations and Yelp counts of businesses with no employees. Finally, we limit our analysis to ZIP codes with at least one business in CBP and Yelp in 2009, and examine a balanced sample of ZIP codes from 2009 to 2015. **Table 1** shows summary statistics of all variables in our data set across this time period.

In the sections that follow, we use this data set to describe Yelp's coverage over time and geography in greater detail, as well as the findings of our analyses.

Comparing Restaurant Coverage on Yelp and County Business Patterns

We first compare Yelp and CBP restaurant numbers to paint a more detailed picture of Yelp coverage across geography. In 2015 (the last year of CBP data available), 27,074 ZIP codes out of 33,120 ZCTAs listed in the U.S. in 2010 had at least one restaurant in either CBP or Yelp.[6] CBP listed 542,029 restaurants in 24,790 ZIP codes, and Yelp listed 576,233 restaurants in 22,719 ZIP codes. There were 2,284 ZIP codes with at least one Yelp restaurant but no CBP restaurants, and 4,355 ZIP codes with at least one CBP restaurant and no Yelp restaurants.

We focus on Yelp coverage ratios, which are defined as the ratio of Yelp restaurants to CBP restaurants. Since we match the data by geography, not by establishment, there is no guarantee that the same establishments are being counted in the two data sources. Nationwide, the Yelp coverage ratio is 106.3%, meaning that Yelp captures more establishments, presumably disproportionately smaller ones, than it misses.[7] Approximately, 95 percent of the population in our sample live in ZIP codes where the number of Yelp restaurants is at least 50% of the number of CBP restaurants, and over 50 percent of the population in our ZIP code sample live in ZIP codes with more Yelp restaurants than CBP restaurants. (see **Figure 3**).

Yelp coverage of CBP restaurants is strongly correlated with population density. In the 1000 most sparsely populated ZIP codes covered by CBP, mean Yelp coverage is 88% (median coverage = 67%), while in the 1000 densest ZIP codes, mean coverage is 126% (median coverage = 123%). **Figure 4** shows the relationship between Yelp coverage of CBP restaurants and population density across all ZIP codes covered by CBP, plotting the average Yelp/CBP ratio for each equal-sized bin of population density. The relationship is at first negative and then positive for population density levels above 50 people per square mile.

The non-monotonicity may simply reflect a non-monotonicity in the share of restaurants with no employees, which in turn reflects offsetting supply and demand side effects. In ZIP codes with fewer than 50 people per square mile, Yelp tends to report one or two restaurants in many of these areas where CBP reports none. Extremely low density levels imply limited restaurant demand, which may only be able to support one or two small establishments. High density levels generate robust demand for both large and small establishments, but higher density areas may also have a disproportionately abundant supply of small-scale, often immigrant entrepreneurs. High

[6] We note that ZCTAs are only revised for the decennial census.

[7] These ratios refer to the total counts of CBP and Yelp restaurants; we can make no claims about whether the two sources are counting the same businesses.

density levels may also have greater Yelp usage, which helps explain the upward sloping part of the curve.

ZIP code 93634 in Lakeshore, California exemplifies low density America. The total population is 33 people, over an area of 1,185 square miles that is mountainous. Yelp lists two restaurants, while CBP lists zero. The two restaurants are associated with a resort that may be counted as part of lodging establishments in CBP. ZIP Code 45346 in New Madison, Ohio is near the threshold of 50 people per square mile. This large rural area includes 42 square miles and a small village with 2,293 people. Both Yelp and CBP track exactly one restaurant, which is a snack shop in the Yelp data. A very dense ZIP code like 10128 in Manhattan, New York City's Upper East Side, with a population of 60,453 in an area of 0.471 square miles, lists 177 Yelp restaurants and 137 CBP restaurants, for a Yelp coverage ratio of 129%. While this neighborhood contains many large eating establishments, it also contains an abundance of smaller eateries, including food trucks, that are unlikely to be included in County Business Patterns.

III. Nowcasting CBP

We now evaluate the potential for Yelp data to provide informative measures of the local economy by exploring its relationship with CBP measures, first using regression analysis and then turning to a more flexible forecasting exercise.

Regression Analysis

Table 2 shows results from regressing changes in CBP business numbers on prior CBP and Yelp measures. Column (1) regresses changes in CBP's number of businesses in year *t* on two lags of CBP. The addition of one CBP establishment in the previous year is associated with an increase in 0.3 businesses in year *t*, showing that there is positive serial correlation in the growth of businesses at the ZIP code level. The correlation is also strongly positive with a two-year lag of CBP business openings. Together, the two lags of changes in CBP establishments explain 14.8% of the variance (as measured by adjusted r-squared).

Column 2 of **Table 2** regresses changes in CBP business numbers in year *t* on two lags of CBP and the contemporaneous change in Yelp business numbers. Adding contemporaneous Yelp business numbers increases the variance explained to 22.5%. A one-unit change in the number of Yelp businesses in the same year is associated with an increase in the number of CBP businesses of six-tenths. This coefficient is fairly precisely estimated, so that with 99 percent confidence, a one unit increase in the number of Yelp establishments is associated with between .55 and .66 CBP establishments in the same year, holding two years of lagged CBP establishment growth constant.

The prediction of a purely accounting model of establishments is that the coefficient should equal one, but there are at least two reasons why that prediction will fail. First, if there is measurement error in the Yelp variable, that will push the coefficient below one due to attenuation bias. Second, Yelp does not include many CBP establishments, especially in industries other than retail. If growth in retail is associated with growth in other industries, then the coefficient could be greater than one, which we term spillover bias and expect to be positive. The estimated coefficient of .61 presumably reflects a combination of attenuation and spillover bias, with spillover bias dominating.

Columns 3 and 4 show that lagged Yelp data, as well as other Yelp variables including the number of closures and reviews, are only mildly informative in explaining the variance of CBP business number growth. Growth in CBP establishments is positively associated with one-year lag in the growth in the number of Yelp establishments, and including that variable causes the coefficient on contemporary establishment growth to drop to .44. Regression (4) also shows that increases in the number of Yelp closings is negatively correlated with growth in the number of CBP establishments, and that the number of Yelp reviews is not correlated with growth in the number of CBP establishments. Some of these extra Yelp variables are statistically significant, but they added little to overall explanatory power. The adjusted r-squared only rises from .225 to .229 between regression (2) and regression (4). The real improvement in predictive power comes from the inclusion of contemporaneous Yelp openings, not from the more complex specification. This suggests that simply looking at current changes in the number of Yelp establishments may be enough for most local policy-makers who are interested in assessing the current economic path of a neighborhood.

Table 3 replicates the above analysis for changes in the number of restaurants in a given ZIP code and year. The first specification suggests that there is little serial correlation in CBP restaurant openings, and consequently, past changes in CBP do little to predict current changes. The second regression shows a strong correlation between changes in the number of CBP restaurant openings and contemporaneous Yelp restaurant openings. The r-squared of .11 is lower in this specification than in the comparable regression (2) in **Table 2** (.23), but this is perhaps unsurprising given the much lower baseline r-squared. The improvement in r-squared from adding contemporaneous Yelp data in the restaurant predictions is larger both in absolute and relative terms.

Perhaps more oddly, the coefficient on Yelp openings is .32, which is smaller for the restaurant data than for overall data. We would perhaps expect the measurement bias problem to be smaller for this industrial sub-group, and that would presumably lead us to expect a larger coefficient in **Table 3**. The exclusion of other industries, however, reduces the scope for spillover bias, which probably explains the lower coefficient. This shift implies that both attenuation and spillover biases are likely to be large, which pushes against any structural interpretation of the coefficient.

Regression (3) includes a one-year lag of Yelp openings, which also has a positive coefficient. Including this lag causes the coefficient on lagged CBP openings to become even more negative. One explanation for this shift could be that actual restaurant openings display mean reversion, but restaurants appear in Yelp before they appear in County Business Patterns. Consequently, last year's growth in Yelp restaurants predicts this year's growth in CBP restaurants. Including this lag improves the r-squared to .123.

In regression (4), we also include our measure of closures in the Yelp data and the number of Yelp reviews. The coefficients on both variables are statistically significant and both have the expected signs. More Yelp closures are associated with less growth in CBP establishments. More Yelp reviews imply more restaurant openings, perhaps because more reviews are associated with more demand for restaurants. Including these extra variables improves the r-squared to .139. These regressions suggest that there is more advantage in using a more complicated Yelp-based model to assess the time series of restaurants than to assess the overall changes in the number of establishments.

While these results suggest that Yelp data has the potential to serve as a useful complement to official data sources, these regression analyses are hardly a comparison of best possible predictors. To provide a more robust evaluation of the potential for Yelp data to provide informative measures of the local economy, we now turn to out-of-sample forecasting of CBP measures using a more sophisticated prediction algorithm.

Forecasting with A Random Forest Algorithm

We leverage a random forest algorithm to evaluate whether Yelp measures can provide gains in nowcasting CBP measures before the release of official statistics. We are interested in the ability of Yelp to predict changes in overall CBP establishments and restaurants over and above the prediction power generated by lagged CBP data. Consequently, we begin our prediction task by regressing the change in CBP establishments on the two lags of changes in CBP establishments and ZIP code and year fixed effects. We then work with the residual quantity. Given the two lags of CBP, our sample spans years 2012 to 2015. We use a relatively simple first stage regression because we have a limited number of years, and because modest increases in complexity add little predictive power.

We assign the last year of our data set (2015) to the test set, which represents 25% of our sample, and the rest to the training set. We then examine the ability of lagged and contemporaneous Yelp data to predict residual changes in CBP number of establishments in a given year and ZIP code. We include the following Yelp measures in the feature set: contemporaneous and lagged changes in, and absolute count of, the total number of open, opened, and closed businesses, aggregate review count, and the average rating of businesses, all in terms of total numbers and broken down by lowest and highest price level, along with year and the total number of businesses that closed within one year. The number of trees in the forest is set to 300, and the gains to increasing this number are marginal, yielding very similar results. Using an off-the-shelf random forest algorithm on models with limited feature sets, our analyses represent basic exercises to evaluate the usefulness of Yelp data, rather than to provide the most precise forecasts.

Table 4 shows prediction results. The first column shows our results for CBP establishments overall. The second column shows the results for restaurants. We evaluate the predictive power of our model in two ways. Using the 2012-2014 data, we can use an “out-of-bag” estimate of the prediction accuracy. We also use the 2015 data as a distinct testing sample.

The first row shows that the model has an r-squared of .29 for predicting the 2014-2015 CBP openings for all businesses and an r-squared of .26 for restaurants. Since the baseline data had already orthogonalized with respect to year, this implies that the Yelp-based model can explain between one-quarter and one-third of the variation across ZIP code in the residualized CBP data.

The second row shows the out-of-bag estimates of r-squared, based on the training data. In this case, the r-squared is .21 for both data samples. The lower r-squared is not surprising given that out-of-bag estimates can often understate the predictive power of models. Nonetheless, it is useful to know that the fit of the model is not particular to anything about 2015.

There appears to be a wide range of predictive ability – but on average bounded within approximately half a standard deviation for businesses, with 8.0 mean absolute error (MAE) and 3.9 median absolute error, compared to a mean of 3.4 and a standard deviation of 15.1. The mean and median absolute errors for restaurants are substantially smaller than for businesses, at 1.7 and 1.1, respectively, but the mean and standard deviation for restaurant growth are also substantially lower than for businesses, at .5 and 2.9, respectively.

Yelp's predictive power is far from perfect, but it does provide significant improvement in our knowledge about the path of local economies. Adding Yelp data can help marginally improve predictions compared to using only prior CBP data.

IV. The Limits to Nowcasting by Geographic Area and Industry

We now examine where Yelp data is better or worse at predicting local economic change, looking across geographic traits and industry categories. As discussed earlier, we believe that Yelp is likely to be more accurate when population densities are higher and when Yelp use is more frequent. We are less sure why Yelp should have more predictive power in some industries than in others, but we still test for that possibility. We first use a regression framework to examine the interaction between Yelp changes and local economic statistics on population density, median household income, and education. We then run separate regression analyses by industry categories. Finally, we reproduce our random forest approach for geographic and industrial sub-groups.

Table 5: Interactions with Area Attributes

Table 5 shows results from regressions where changes in Yelp's open business numbers are interacted with indicators for geographic characteristics. We use indicator variables that take on a value of one if the area has greater than the median level of population density, education, and income, and zero otherwise. Population density estimates are from the 2010 Census, while measures of median household income and percent with a Bachelor's degree are from the 2015 American Community Survey 5-year estimates. We present results just for total establishments, and begin with the simple specification of regression (2) in **Table 2**.

In this first regression, we find that all three interactions terms are positive and statistically significant. The interaction with high population density is .14. The interaction with high income is .30. The interaction with high education is .09. Together, these interactions imply that the coefficient on contemporaneous Yelp openings is .2 in a low density, low education and low income ZIP code, and .73 in a high density, high education, and high income ZIP code. This is an extremely large shift in coefficient size, perhaps best explained by far greater usage of Yelp in places with more density, education and income. If higher usage leads to more accuracy, this should cause the attenuation bias to fall and the estimated coefficient to increase.

In the second regression, we also add lagged Yelp openings. In this case, the baseline coefficient is negative, but again all three interactions are positive. Consequently, the estimated coefficient on lagged Yelp openings is -.1 in low density, low income, low education locales, but .24 in high density, high income, high education areas. Again, decreased attenuation bias is one possible interpretation of this change. The third regression includes changes in Yelp closings and the number of Yelp reviews.

These interactions suggest that the predictive power of Yelp is likely to be higher in places with more density, education and income. However, it is not true that adding interactions significantly improves the overall r-squared. There is also little increase in r-squared from adding the lag of Yelp openings or the other Yelp variables, just as in **Table 2**. While contemporaneous Yelp openings is the primary source of explanatory power, if policy-makers want to use Yelp openings to predict changes in establishments, they should recognize that the mapping between contemporaneous Yelp openings and CBP openings is different in different places.

Table 6: The Predictive Power of Yelp and Area Attributes

Table 5 examined how the coefficient on Yelp openings changed with area attributes. **Table 6** examines whether the predictive power of Yelp differs with the same attributes. To test this hypothesis, we replicate **Table 4** on different subsamples of the data. We split the data into two groups based on first density, then income, and then education. The split is taken at the sample median. For each split, we replicate our previous analysis using a random forest algorithm. Once again, we omit the 2015 data in our training sample and use that data to test the model's predictive power.

The first panel of **Table 6** shows the split based on density. Our two primary measures of goodness of fit are the r-squared for 2014-2015 CBP openings and the out-of-bag r-squared estimated for the earlier data. In the high-density sample, the r-squared for the out-of-sample data is .24. In the low-density sample, the r-squared is .06. The out-of-bag r-squared is .19 in the high-density sample and .03 in the low-density sample. As the earlier interactions suggest, Yelp openings have far more predictive power in high-density ZIP codes than in low-density ZIP codes. One natural interpretation of this finding is that there is much more Yelp usage in higher density areas, and, consequently, Yelp provides a more accurate picture of the local economy when density is high.

The mean and median absolute errors are higher in high-density ZIP codes than in low-density ZIP codes. Yet, the mean and standard deviation of CBP establishment growth are also much higher in such areas. Relative to the mean and standard deviation of CBP openings, the standard errors are smaller in higher density locations. The mean and median absolute errors are 12.7 and 8.0 in the high-density sample, compared to a mean CBP growth of 7.0 and standard deviation of 20.5. In low-density locations, the mean and median absolute errors are 3.9 and 2.5, compared to a mean CBP growth of .5 with a 6.5 standard deviation.

In the second panel, we split based on income. In the higher income sample, the r-squared for 2014-2015 data is .33 and the out-of-bag r-squared is .26. In the lower income sample, the r-squared for the later data is .15 and the out-of-bag r-squared is .08. Once again, in higher income areas where Yelp usage is more common, Yelp provides better predictions. In higher income areas, the median absolute error (5.1) is lower than the mean CBP growth (6.1), compared to lower income areas where the median absolute error at 3.5 is two and half times the mean CBP growth of 1.4.

In the final panel, we split based on education and the results are again similar. The r-squared using the 2014-2015 data is .29 in the high education sample and .06 in the low education sample. The out-of-bag r-squared is .23 in the high education sample and .03 in the low education sample. Similar to the density split, the mean and median absolute errors are much higher in high education ZIP codes than in low education ZIP codes, but smaller relative to the mean and standard deviation of CBP establishment growth. The median absolute error in high education ZIP codes is 6.0, slightly lower than the mean CBP growth of 6.5 and approximately a third of the standard deviation of CBP growth (19.1). In low education ZIP codes, the median absolute error is 3.0, more than three times the mean CBP growth (.9) and approximately a third of the standard deviation (10.2).

Table 6 shows that the predictive power of Yelp is much lower in lower education or lower density locations. Yelp does a bit better in lower income areas. Yelp is more effective at predicting the local economy when education, density and income is high. This suggests that using Yelp to understand the local economy makes more sense in richer coastal cities, than in poorer places.

Yelp appears to complement income, education, and population density, perhaps because higher density areas have more restaurant options. Consequently, Yelp is just a better source for data in these areas and may be able to do more to improve local policy-making. This provides yet another example of a setting where new technology favors areas with initial advantages.

Tables 7, 8 and 9: Cross Industry Variation

We now examine whether Yelp is more predictive in some industries than others. We define industry categories loosely based on NAICS supersectors, creating six industry categories described in **Table 7**. These sectors include "retail, leisure and hospitality," which is the sector that has the most overlap with Yelp coverage, "goods production," "transportation and wholesale trade," "information and financial activities," "professional and business services," and "public services."

We expect that Yelp's predictive power will be higher in those industries where Yelp has more coverage. Yelp covers local restaurants and services businesses, including hospitality, real estate, home services, and automotive repair, as well as local landmarks including museums and religious buildings. These industries mostly fall into two of our industry categories – retail, leisure, and hospitality and professional and business services, with real estate and leasing falling into the information and financial activities category.

For each industrial supersector, we regress changes in CBP business numbers in year *t* on two lags of CBP in that industry group, contemporaneous and lagged changes in Yelp business numbers, and changes in business closures and aggregate review counts in Yelp. We include the CBP lags in each specific industry, but we do not try to distinguish Yelp listings by industry, primarily because Yelp coverage in most of the industries is modest.

The first regression in **Table 8** shows that the coefficients for the retail, leisure, and hospitality industries are relatively large. A one-unit contemporaneous change in the number of Yelp businesses is associated with a .21 change in the number of CBP businesses in that sector. The coefficients on Yelp closings and total Yelp reviews are also significant. As in **Table 3**, lagged CBP establishment openings are statistically insignificant in this sector.

The coefficient on contemporary Yelp openings for all of the other five industrial supersectors can essentially be grouped into two sets. For professional and business services and for information and finance, the coefficient is close to .1, and the other Yelp variables are strongly significant as well. For the other three supersectors, the coefficient on the Yelp variables is much smaller. The r-squared mirrors the coefficient sizes. In retail, leisure, and hospitality and professional and business services categories, we can explain 8.5 to 10.2 percent of the variation in CBP measures using lagged CBP and Yelp data, compared to 0.9 to 8.2 percent in the other industry categories. These results suggest that Yelp is most likely to be useful for retail and professional services industries and less likely for public services, goods manufacturing or transportation and wholesale trade.

Finally, **Table 9** replicates our random forest approach for each of the industrial supersectors. Again, we follow the same two stage structure of first orthogonalizing with respect to year, ZIP code, and past CBP changes. We again exclude the 2014-2015 CBP data from the training data. We again calculate both the out-of-sample r-squared for that later year and we calculate the out-of-bag r-squared based on earlier data.

The cross-industry pattern here is similar to the pattern seen in the regressions. Yelp has the greatest predictive power for hospitality and leisure, professional and business services, and information and finance. Among this group, however, Yelp data has the greatest ability to predict movement in professional and business services, perhaps because that sector is less volatile than restaurants. In this group, the r-squared for 2014-2015 data ranges from .11 for information and finance to .17 for professional and business services. The out-of-bag r-squared values range from .08 to .16.

Goods production and public services show less predictability from Yelp data. The 2014-2015 r-squared for both these two groups is approximately .07. The out-of-bag r-squared is less than .01 for goods production and .03 for public services. Finally, Yelp shows little ability to predict transportation and wholesale trade.

Our overall conclusion from this exercise is that Yelp does better at predicting overall changes in the number of establishments than in predicting changes within any one industry. The safest industries to focus on relatively fall either to hospitality or to business services. For manufacturing and wholesale trade, Yelp does not seem to offer much predictive power.

V. Conclusion

Recent years have witnessed ongoing discussions about how to update or replace the national census across many countries. For example, the United Kingdom considered replacing the census with administrative data as well as third-party data from search engines like Google (Hope 2010, Sanghani 2013). One of the areas that the U.S. Census Bureau has been considering in its new plan to pare $5.2 billion dollars from its cost of $20 billion for the decennial census is to utilize administrative records and third-party data (U.S. Census Bureau 2015a, Mervis 2017).

Our analyses of one possible data source, Yelp, suggests that these new data sources can be a useful complement to official government data. Yelp can help predict contemporaneous changes in the local economy, and also provide a snapshot of economic change at the local level. It thus provides a useful addition to the data tools that local policy-makers can access.

In particular, we see two main ways in which new data sources like Yelp may potentially help improve official business statistics. First, they can improve forecasting at the margin for official Census products such as the County Business Patterns and the Business Dynamics Statistics that measure the number of businesses. While these products provide invaluable guidance across the economy, there can be a considerable lag in their getting information about new businesses and business deaths. Data sources like Yelp may be able to help identify these events earlier, or provide a basis for making real-time adjustments to the statistics. Second, these data sources can help provide a cross-check for the microdata underlying these statistics products and help reconcile missing or inconsistent data. For example, it may take the Census time to classify businesses correctly, especially for small and new businesses that they under-sample due to respondent burden, and new data sources can provide a source of validation.

Yet, our analysis also highlights the challenges with the idea of replacing the Census altogether at any point in the near future. Government statistical agencies invest heavily in developing relatively complete coverage, for a wide set of metrics. The variation in coverage inherent in data from online platforms make it difficult to replace the role of providing official statistics that government data sources play.

Ultimately, data from platforms like Yelp –combined with official government statistics – can provide valuable complementary datasets that will ultimately allow for more timely and granular forecasts and policy analyses, with a wider set of variables and more complete view of the local economy.

References

Cavallo, Alberto (2012). “Scraped Data and Sticky Prices.” MIT Sloan Working Paper.

Bureau of Labor Statistics (2010). “Mission Statement.” Accessed June 10, 2017. https://www.bls.gov/bls/blsmissn.htm

Census Bureau (2017). “About the Bureau.” Accessed June 10, 2017. https://www.census.gov/about/what.html

Choi, Hyunyoung, and Hal Varian (2012). “Predicting the Present with Google Trends.” Economic Record 88:2–9

Einav, Liran, and Jon Levin (2014). "The Data Revolution and Economic Analysis," *Innovation Policy and the Economy* 14: https://doi.org/10.1086/674019

Glaeser, Edward L., Scott Duke Kominers, Michael Luca, and Nikhil Naik (2017). "Big Data and Big Cities: The Promises and Limitations of Improved Measures of Urban Life." Economic Inquiry (forthcoming).

Goel, Sharad, Jake M. Hofman, Sebastien Lahaie, David M. Pennock, and Duncan J. Watts (2010). “Predicting consumer behavior with Web search.” PNAS 107(41): 17486-17490.

Guzman, Jorge, and Scott Stern (2016). “Nowcasting and Placecasting Entrepreneurial Quality and Performance.” Working Paper.

Hope, Christopher (2010). “National Census to be axed after 200 years.” The Telegraph, July 9, 2010. <http://www.telegraph.co.uk/news/politics/7882774/National-census-to-be-axed-after-200-years.html> Accessed July 6, 2017.

Mervis, Jeffrey (2017). “Scientists fear pending attack on federal statistics collection.” Science Magazine, January 3, 2017. <http://www.sciencemag.org/news/2017/01/scientists-fear-pending-attack-federal-statistics-collection> Accessed July 6, 2017.

Owens, Brian (2015). “Canada reinstates mandatory census, to delight of social scientists.” Science Magazine, November 5, 2015. <http://www.sciencemag.org/news/2015/11/updated-canada-reinstates-mandatory-census-delight-social-scientists> Accessed July 6, 2017.

Sanghani, Radhika (2013). “Google could replace national census.” The Telegraph, June 26, 2013. <http://www.telegraph.co.uk/technology/google/10142641/Google-could-replace-national-census.html> Accessed July 6, 2017.

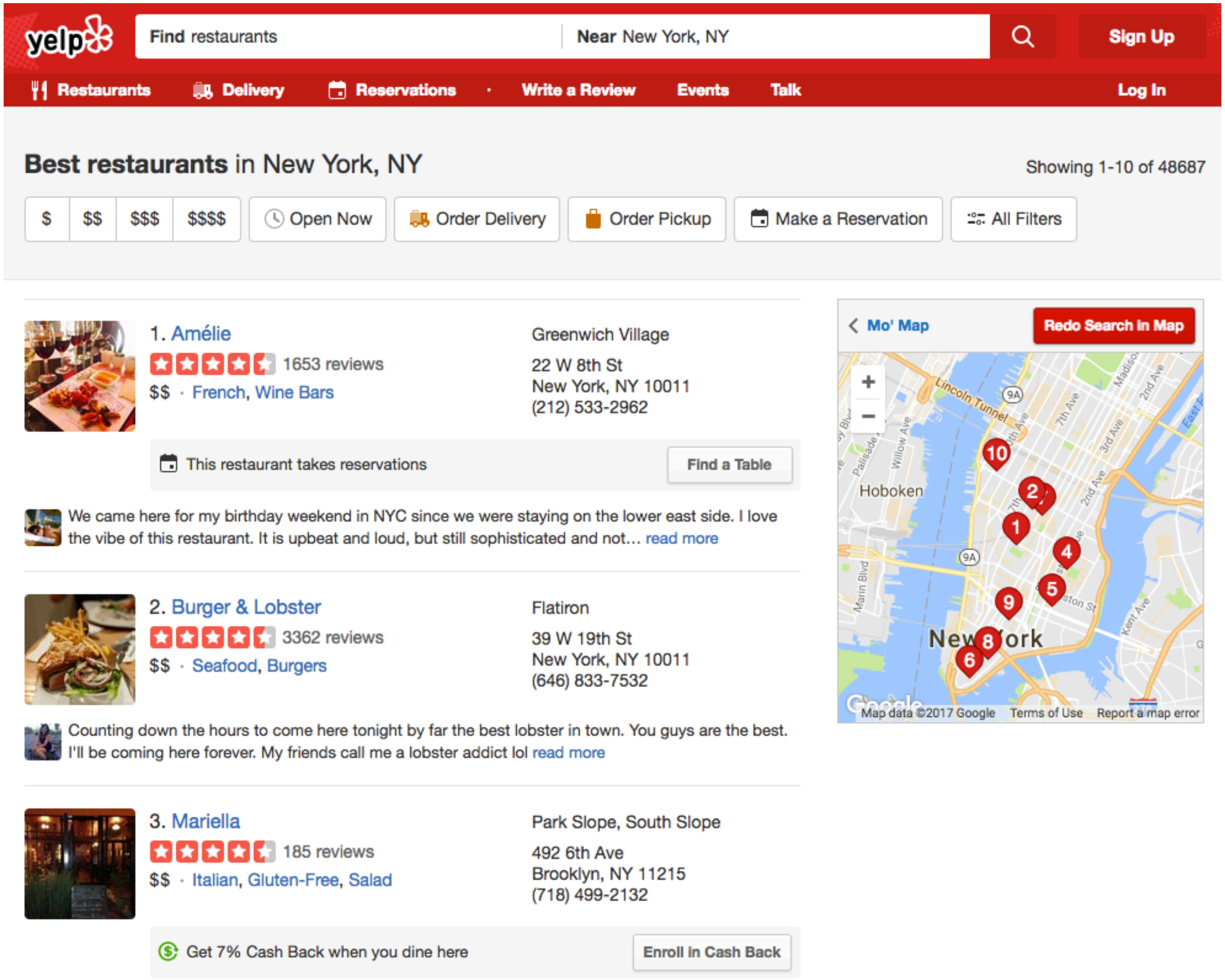
U.S. Census Bureau (2015a). “2020 Census Operational Plan Overview and Operational Areas.” <https://censusproject.files.wordpress.com/2015/12/2020-census-opplan-conference-call_the-census-project_10-21-15_final-1.pdf> Accessed July 6, 2017.

U.S. Census Bureau (2015b). "Potential data sources to replace or enhance the question on condominium status on the American Community Survey." <https://www.census.gov/content/dam/Census/library/working-papers/2015/acs/2015_Flanagan_Doyle_01.pdf> Accessed July 6, 2017.

Wu, Lin, and Erik Brynjolfsson (2015). "The Future of Prediction: How Google Searches Foreshadow Housing Prices and Sales" in Economic Analysis of the Digital Economy, eds. Avi Goldfarb, Shane M. Greenstein, and Catherine E. Tucker. Chicago: University of Chicago Press, 2015.

Yelp (2017). News Release, "Yelp to Participate in the J.P. Morgan Global Technology, Media and Telecom Conference," May 24, 2017. Accessed June 10, 2017. http://www.yelp-ir.com/phoenix.zhtml?c=250809&p=irol-newsArticle&ID=2272465

Figure 1 Example of a Yelp Restaurant Listing



Best restaurants in New York, NY

Showing 1-10 of 48687

1. Amélie

1653 reviews

$$ · French, Wine Bars

Greenwich Village

22 W 8th St
New York, NY 10011
(212) 533-2962

This restaurant takes reservations

We came here for my birthday weekend in NYC since we were staying on the lower east side. I love the vibe of this restaurant. It is upbeat and loud, but still sophisticated and not… read more

2. Burger & Lobster

3362 reviews

$$ · Seafood, Burgers

Flatiron

39 W 19th St
New York, NY 10011
(646) 833-7532

Counting down the hours to come here tonight by far the best lobster in town. You guys are the best. I'll be coming here forever. My friends call me a lobster addict lol read more

3. Mariella

185 reviews

$$ · Italian, Gluten-Free, Salad

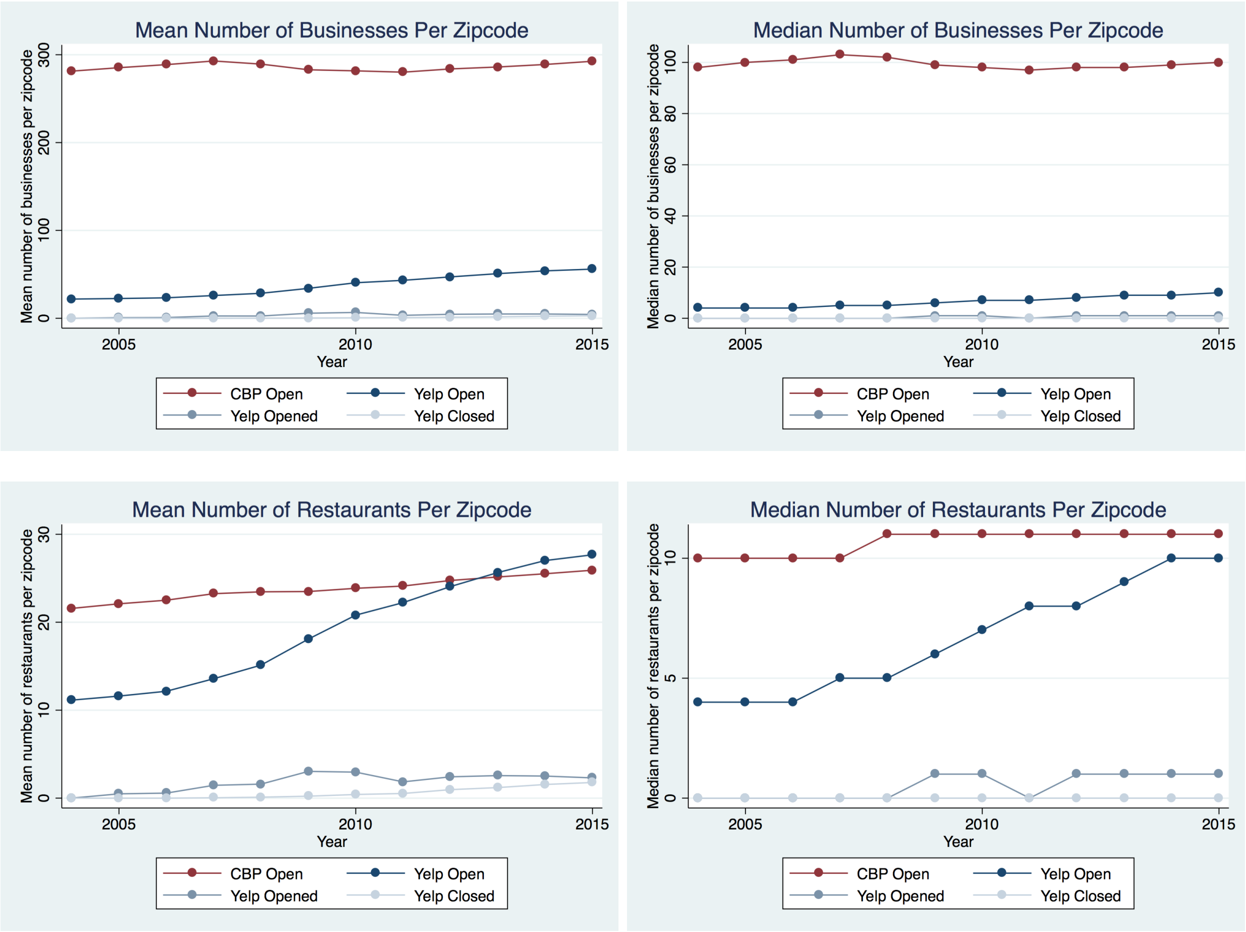
Park Slope, South Slope

492 6th Ave
Brooklyn, NY 11215
(718) 499-2132

Get 7% Cash Back when you dine here

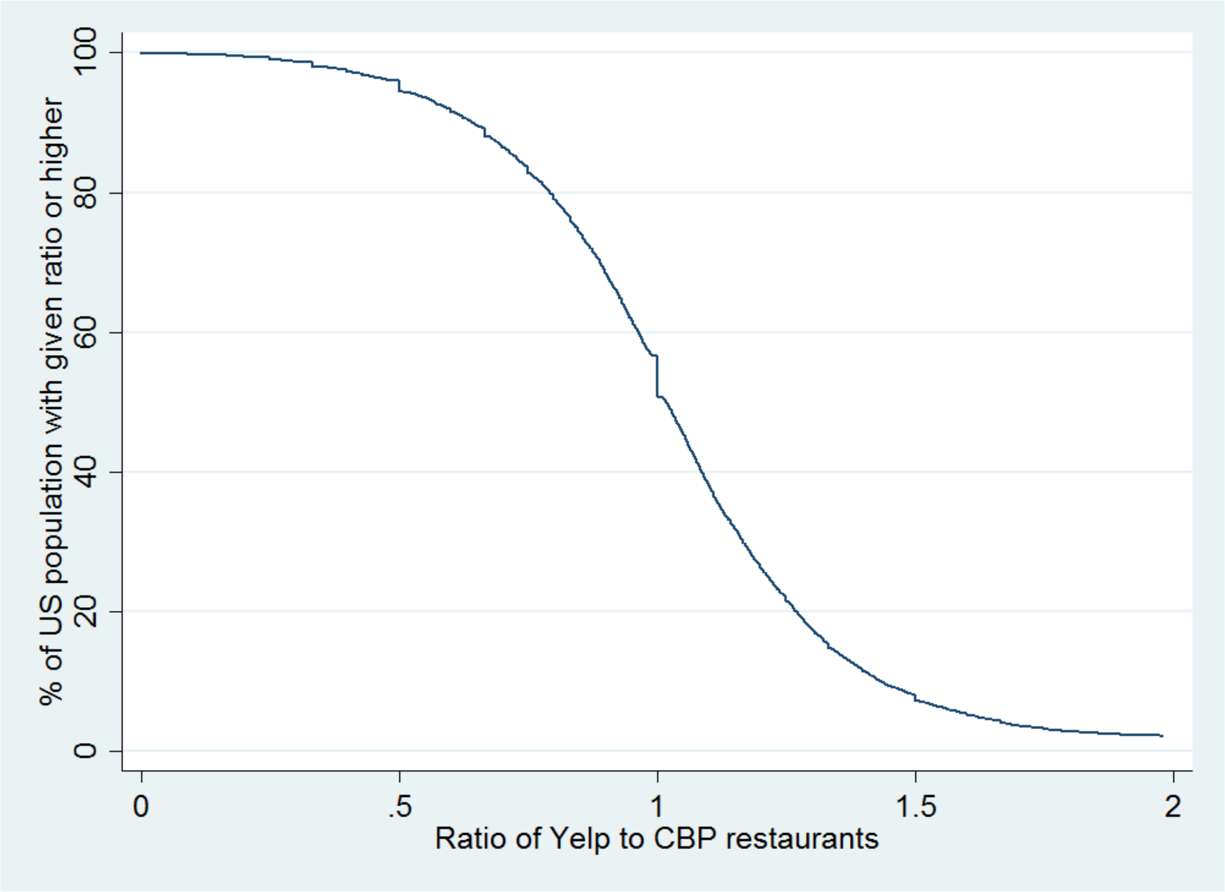
This figure shows a screenshot of a search of restaurants in New York, NY on the Yelp platform.

Figure 2 Number of Businesses and Restaurants Recorded by CBP vs. Yelp 2004-2015



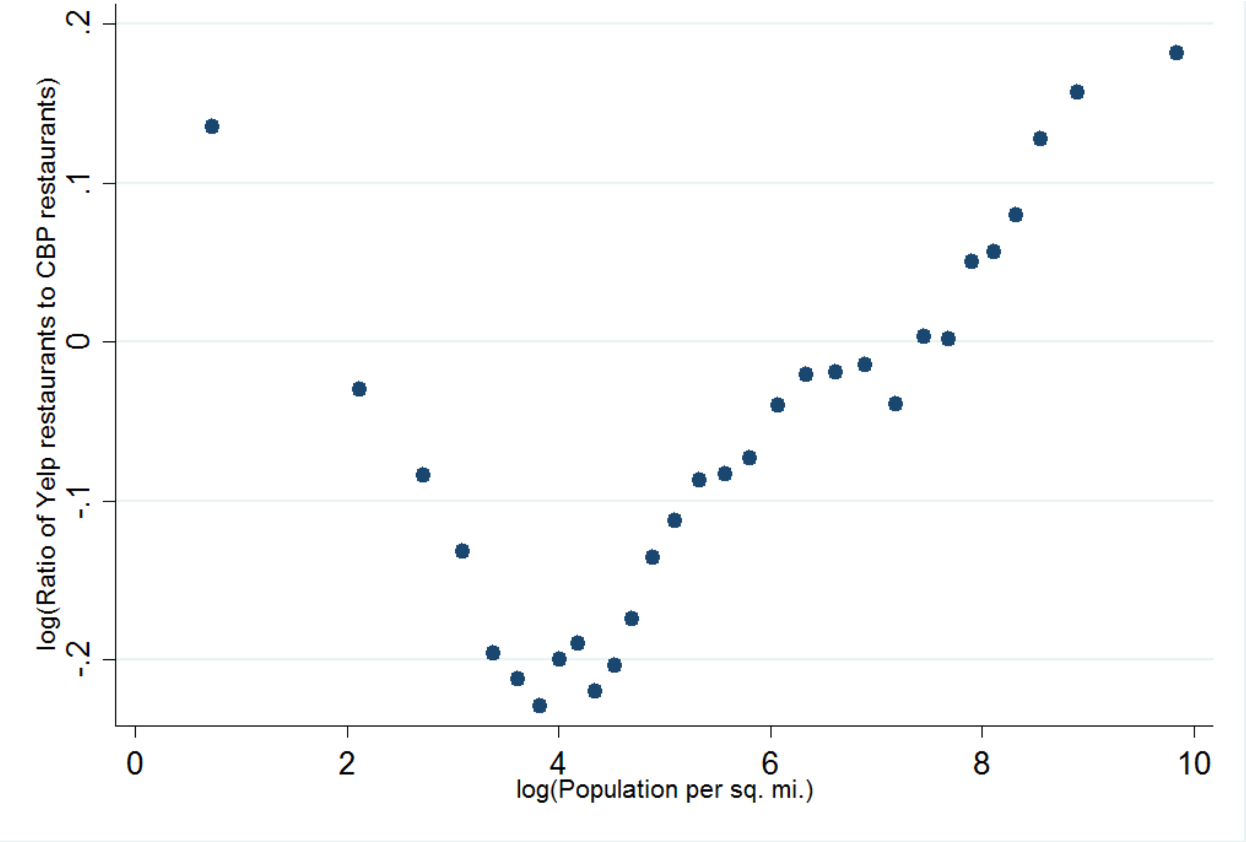
These figures compare the mean and median number of businesses (top) and restaurants (bottom) per ZIP code as recorded by Yelp and CBP between 2004 (when Yelp was founded) to 2015, in all ZIP codes covered by both sources. Yelp Opened shows the mean and median number of restaurants opened that year per ZIP code as recorded by Yelp. Yelp Closed represents the mean and median number of restaurants closed that year per ZIP code as recorded by Yelp.

Figure 3 Distribution of Yelp Coverage Across ZIP codes (Weighted by Population)



This figure shows the cumulative density function of Yelp coverage weighted by population, across all ZIP codes that CBP covers. For each ratio of Yelp to CBP restaurants, this figure shows the percentage of ZIP codes that has that ratio or higher. This figure has been truncated at Yelp/CBP ratio = 2.

Figure 4 Yelp Coverage by Population Density



This figure shows the conditional expectation function of the ratio of Yelp to CBP restaurants on population density across all ZIP codes covered by CBP, plotting the average Yelp/CBP ratio for each equal-sized bin of population density.

Table 1 Summary Statistics

	Businesses		Restaurants	
	Number	*Annual Growth*	*Number*	*Annual Growth*
CBP Number of Open Establishments	317.920	1.717	27.723	0.484
	(432.933)	(14.503)	(34.026)	(2.852)
Yelp Number of Open Businesses	52.274	4.071	26.679	1.811
	(99.450)	(9.159)	(38.880)	(3.571)
Yelp Number of Closed Businesses	1.534	0.476	1.076	0.294
	(4.878)	(2.221)	(2.745)	(1.622)
Number of Yelp Reviews	272.051	69.266	247.470	63.386
	(1218.273)	(260.433)	(984.581)	(214.393)
Average Yelp Rating	3.000	0.162	3.104	0.144
	(1.547)	(1.560)	(1.350)	(1.405)
Yelp Number of Businesses that Closed Within 1 Year	0.038	-0.268	0.032	-0.140
	(0.235)	(8.157)	(0.204)	(3.386)
Yelp Number of Opened Businesses	5.497	0.012	2.831	0.010
	(11.697)	(0.271)	(4.831)	(0.252)
Observations	*159369*	*136602*	*127176*	*109008*
Population Density per Sq. Mile	1756.609		2034.598	
	(5634.997)		(6035.183)	
% Bachelor's Degree or Higher	26.556		27.686	
	(16.249)		(16.438)	
Median Household Income in Past 12 Months (in 2015 dollars)	56533.953		57271.358	
	(23725.879)		(24219.673)	
Observations	*145425*		*122976*	

Means and standard deviations (in parentheses) are displayed for each variable, for absolute numbers and annual changes of both businesses and restaurants. Each observation is at the ZIP code – year level, across years 2009-2015. Population Density estimates are from the 2010 Census. Percent with a Bachelor's Degree or Higher and Median Household Income are from the 2015 American Community Survey 5-year estimates.

Table 2 Predicting CBP Establishment Growth Using Regression Analysis

	(1)	(2)	(3)	(4)
	CBP Establishment Growth	CBP Establishment Growth	CBP Establishment Growth	CBP Establishment Growth
CBP Establishment Growth (lag1)	0.271^{***}	0.197^{***}	0.189^{***}	0.188^{***}
	(0.018)	(0.017)	(0.017)	(0.017)
CBP Establishment Growth (lag2)	0.219^{***}	0.190^{***}	0.185^{***}	0.184^{***}
	(0.010)	(0.011)	(0.011)	(0.011)
Yelp Business Growth		0.605^{***}	0.443^{***}	0.495^{***}
		(0.023)	(0.029)	(0.029)
Yelp Business Growth (lag1)			0.194^{***}	0.169^{***}
			(0.025)	(0.025)
Yelp Growth in Closed Businesses				-0.264^{***}
				(0.048)
Yelp Reviews Growth (divided by 100)				0.094
				(0.081)
Constant	4.542^{***}	1.782^{***}	1.854^{***}	1.822^{***}
	(0.127)	(0.148)	(0.149)	(0.144)
Year FE	Yes	Yes	Yes	Yes
Observations	91068	91068	91068	91068
Adjusted R^2	0.148	0.225	0.228	0.229

All regressions include a full set of calendar year dummies and cluster standard errors at the ZIP Code level. * p<0.10, ** p<0.05, *** p<0.01.

Table 3 Predicting CBP Restaurant Growth Using Regression Analysis

	(1)	(2)	(3)	(4)
	CBP Restaurant Growth	CBP Restaurant Growth	CBP Restaurant Growth	CBP Restaurant Growth
CBP Restaurant Growth (lag1)	-0.049***	-0.127***	-0.157***	-0.165***
	(0.010)	(0.009)	(0.009)	(0.009)
CBP Restaurant Growth (lag2)	0.059***	-0.012	-0.034***	-0.048***
	(0.008)	(0.007)	(0.007)	(0.007)
Yelp Restaurant Growth		0.319***	0.257***	0.274***
		(0.008)	(0.008)	(0.009)
Yelp Restaurant Growth (lag1)			0.132***	0.088***
			(0.009)	(0.009)
Yelp Growth in Closed Restaurants				-0.119***
				(0.013)
Yelp Reviews Growth (divided by 100)				0.164***
				(0.020)
Constant	0.783***	0.160***	0.099***	0.166***
	(0.025)	(0.024)	(0.025)	(0.024)
Year FE	Yes	Yes	Yes	Yes
Observations	72672	72672	72672	72672
Adjusted R^2	0.009	0.110	0.123	0.139

All regressions include a full set of calendar year dummies and cluster standard errors at the ZIP Code level. * p<0.10, ** p<0.05, *** p<0.01.

Table 4 Predicting CBP Establishment and Restaurant Growth Using Random Forest

	Establishments	Restaurants
R-squared	0.292	0.264
Out-of-bag R-squared	0.214	0.212
Mean Absolute Error	7.989	1.713
Mean Squared Error	222.067	7.200
Median Absolute Error	3.871	1.062
Mean CBP Growth	*3.393*	*0.539*
St. Dev CBP Growth	*15.078*	*2.913*
Observations	*91068*	*72672*

All analyses predict residual variance in the change in CBP establishments after regressing two lags of changes in CBP establishments with ZIP code and year fixed effects. Features include year and the change in and absolute number of total open, opened, and closed businesses as recorded by Yelp, aggregate review count, and average rating, and broken down by lowest and highest business price level. The sample covers the time period 2012-2015, and all observations for 2015 are assigned to the test set, and the rest to training. The number of trees in the forest is set to 300. The number of observations, means and standard deviations of CBP Growth are reported using the full set of observations across both training and test sets.

Table 5 Predicting CBP Establishment Growth by Area Attributes Using Regression Analysis

	(1)	(2)	(3)
	CBP Establishment Growth	CBP Establishment Growth	CBP Establishment Growth
CBP Establishment Growth (lag1)	0.188***	0.179***	0.179***
	(0.018)	(0.018)	(0.017)
CBP Establishment Growth (lag2)	0.182***	0.177***	0.175***
	(0.011)	(0.011)	(0.011)
Yelp Business Growth	0.195***	0.302***	0.339***
	(0.047)	(0.060)	(0.060)
High Density * Yelp Business Growth	0.144**	0.016	0.021
	(0.047)	(0.065)	(0.065)
High Income * Yelp Business Growth	0.295***	0.222**	0.224**
	(0.037)	(0.072)	(0.072)
High Education * Yelp Business Growth	0.092**	-0.022	-0.004
	(0.035)	(0.068)	(0.067)
Yelp Business Growth (lag1)		-0.106*	-0.112*
		(0.047)	(0.047)
High Density * Yelp Business Growth (lag1)		0.139**	0.136**
		(0.047)	(0.047)
High Income * Yelp Business Growth (lag1)		0.086	0.084
		(0.073)	(0.073)
High Education * Yelp Business Growth (lag1)		0.125*	0.115
		(0.062)	(0.061)
Yelp Growth in Closed Businesses			-0.281***
			(0.048)
Yelp Reviews Growth (divided by 100)			0.056
			(0.074)
Constant	2.066***	2.095***	2.038***
	(0.154)	(0.156)	(0.153)
Year FE	Yes	Yes	Yes
Observations	83100	83100	83100
Adjusted R^2	0.230	0.233	0.235

All regressions include a full set of calendar year dummies and cluster standard errors at the ZIP Code level. Indicators High Density, High Income, and High Education equal 1 if a ZIP Code is above the median across all ZIP Codes in population density, median household income, and percent with a bachelor's degree, respectively. * p<0.10, ** p<0.05, *** p<0.01.

Table 6 Predicting CBP Establishment Growth by Area Attributes Using Random Forest

	Population Density		Income		Education	
	High	*Low*	*High*	*Low*	*High*	*Low*
R-squared	0.244	0.056	0.328	0.149	0.291	0.064
Out-of-bag R-squared	0.194	0.029	0.256	0.075	0.234	0.023
Mean Absolute Error	12.731	3.922	9.806	6.997	11.111	5.593
Mean Squared Error	427.918	42.065	292.104	186.273	363.237	110.182
Median Absolute Error	7.966	2.492	5.0785	3.476	6.030	3.034
Mean CBP Growth	*6.799*	*0.494*	*6.106*	*1.370*	*6.453*	*0.900*
St. Dev CBP Growth	*20.484*	*6.485*	*17.654*	*13.011*	*19.137*	*10.153*
Observations	*42644*	*42648*	*41548*	*41552*	*42224*	*42568*

Broken down by subsamples of the data based on population density, median household income, and percent with a Bachelor's degree, all analyses predict residual variance in the change in CBP establishments after regressing two lags of changes in CBP establishments with ZIP code and year fixed effects. Features include year and the change in and absolute number of total open, opened, and closed businesses as recorded by Yelp, aggregate review count, and average rating, and broken down by lowest and highest business price level. The sample covers the time period 2012-2015, and all observations for 2015 have been assigned to the test set, and the rest to training. The number of trees in the forest is set to 300. Each column indicates which subsample of the data was analyzed. The number of observations, means and standard deviations of CBP Growth are reported for each column using the full set of observations across both training and test sets.

Table 7 Industry Category Definitions

Category	NAICS sectors	Description
Retail, Leisure, and Hospitality	44, 45, 71, 72	Retail stores and dealers, arts, entertainment, recreation, accommodation, and food services
Goods Production	11, 21, 22, 23, 31, 32, 33	Agriculture, forestry, fishing, hunting, mining, quarrying, oil and gas extraction, utilities, construction, and manufacturing
Transportation and Wholesale Trade	42, 48, 49	Wholesale traders, markets, and agents; transportation and support activities; postal and delivery services; and warehousing
Information and Financial Activities	51, 52, 53	Publishing, media production, telecommunications, finance, insurance, real estate, and leasing
Professional and Business Services	54, 55, 56, 81	Professional, scientific, technical, administrative, and support services; management of companies; waste management; repair and maintenance; personal and laundry services; religious and other organizations
Public Services	61, 62, 92, 99	Education, health care, social assistance, public administration, and government

All CBP establishments are classified by NAICS codes, and each NAICS code was categorized into an industry category, based loosely on NAICS supersectors.

Table 8 Predicting CBP Establishment Growth by Industry Category Using Regression Analysis

	(1)	(2)	(3)	(4)	(5)	(6)
	Retail, Leisure, and Hospitality	Goods Production	Transportation and Wholesale Trade	Information and Financial Activities	Professional and Business Services	Public Services
CBP Establishment Growth (own industry, lag1)	-0.077	-0.010	0.006	-0.065	0.068^{***}	0.180^{***}
	(0.055)	(0.007)	(0.018)	(0.067)	(0.014)	(0.043)
CBP Establishment Growth (own industry, lag2)	0.003	0.044^{***}	0.039^{*}	0.038^{*}	0.103^{***}	0.095^{***}
	(0.060)	(0.006)	(0.015)	(0.019)	(0.013)	(0.028)
Yelp Business Growth	0.214^{***}	0.015^{**}	0.035^{***}	0.090^{***}	0.112^{***}	0.039^{***}
	(0.016)	(0.006)	(0.007)	(0.010)	(0.013)	(0.009)
Yelp Business Growth (lag1)	0.025	0.034^{***}	-0.007	0.068^{***}	0.102^{***}	0.054^{***}
	(0.013)	(0.005)	(0.006)	(0.011)	(0.012)	(0.010)
Yelp Growth in Closed Businesses	-0.112^{***}	-0.018	-0.038^{***}	-0.055^{***}	-0.041^{*}	-0.037^{*}
	(0.030)	(0.010)	(0.011)	(0.016)	(0.020)	(0.018)
Yelp Reviews Growth (divided by 100)	0.086^{**}	0.035^{**}	0.013	-0.039	0.083^{*}	0.084^{***}
	(0.030)	(0.011)	(0.017)	(0.033)	(0.033)	(0.019)
Constant	-0.139	-0.139^{***}	0.397^{***}	0.151^{*}	0.461^{***}	0.034
	(0.102)	(0.029)	(0.030)	(0.071)	(0.048)	(0.033)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	91068	91068	91068	91068	91068	91068
Adjusted R^2	0.085	0.020	0.009	0.051	0.102	0.082

All regressions include a full set of calendar year dummies and cluster standard errors at the ZIP Code level. * p<0.10, ** p<0.05, *** p<0.01.

Table 9 Predicting CBP Establishment Growth by Industry Category Using Random Forest

	Retail, Leisure, and Hospitality	Goods Production	Transportation and Wholesale Trade	Information and Financial Activities	Professional and Business Services	Public Services
R-squared	0.131	0.066	0.014	0.109	0.172	0.072
Out-of-bag R-squared	0.147	0.004	0.007	0.079	0.158	0.034
Mean Absolute Error	3.161	2.315	1.759	2.205	3.437	2.448
Mean Squared Error	36.203	13.300	10.468	17.752	38.502	36.945
Median Absolute Error	1.616	1.392	0.967	0.982	1.659	1.161
Mean CBP Growth	*0.648*	*0.280*	*0.193*	*0.469*	*1.030*	*0.774*
St. Dev CBP Growth	*5.755*	*3.585*	*3.231*	*4.498*	*6.303*	*5.097*
Observations	*91068*	*91068*	*91068*	*91068*	*91068*	*91068*

Broken down by subsamples of the data based on industry categories, all analyses predict residual variance in the change in CBP establishments after regressing two lags of changes in CBP establishments with ZIP code and year fixed effects. Features include year and the contemporaneous and lagged change in and absolute number of total open, opened, and closed businesses as recorded by Yelp, aggregate review count, and average rating, and broken down by lowest and highest business price level. The sample covers the time period 2012-2015, and all observations for 2015 have been assigned to the test set, and the rest to training. The number of trees in the forest is set to 300. Each column indicates which subsample of the data was analyzed. The number of observations, means and standard deviations of CBP Growth are reported for each column using the full set of observations across both training and test set.