Survival of Gentrification / Depreciation in Restaurants

F. M. Marsh and J. A. Clithero

September 18, 2016

1 Abstract

2 Introduction

Yelp Dataset Papers:

[Alghunaim, 2015, Byers et al., 2012, Cawkwell et al., 2015, Chepurna and Makrehchi, 2015, Feng and Qian, 2013, Gutierrez, 2014, Hajas et al., 2014, Hu et al., 2014, Liu et al., 2015, Mashhadi et al., 2012, Quattrone et al., 2015]

Zillow Dataset Papers:

Has previously been combined with the yelp dataset Bonnar et al..

Our goal is to use the correlation of two time-series:

1. The monthly median rent, as tracked by Zillow Rental Data. 2. The median restaurant review rating (stars) for each restaurant in a neighborhood.

Zillow rental data can be used to detect appreciating, and depreciating neighborhoods.

As rents rise in a given neighborhood, which types of businesses fare / worse better in the reviews? As rents fall in a given neighborhood, which types of business fare / worse better in the reviews?

We hope to present concrete suggestions to restaurant owners to improve the survivability of their businesses in times of strong appreciation / depreciation in the housing market.

3 Data

3.1 Yelp Academic Dataset

The Yelp Academic Dataset contains five files:

1) `yelp_academic_dataset_business.json`

2) `yelp_academic_dataset_review.json`

3) `yelp_academic_dataset_user.json`

4) `yelp_academic_dataset_checkin.json`

3.2 Zillow Public Dataset

The Zillow Public Dataset (hereafter Zillow dataset) contains many files.

Zillow divides homes into geographic “neighborhoods” with well defined boundaries. The Zillow Home Value Index (ZHVI) is Zillow’s best estimate of median home price in a neighborhood. ZHVI is reported on a monthly basis for 6,958 neighborhoods across the US.

Median rental price for studio, one, two, three, four and five or more bedroom apartments are reported for a smaller set of about 300 neighborhoods.

Each Zillow neighborhood has geographic boundaries, defined in an associated ESRI arcGIS shape file.

4 Methods

4.1 Combination of Datasets

Each Yelp business is tagged with a geographic (latitude, longitude) coordinate. In this section, we describe how we sort each Yelp business into its appropriate Zillow neighborhood.

To perform this sorting, we employ a two-step approach. In the first step, we test every Yelp business for inclusion in the bounding box of every Zillow neighborhood. In the second step, we test every Yelp business for polygon inclusion in the neighborhoods which bounding boxes it lies within. We use this two-step approach because the first step can rule out all but two or three of the

We introduce the concept of the bounding box which we will define as the smallest range of latitudes and longitudes that include the whole neighborhood polygon. We test each Yelp business for inclusion in the set of 6,958 bounding boxes. In Fig. 2, we see a randomly selected Yelp business, displayed as a red point. We see that this business is included in the bounding boxes of two Zillow neighborhoods.

We then test for point-in-polygon inclusion using an implementation of a ray-casting method in `Python` [**?**]. For each Yelp business, we only test the Zillow neighborhoods whose bounding boxes it lies within.

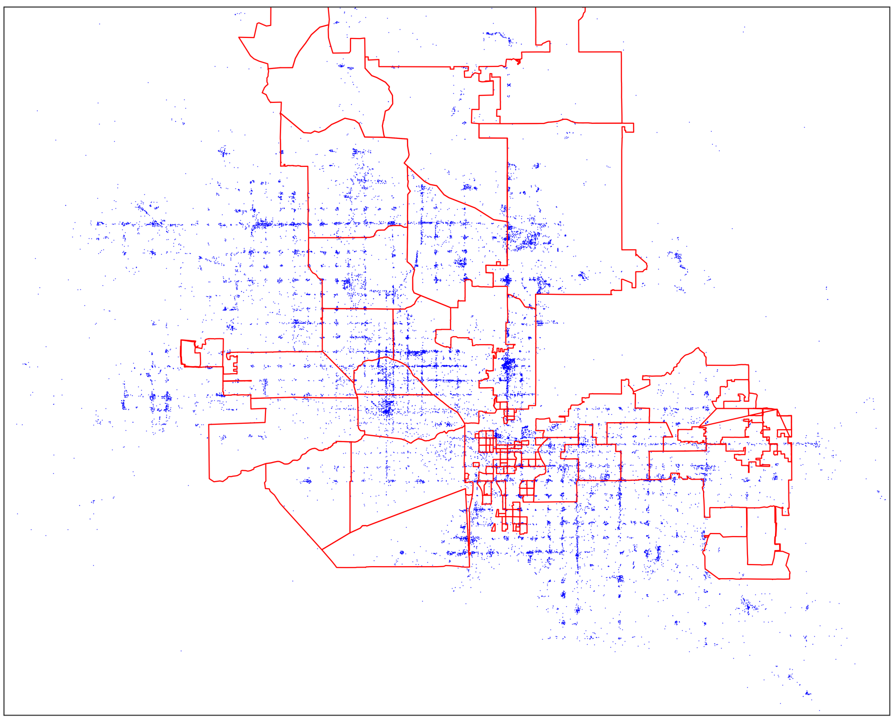

Figure 1: Yelp Businesses (points in blue) and Zillow neighborhood boundaries (lines in red) for the Phoenix, AZ metro area. In §4.1, we describe how we sort each Yelp Business into its appropriate Zillow neighborhood.

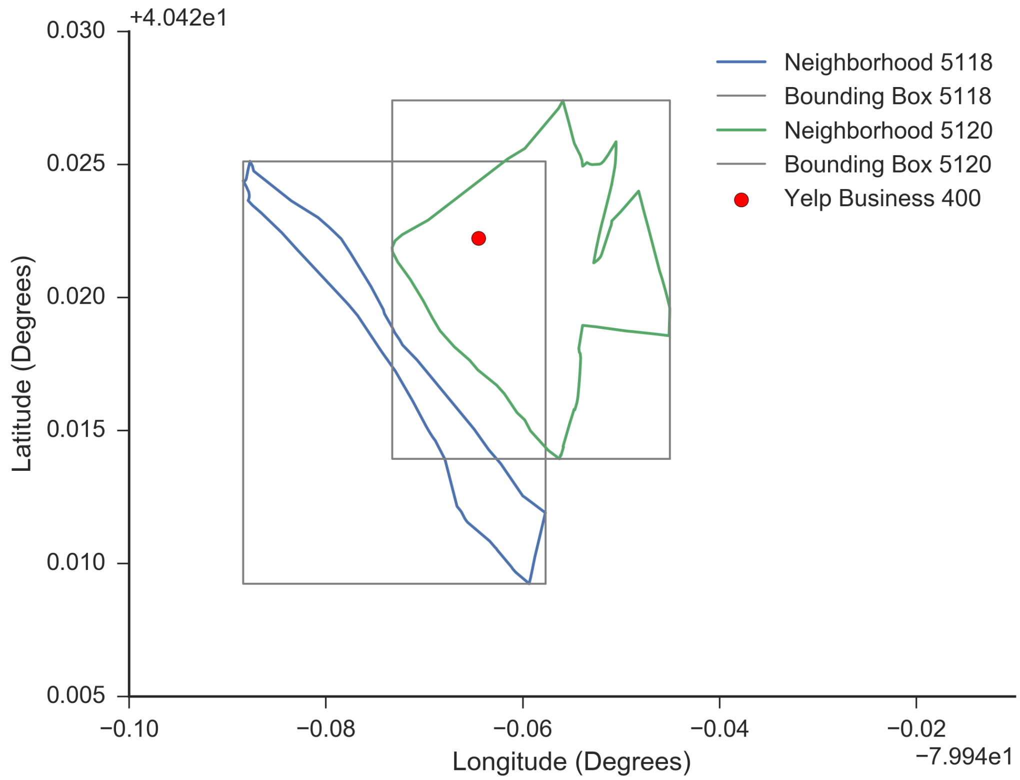

Figure 2: Example of neighborhood and neighborhood bounding box inclusion method. Yelp business 400 (the red point) is included in the bounding boxes of two Zillow neighborhoods. It is only included in one neighborhood polygon, however.

city	median	mean	sum	len
Charlotte	43.0	66.092308	4296	65
Henderson	105.0	153.000000	2907	19
Las Vegas	241.5	242.041667	5809	24
Madison	36.0	71.400000	1071	15
Mesa	491.5	509.500000	3057	6
North Las Vegas	903.0	903.000000	903	1
Phoenix	625.0	763.333333	11450	15
Pittsburgh	43.5	94.769231	2464	26
Scottsdale	2100.0	1531.666667	4595	3
Tempe	26.5	29.888889	538	18

state	median	mean	sum	len
AZ	116.5	467.619048	19640	42
NC	43.0	66.092308	4296	65
NV	178.0	218.613636	9619	44
PA	43.5	94.769231	2464	26
WI	36.0	71.400000	1071	15

4.2 Determination of Most Common Yelp Tags

Each Yelp business has user-generated tags, that allow other users to determine what genre the business is. For restaurants, common tags are "Mexican", "Chinese", etc.

4.3 Computation of Price Trend

References

A. Alghunaim. *A Vector Space Approach for Aspect-Based Sentiment Analysis*. PhD thesis, Massachusetts Institute of Technology, 2015.

C. Bonnar, F. Cordeiro, and J. Michelman. With a little help from yelp.

J. W. Byers, M. Mitzenmacher, and G. Zervas. The groupon effect on yelp ratings: a root cause analysis. In *Proceedings of the 13th ACM conference on electronic commerce*, pages 248–265. ACM, 2012.

P. B. Cawkwell, L. Lee, M. Weitzman, and S. E. Sherman. Tracking hookah bars in new york: Utilizing yelp as a powerful public health tool. *JMIR public health and surveillance*, 1(2), 2015.

I. Chepurna and M. Makrehchi. Exploiting class bias for discovery of topical experts in social media. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pages 64–71. IEEE, 2015.

H. Feng and X. Qian. Recommendation via user's personality and social contextual. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 1521–1524. ACM, 2013.

L. A. Gutierrez. *Noise reduction in user generated datasets*. PhD thesis, RENSSELAER POLYTECHNIC INSTITUTE, 2014.

P. Hajas, L. Gutierrez, and M. S. Krishnamoorthy. Analysis of yelp reviews. *arXiv preprint arXiv:1407.1443*, 2014.

L. Hu, A. Sun, and Y. Liu. Your neighbors affect your ratings: on geographical neighborhood influence to rating prediction. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 345–354. ACM, 2014.

J. Liu, J. Shang, C. Wang, X. Ren, and J. Han. Mining quality phrases from massive text corpora. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pages 1729–1744. ACM, 2015.

A. Mashhadi, G. Quattrone, L. Capra, and P. Mooney. On the accuracy of urban crowdsourcing for maintaining large-scale geospatial databases. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*, page 15. ACM, 2012.

G. Quattrone, L. Capra, and P. De Meo. There's no such thing as the perfect map: Quantifying bias in spatial crowd-sourcing datasets. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 1021–1032. ACM, 2015.