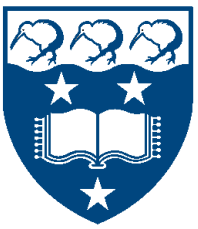
Explaining Business Satisfaction: Identifying Local Factors with Geographic Weighted Regression



THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

Douglas Callaway

October 2016

Department of Computer Science

Supervisor: Professor Mark Gahegan

A dissertation submitted in partial fulfilment of the requirements for the degree of Master of Professional Studies in Data Science, The University of Auckland, 2016.

Abstract

Last

Contents

1. Introduction

Sixth

- Motive (why)
- Global model can obscure interesting local relationships (Brunsdon, Fotheringham, & Charlton, 1996)
- Purpose (research question, goals, aims, etc.)

Related Work

- Science (soil science, disease mapping)
- Public sector (social trends)
- Business? -> lack of global data

- Structure

2. Background

2.1. Simple Linear Regression

A popular technique for both explanatory and predictive data analysis is simple linear regression. Unlike "black-box" prediction tools such as Artificial Neural Networks (ANN), simple linear regression has the advantage of quantifying the individual effects of each predictor variable on the response. This quality is important for understanding why a particular outcome is predicted, as well as using regression as an exploratory tool. This general model is formalised by the following equation:

$$y_i = a_0 + \sum_{k=1}^{m} a_k x_{ik} + \varepsilon_i \qquad (1)$$

where:
y_i is the ith observation of the dependent variable y,
a_0 is the intercept,
m is the number of predictor variables,
a_k is slope/effect of the independent variable x_k,
and ε_i is the error term where ε is independent and normally distributed with mean = 0.
(Dobson & Barnett, 2008, p. 89)

For example, a linear regression model that predicts salary from years of experience and certification by professional body would include the slope of each, which can be interpreted as a predictor variable's effect on salary with respect to the other variables in the model. So if the model was: $salary = 33 + 3(experience) + 10(certification) + \varepsilon$, then the slope/effect of experience is +3, meaning an additional year of experience at the same

certification level would yield an additional 3 units of salary, plus or minus some random error, ε (see Figure 1).

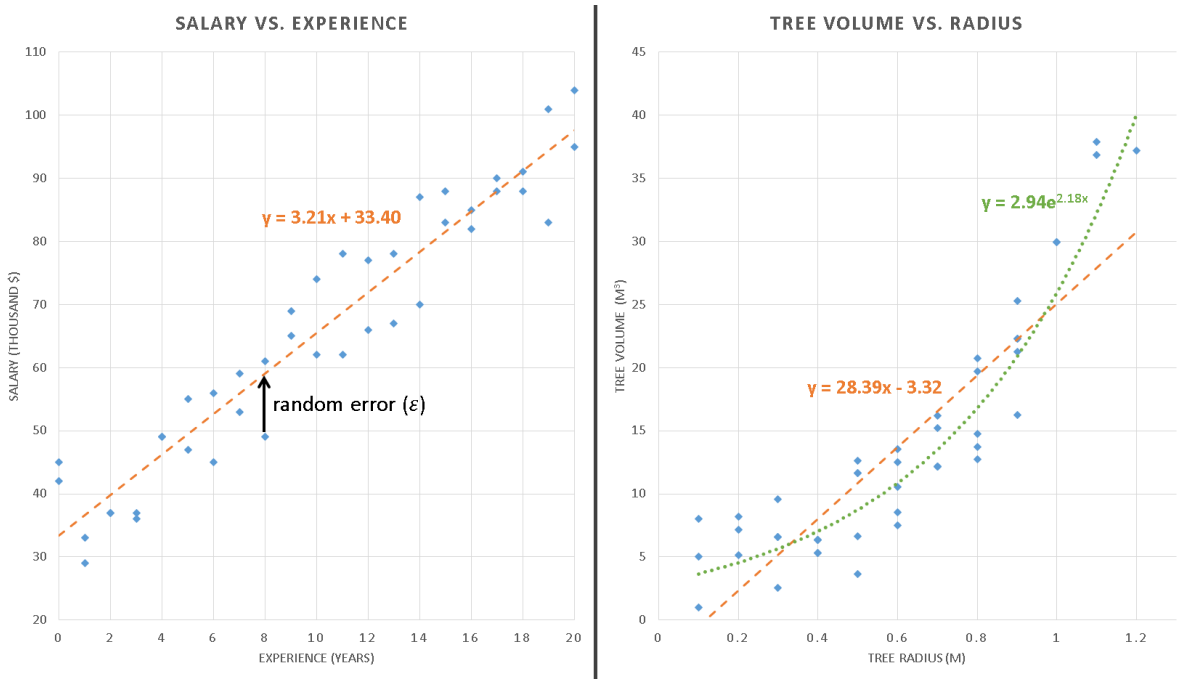


Figure 1: examples of proper (left plot) and improper (right plot) linear models using hypothetical data. Observations are depicted by blue diamonds while expected values lie along the dashed/dotted lines. The relationship between tree volume and radius (right) is clearly exponential, so the linear model (orange line) tends to underestimate volume at radii extremes and overestimate volume near average radii. The exponential model (green line) is clearly a better model. As shown in the left-hand plot, error values should be random and normally distributed about the expected value.

A key assumption in linear regression is a linear relationship between the response and each predictor. If there is a non-linear relationship, a linear model can still be fit, but the error terms (ε_is) will no longer be normally distributed, causing instability in the model (see Figure 1).

It also is important to observe that simple linear regression produces a *global* model (Brunsdon et al., 1996). Any local patterns or relationships among the data become generalized by a single line described only by an intercept and slopes for each variable. This can simplify the interpolation process where an unknown response must be predicted.

2.2. Spatial non-stationarity

The simplicity of such a global model is useful when studying truly global trends, or when the details of local phenomenon are not of interest. However, spatial data are inherently influenced by their locality. Tobler asserts that "everything is related to everything else, but near things are more related than distant things" (1970). Therefore, it makes sense to extend simple linear regression to identify interesting local relationships and how those relationships change over geographic space. These variations are referred to as spatial non-stationarity (Brunsdon et al., 1996). To extend the global model defined by equation $\boldsymbol{y_i = a_0 +}$

$\sum_{k=1}^{m} a_k x_{ik} + \varepsilon_i$

(1), a different model must be fit for each location of interest, formalised by the following equation:

$$y_i = a_{i0} + \sum_{k=1}^{m} a_{ik} x_{ik} + \varepsilon_i \quad (2)$$

where a_{ik} is slope/effect of the independent variable x_k *at the local observation*, i (Brunsdon et al., 1996).

This approach is referred to as Geographic Weighted Regression (GWR). However, GWR introduces additional complexity over simple regression – specifically regarding how to calculate the individual a_{ik} terms. Like simple linear regression, GWR generally uses a least squares method to find a_k terms for equation $y_i = a_0 + \sum_{k=1}^{m} a_k x_{ik} + \varepsilon_i$ (1) that minimize S in the following equation:

$$S = \sum_{i=1}^{n} [y_i - \mu_i(a_k)]^2 \quad (3)$$

where:
y_i is the actual response at point i,
and μ_i is the expected (mean) response at point i (Dobson & Barnett, 2008, p. 14).

The minimization of error, S simultaneously for all points, i makes simple linear regression a global model. To model spatial non-stationarity then, a different model must be fit for each location, with individual errors S_i, minimized only among each i's "local" points, j. Therefore, the concept of what is considered "local" must be made explicit. In other words, defining a locality implies devising a spatial weighting function that excludes distant, non-local points by either specifying a distance buffer, or effectively reducing their weights to zero via a decay function. Brunsdon et al. recommends the following "bisquare function" which provides a balance between the lower computational cost offered by the former and the parameter surface smoothness offered by the latter:

$$w_{ij} = \left[1 - \frac{d_{ij}^2}{d^2}\right]^2 \; if \; d_{ij} < d; \; w_{ij} = 0 \; otherwise \quad (4)$$

where:
i is a point where model parameters are estimated,
j is an observed point,
d_{ij} is the distance between points i and j,
and d is the spatial neighborhood distance (1996).

With this method, only points within the distance buffer, d will be considered, saving computation cost. Additionally, distant points, j near the buffer boundary will have relatively low influence on w_{ij} which will prevent the resulting parameter surface from having abrupt changes where those points are included or not (see Figure 2). Such a decay function also better respects Tobler's (1970) law by giving higher weight to closer points, j.

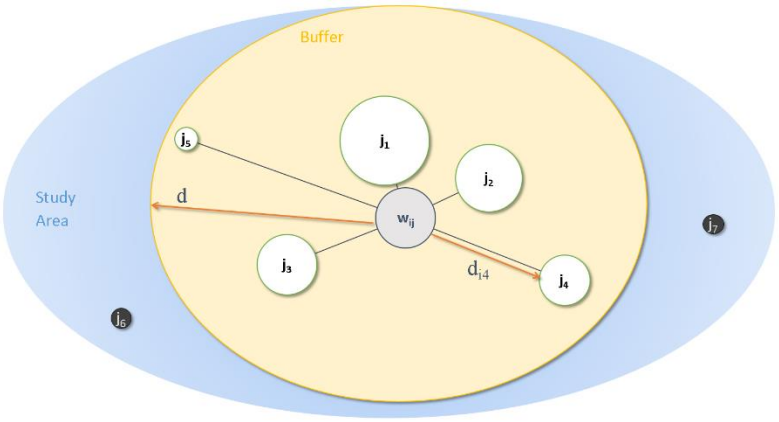


Figure 2: The spatial neighbourhood of an unmeasured parameter, w_{ij}, defined by a buffer with distance, d from w_{ij}. Nearby, measured locations ($j_{1\text{-}5}$) are used to estimate the value at w_{ij}. The sizes of $j_{1\text{-}5}$ illustrate their relative weight in determining w_{ij}. Closer points are assumed to be more closely related, so they are weighted heavier. Points 6 and 7 are outside the buffer, so they are not used to calculate w_{ij}. Adapted from Mitchell (1999).

However, equation $w_{ij} = \left[1 - \frac{d_{ij}^2}{d^2}\right]^2 \; if \; d_{ij} < d; \; w_{ij} = 0 \; otherwise$ (4) assumes a fixed buffer distance, d and decay function, $[1 - \mathrm{d}_{\mathrm{ij}}^2/\mathrm{d}^2]$. This may not be a valid assumption if the concept of "local" also varies over geographic space. Like simple linear regression, fixing the weighting function for all data may over-generalise the data and obscure local patterns. Brunsdon et al. concede that such a general weighting function may not always be appropriate, especially in economic applications where market areas vary widely (1996). For example, if a study seeks to compare customer behaviour across a large metropolitan region, market areas will likely be larger in more rural suburbs where customers are accustomed to traveling further for various services.

A possible alternative to equation $w_{ij} = \left[1 - \frac{d_{ij}^2}{d^2}\right]^2 \; if \; d_{ij} < d; \; w_{ij} = 0 \; otherwise$ (4) is a modified inverse-distance weighting (IDW) scheme. IDW is a generalized form of equation $w_{ij} = \left[1 - \frac{d_{ij}^2}{d^2}\right]^2 \; if \; d_{ij} < d; \; w_{ij} = 0 \; otherwise$ (4):

$$w_{ij} = \frac{1}{d_{ij}^{\alpha}} \qquad (5)$$

where:
α is the power (distance-decay parameter),
and $\sum_i^n w_i = 1$ (Lu & Wong, 2008).

This

(4), so all points would be considered for every w_{ij}. This is computationally expensive, especially considering most points not in the vicinity of w_{ij} will be near zero. Therefore, a maximum distance buffer can be added (Lu & Wong, 2008), as in equation $wij = \left[1 - \frac{d_{ij}^2}{d^2}\right]^2 \; if \; d_{ij} < d; \; w_{ij} = 0 \; otherwise$
(4).

Lu and Wong (2008) extend IDW into an "Adaptive" IDW (AIDW) technique to improve interpolation over study areas in which the spatial pattern of observations vary widely, such as with the metropolitan customer study example described above. In this method, α is calculated for each unmeasured location as a function of a statistic R, defined as:

$$R = \frac{r_{obs}}{r_{exp}} \tag{6}$$

where:
r_{obs} is the average nearest-neighbor distance for w_i's 5 nearest observed points, j,
and r_{exp} is the expected (mean) nearest-neighbor distance for a random point pattern (Lu & Wong, 2008).

This statistic provides a simple means for measuring relative clustering or dispersion by comparing the observed clustering, r_{obs} to what would be expected at random, r_{exp}. The r_{exp} term only must be calculated once for the study area with:

$$r_{exp} = \frac{1}{2(n/A)^{0.5}} \tag{7}$$

where A is the area of the study region (Lu & Wong, 2008).

Finally, the R_is are mapped to values of α. Areas with high dispersion (low R) should have lower α values to give higher weight to nearby points, and vice versa; more clustered areas should not be overly influenced by very-near points (Lu & Wong, 2008). The range of appropriate α values are dependent upon the range of distance-decay values in the study area. For flexibility, Lu and Wong (2008) recommend a triangular membership function as proposed by Kantardzic (2011, p. 418) for this mapping. Equation $w_{ij} = \frac{1}{d_{ij}^{\alpha}}$
(5) can then be modified as:

$$w_{ij} = \frac{1}{d_{ij}^{\boldsymbol{m(R_i)}}}$$

(8)

where $\mathrm{m}(\mathrm{R_{ij}})$ is the mapping function from R to α.

Equation $w_{ij} = \frac{1}{d_{ij}^{m(R_i)}}$ (8) can now be applied to generate adaptive, neighborhood-weighted estimates of a_{ik} throughout the study area. Typically, a least-squares approach using equation $\boldsymbol{S} = \sum_{i=1}^{n}[\boldsymbol{y_i} - \boldsymbol{\mu_i}(\boldsymbol{a_k})]^2$ (3) will be used to fit values for a_{ik} that minimize the $\boldsymbol{S_i}$s with respect to the neighbourhood's observations, js. The result is a surface for each coefficient (a_{ik}) in the model representing that variable's change in effect over geographic space.

3. Design

3.1. Study area & related data

Conducting spatial analysis, such as GWR, of business performance is fundamentally difficult. Business data is generally confidential or expensive to obtain, and data types and formats can vary significantly from one business to another. Many smaller businesses may not even collect or maintain useful performance metrics.

The recent rise of crowdsourced review services such as Yelp and TripAdvisor provide a useful solution to these problems. Additionally, Yelp provides the *Yelp Challenge Academic Dataset* (2016) which can be downloaded free for non-commercial research. The 2016 dataset includes 2.7 million reviews for 86 thousand businesses over ten cities. Given the common schemas for reviews and business profiles (including locations), spatial analysis can easily be conducted with minimal integration effort.

Phoenix, Arizona restaurants are used for this case study due to having the most data available, and familiarity with the study area and the restaurant business model. There are 622,446 restaurant reviews in the dataset for 9,427 Phoenix restaurants, covering a period from February, 2005 to July, 2016.

Due to the socioeconomic nature of this study, demographics from the U.S. Census Bureau are also considered. Potentially relevant factors include population (2010e), age (2010d), household composition (2010c), race (2010b), and education (2013b). Additional economic indicators include income (2013d), home values (2013c), and proportions of home owners and renters (2010f). These factors are evaluated as regression candidates in the data cleaning and exploration phase.

Road networks likely play a key role in restaurant success, so they are considered well. OpenStreetMap (2016) is used for its state-wide road network differentiating road types and locations.

3.2. Assumptions

Several assumptions underpin the design decisions of this study. First, for any useful regression (global or local) model there must be plausible, linear relationships between the response and predictor variables. Each restaurant's star rating is assumed to be a linear combination of some subset of demographic, review content, or restaurant profile attributes available in, or derived from, the given data. These relations are evaluated in detail in the data cleaning and exploration phase.

Next, I assume a smooth gradient of the various parameters over the study area. This aligns with Tobler's (1970) law of near things being closely related, and the predominant use of distance-decay functions in related studies (e.g. Brunsdon et al., 1996; Lu & Wong, 2008). Further, there are no major geographic or administrative barriers in the study area to disrupt these kinds of relationships which might cause abrupt changes in the parameter surface.

Given the economic nature of the study, there is a high likelihood of variable spatial neighbourhood sizes over the study area (Brunsdon et al., 1996). Therefore, I choose an AIDW (Lu & Wong, 2008) method for determining neighbourhood size when estimating GWR coefficients. This allows for greater flexibility between the dense urban centre and sparse outer suburbs of the study area.

Because roads are used to travel to restaurants, a drive-time buffer would likely provide better market area estimates than the circular, variable distance buffer used by AIDW. However, due to the large dataset, the simpler AIDW buffer is used to allow for reasonable processing times. This is also expected to improve prediction in areas that currently have limited road networks, e.g. outer suburbs where future growth may occur.

3.3. Software tools

Open source software is used throughout the implementation and evaluation phases to ensure easy replication of these results. All scripts are written using Python version 3.5

- Docker
- Python (pandas, geopandas, etc.)
- ArcGIS

3.4. Algorithms

- Sentiment analysis
- Model selection / optimisation
- Weighting function
- Statistics

3.5. Outputs

- Global model

- GWR model
- Coefficient surfaces
- Standard error surfaces (where does model perform well?)

3.6. Statistical tests

- Is GWR model significantly better than global model (Brunsdon et al., 1996)?
- Is the variation of coefficients across space significant (Brunsdon et al., 1996)?

4. Implementation

4.1. Data selection & cleaning

The Yelp dataset (2016) consists of two JSON files used in this study: "yelp_academic_dataset_business," containing attributes for each business, and "yelp_academic_dataset_review," containing user-submitted reviews. The reviews can be linked to their corresponding business by the "business_id" field. Phoenix restaurants are extracted by filtering on the "state" field for "AZ" (Phoenix is the only Arizona city included in the data) and on the "categories" field for "Restaurants". Several traits from the "attributes" field are used to further differentiate the restaurants such as "Price Range" and "Outdoor Seating." The full list of extracted variables is in Appendix A.

The "review.json" file is relatively large (2.3 GB), and is aggregated before combining it with the restaurant data. Review text is reduced to a sentiment score for each review as described in section 4.2. Those scores are then averaged, indicating a predicted overall star-rating from the review text alone. The "review_count" field in "business.json" proved to be unreliable, so the counts of distinct "business_id" are used instead. Finally, a review span is aggregated from the earliest to latest review date.

A potential problem with the review data are the various sample sizes of reviews among restaurants. Review counts range from 1 to 1,610. It is likely that many restaurants with low review counts are significantly biased due to self-reviewing or being negatively reviewed by competitors. Therefore, only restaurants with at least five reviews are considered here.

The US Census data (2010b-f; 2013b-d) are also of high quality with few significant anomalies or omissions. For median home value and household income however, some cases have character values indicating values less-than or greater-than (e.g. > 1,000,000). In these situations, the extra characters were removed, so the observation took the stated value (e.g. 1,000,000). The data come filtered to the study area using the online download tool, so only the relevant variables must be selected. The demographic and economic indicators selected for this study are listed in Appendix A.

Finally, the selected data are converted to spatial features. The Yelp data is converted to point features via the restaurant latitude and longitude provided in "business.json." The census data are joined with their corresponding *Block Groups* shapefiles (U.S. Census Bureau,

2010a, 2013a) on the "Id2" field. The resulting spatial data is projected to UTM Zone 12N (WGS 1984) to ensure consistent distance measurements over the study area.

4.2. Additional candidate variable calculation

Often a relationship between a response and a set of predictors is better modelled by some combination of the predictors rather than the individual predictors themselves (Lumley, 2016). For example, a financial health is better indicated by income minus expenditures rather than income and expenditures considered separately. Therefore, additional candidate variables are calculated from the given data with the hypothesis that they have better predictive power than the raw variables.

One potentially informative metric is restaurant uniqueness. This can be calculated as:

$$u_r = \frac{1}{\sum n_r} \tag{9}$$

where n_r is the set of restaurants with the same name as u_r.

Restaurants with higher uniqueness may have an advantage in areas where customers prefer a more personal, "Mom & Pop" atmosphere. Conversely, lower uniqueness (e.g. for a national chain) may have an advantage in locations where brand recognition dominates.

Another potential factor in review performance is the proximity of competitors. As with the R statistic defined by equation $R = \frac{r_{obs}}{r_{exp}}$ (6), competitor proximity can be modelled as the relative amount of restaurant clustering or dispersion compared to a random point pattern. Thus, each restaurant's competitor score here is defined as R calculated at the restaurant's location.

Perhaps the strongest predictors of star ratings are the contents of the reviews themselves. The processing of text to predict such an outcome is commonly referred to as sentiment analysis. Although sentiment analysis is beyond the scope of this paper, a simple Support Vector Machine (SVM) predictor based on Scikit Learn (2014) is established for the purpose of calculating a "sentiment" score corresponding to the predicted star rating. A summary of these calculated variables, along with the other selected candidate predictors is available in Appendix A.

4.3. Variable selection & normalisation

Given the set of candidate variables, predictors must be chosen that help predict restaurants' star ratings. For linear regression, these relationships must also be linear as illustrated by Figure 1: examples of proper (left plot) and improper (right plot) linear models using hypothetical data. Observations are depicted by blue diamonds while expected values lie along the dashed/dotted lines. The relationship between tree volume and radius (right) is clearly exponential, so the linear model (orange line) tends to underestimate volume at radii

extremes and overestimate volume near average radii. The exponential model (green line) is clearly a better model. As shown in the left-hand plot, error values should be random and normally distributed about the expected value.

- Linear transformations

4.4. Global model

- Iterative model selection

4.5. Geographic Weighted Regression model

- Iterative model selection

5. Evaluation

Fourth

5.1. Visualisation

- Smoothness
- Local patterns
- Sensitive to “edge effects” (Brunsdon et al., 1996)

5.2. Statistical significance

- Is GWR model significantly better than global model (Brunsdon et al., 1996)?
- Is the variation of coefficients across space significant (Brunsdon et al., 1996)?

5.3. Performance

- Overall: standard error, multicollinearity, etc.
- Over study area
- Processing time

6. Conclusion

Fifth

6.1. Results

- Implications (for decision makers, businesses, etc.)
- New insights
- Factors weren’t important, why? (e.g. census data)

6.2. Future work

- New questions stemming from analysis
- Enhancing models

- Better neighbourhood estimators (alternative to IDW, e.g. drive time vs. circular buffers)

References

Brunsdon, C., Fotheringham, A. S., & Charlton, M. E. (1996). Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical analysis, 28*(4), 281-298.

Dobson, A. J., & Barnett, A. (2008). *An introduction to generalized linear models*: CRC press.

Kantardzic, M. (2011). *Data mining : concepts, models, methods, and algorithms* (Second Edition.. ed.): Piscataway, New Jersey : IEEE Press ; Hoboken, NJ : Wiley. 2011 ©2011.

Lu, G. Y., & Wong, D. W. (2008). An adaptive inverse-distance weighting spatial interpolation technique. *Computers & Geosciences, 34*(9), 1044-1055.

Lumley, T. (2016). *STATS 762 Lecture*. University of Auckland.

Mitchell, A. (1999). *The ESRI Guide to GIS Analysis: Geographic patterns & relationships* (Vol. 1): ESRI, Inc.

OpenStreetMap. (2016). *Arizona*. OpenStreetMap. Retrieved from: http://download.geofabrik.de/north-america/us/arizona.html

scikit-learn. (2014). Working With Text Data. Retrieved from http://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic geography, 46*(sup1), 234-240.

U.S. Census Bureau. (2010a). *Arizona 2010 Block Groups*. 2010 TIGER/Line Shapefiles: Block Groups. Retrieved from: https://www.census.gov/cgi-bin/geo/shapefiles/index.php?year=2010&layergroup=Block+Groups

U.S. Census Bureau. (2010b). *Hispanic or Latino, and Not Hispanic or Latino by Race for the Population 18 Years and Over*. 2010 Census. Retrieved from: http://factfinder.census.gov/bkmk/table/1.0/en/DEC/10_SF1/P11/0500000US04007.15000|0500000US04013.15000|0500000US04021.15000|0500000US04025.15000

U.S. Census Bureau. (2010c). *Households and Families*. 2010 Census. Retrieved from: http://factfinder.census.gov/bkmk/table/1.0/en/DEC/10_SF1/QTP11/0500000US04007.15000|0500000US04013.15000|0500000US04021.15000|0500000US04025.15000

U.S. Census Bureau. (2010d). *Median Age by Sex*. 2010 Census. Retrieved from: http://factfinder.census.gov/bkmk/table/1.0/en/DEC/10_SF1/P13/0500000US04007.15000|0500000US04013.15000|0500000US04021.15000|0500000US04025.15000

U.S. Census Bureau. (2010e). *Total Population*. 2010 Census. Retrieved from: http://factfinder.census.gov/bkmk/table/1.0/en/DEC/10_SF1/P1/0500000US04007.15000|0500000US04013.15000|0500000US04021.15000|0500000US04025.15000

U.S. Census Bureau. (2010f). *Total Population in Occupied Housing Units by Tenure*. 2010 Census. Retrieved from: http://factfinder.census.gov/bkmk/table/1.0/en/DEC/10_SF1/H11/0500000US04007.15000|0500000US04013.15000|0500000US04021.15000|0500000US04025.15000

U.S. Census Bureau. (2013a). *Arizona 2013 Block Groups*. 2013 TIGER/Line Shapefiles: Block Groups. Retrieved from: https://www.census.gov/cgi-bin/geo/shapefiles/index.php?year=2013&layergroup=Block+Groups

U.S. Census Bureau. (2013b). *Educational Attainment for the Population 25 Years and Over*. 2009-2013 American Community Survey 5-Year Estimates. Retrieved from: http://factfinder.census.gov/bkmk/table/1.0/en/ACS/13_5YR/B15003/0500000US04007.15000|0500000US04013.15000|0500000US04021.15000|0500000US04025.15000

U.S. Census Bureau. (2013c). *Median [Home] Value (Dollars)*. 2009-2013 American Community Survey 5-Year Estimates. Retrieved from: http://factfinder.census.gov/bkmk/table/1.0/en/ACS/13_5YR/B25077/0500000US04007.15000|0500000US04013.15000|0500000US04021.15000|0500000US04025.15000

U.S. Census Bureau. (2013d). *Median Household Income in the Past 12 Months (In 2013 Inflation-Adjusted Dollars)*. 2009-2013 American Community Survey 5-Year Estimates. Retrieved from: http://factfinder.census.gov/bkmk/table/1.0/en/ACS/13_5YR/B19013/0500000US04007.15000|0500000US04013.15000|0500000US04021.15000|0500000US04025.15000

Yelp. (2016). *Yelp Challenge Academic Dataset*. Retrieved from: https://www.yelp.com/dataset_challenge/dataset

Appendix A: Selected Candidate Variables

Dataset	Source Name	Target Name	Transformation
"business.json" (Yelp, 2016)	'name'	'uniqueness'	$u_r = \frac{1}{\sum n_r}$ (9)
	'latitude', 'longitude'	'competitor_proximity'	$R = \frac{r_{obs}}{r_{exp}}$ (6)
	'Alcohol'	'beer_and_wine', 'full_bar'	n/a
	'Price Range'	'price_range'	n/a
	'Attire'	'attire'	n/a
	'Take-out'	'takeout'	n/a
	'Waiter Service'	'waiter_service'	n/a
	'Outdoor Seating'	'outdoor_seating'	n/a
"review.json" (Yelp, 2016)	'business_id'	'review_count'	count
	'stars'	'stars'	mean
	'text'	'sentiment'	See "review_data_ classify.ipynb"
	'date'	'review_span'	max(date) – min(date)
U.S. Census Bureau (2010e)	'Total'	'population_density'	$\frac{count}{kilometer^2}$
U.S. Census Bureau (2010f)	'Owned with a mortgage or a loan'	'home_mortgage_density'	
	'Owned free and clear'	'home_owner_density'	
	'Renter occupied'	'renter_density'	
(U.S. Census Bureau, 2013b)	'Estimate; Total: - Regular high school diploma' + 'Estimate; Total: - GED or alternative credential'	'density_education_highschool'	
	'Estimate; Total: - Associate\'s degree' + 'Estimate; Total: - Bachelor\'s degree'	'density_education_undergraduate'	
	'Estimate; Total: - Master\'s degree' + 'Estimate; Total: - Professional school degree' + 'Estimate; Total: - Doctorate degree'	'density_education_postgraduate'	
U.S. Census Bureau (2010b)	'Hispanic or Latino'	'hispanic_latino_population_density'	
	'Not Hispanic or Latino: - Population of one race: - White alone'	'white_population_density'	
	'Not Hispanic or Latino: - Population of one race: - Black or African American alone'	'black_population_density'	
	'Not Hispanic or Latino: - Population of one race: - American Indian and Alaska Native alone'	'native_american_population_density'	
	'Not Hispanic or Latino: - Population of one race: - Asian alone'	'asian_population_density'	
U.S. Census Bureau (2010c)	'Number; HOUSEHOLD TYPE - Total households'	'household_density'	
	'Number; HOUSEHOLD TYPE - Total households - Family households [1]'	'family_household_density'	
	'Number; HOUSEHOLD SIZE - Total households - 1-person household'	'single_household_density'	
	'Number; HOUSEHOLD SIZE - Total households - Average household size'	'average_household_size'	n/a
U.S. Census Bureau (2010d)	'Median age -- - Both sexes'	'median_age'	n/a
(U.S. Census Bureau, 2013c)	'Estimate; Median value (dollars)'	'median_home_value'	n/a

(U.S. Census Bureau, 2013d)	'Estimate; Median household income in the past 12 months (in 2013 inflation-adjusted dollars)'	'median_household_income'	n/a

Appendix B: