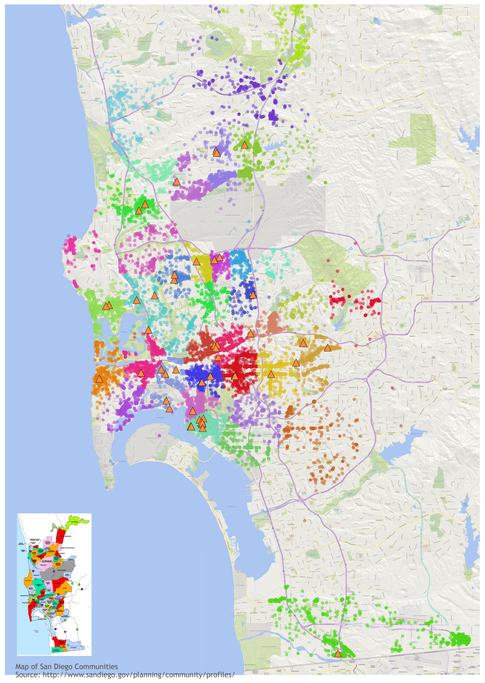
Gentrification Index in San Diego Using Yelp Data

Gene Burinsky, Jianwei Li, Tianyi Wang & Xiaoxian Wu

Abstract

We aim to make an index for the relative likelihood of a San Diego community to be gentrified. With abundant data scraped from Yelp on all businesses in San Diego, CA, as well as, data on almost all of the users who have reviewed restaurants found among the business data. We first group all the businesses into 40 areas using k-means algorithm and classify business data by their pricing level. We then use a support vector machine (SVM) and a radial-nearest neighbor (RNN) classifier to categorize either a cluster or each business based on the surrounding members. Since the SVM does not generate satisfying results, only the RNN output is reserved. We categorize users into 9 categories based on Latent-Dirichlet allocation (LDA) model on business categories and prices. We assign each user to one of the 9 topics based on their consumption history and then, using a logistic model, calculate the relative likelihood of each group to patronize a gentrifying business. This gives us the relative tendency of each the 9 consumer preference bundles to consume in a gentrifying business. Finally, given the 40 communities into which we divided San Diego, we use a multivariate logistic regression to assign a probability of each of the 9 categories to consume in a given community. To attain an index aggregated over the 9 groups for each community, we plug in the returned probabilities into the fitted likelihood of gentrification model.

Radial-Nearest-Neighbor Classification



- 10% of the entire businesses are identified as expensive
- Radial-Nearest-Neighbor (RNN) algorithm suggests 20% of expensive businesses are gentrifiers

- Across all categories in total, we have 26,000 businesses scraped from Yelp. The big map above shows the 40 groups of areas into which all our businesses have been divided into using a k-means algorithm. We could have used spectral but it doesn't make much of a difference in the output. The grouping results using k-means approximately correspond to the community divisions of San Diego shown in the corner.
- The triangles in the big map designate whether a business is a gentrifier or not. The Radial-Nearest-Neighbor algorithm (RNN) is applied to identify those gentrifiers. The analysis is based on a distance of .00025, which is approximately a quarter of a block in degrees (map metric).

Latent Dirichlet Allocation for User Clustering

$K = 9, M = 1$

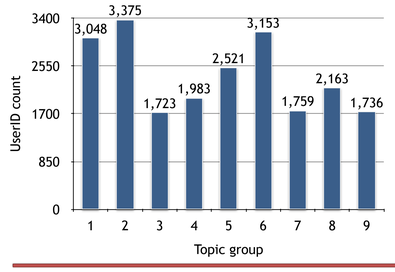
$\varphi_{k=1,\dots,K} \sim Dirichlet_V(\beta)$

$\theta_{d=1} \sim Dirichlet_K(\alpha)$

$z_{d=1,w=1,\dots,N} \sim Categorical_K(\theta_{d=1})$

$w_{d=1,w=1,\dots,N} \sim Categorical_V(\varphi_{z_{dw}})$

topic group	userID count
1	3048
2	3375
3	1723
4	1983
5	2521
6	3153
7	1759
8	2163
9	1736



Top 5 Words in 9 Topics

Topic 1: mexican1.0 sandwiches1.0 burgers1.0 hotdogs1.0 fastfood1.0

Topic 2: breakfast2.0 breakfastbrunch2.0 brunch2.0 newamerican2.0 american(new)2.0

Topic 3: divebars1.0 pubs1.0 grocery2.0 bars1.0 barbers2.0

Topic 4: japanese2.0 sushi2.0 sushibars2.0 seafood2.0 asianfusion2.0

Topic 5: coffee1.0 tea1.0 sandwiches1.0 cafes1.0 bakeries1.0

Topic 6: bars2.0 sportsbars2.0 burgers2.0 tradamerican2.0 american(traditional)2.0

Topic 7: pizza1.0 italian1.0 sandwiches1.0 seafood3.0 salad1.0

Topic 8: coffee2.0 icecream1.0 chinese1.0 bakeries2.0 vietnamese1.0

Topic 9: italian2.0 vegetarian2.0 pizza2.0 vegetarian1.0 vegan2.0

Note: the number following each word is its associated price level

Logistic Regression

$y_i = \{0,1\}_{i=1}^{n=\#\ of\ businesses}$

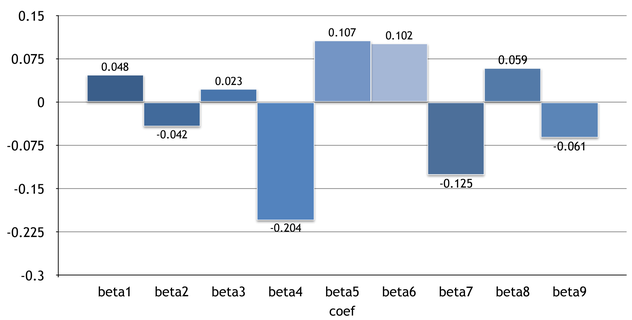
$RNN\ output \Rightarrow \begin{cases} y_i = 0\ represents\ for\ non-gentrifying\ business \\ y_i = 1\ represents\ for\ gentrifyiing\ business \end{cases}$

$x_{ij} = \#\ of\ users\ from\ group\ j\ who\ reviewed\ business\ i,\ \text{j} = 1,\dots,9$

$\Rightarrow Index_k = \log\left(\frac{p}{1-p}\right) = \beta_0 + \sum_{j=1}^{9} \beta_j x_{kj}$

⇒ RNN logistic regression result:

	var	coef
0	X0	-1.707240
1	X1	0.047583
2	X2	-0.042175
3	X3	0.022985
4	X4	-0.203906
5	X5	0.107342
6	X6	0.102000
7	X7	-0.125309
8	X8	0.058917
9	X9	-0.060631



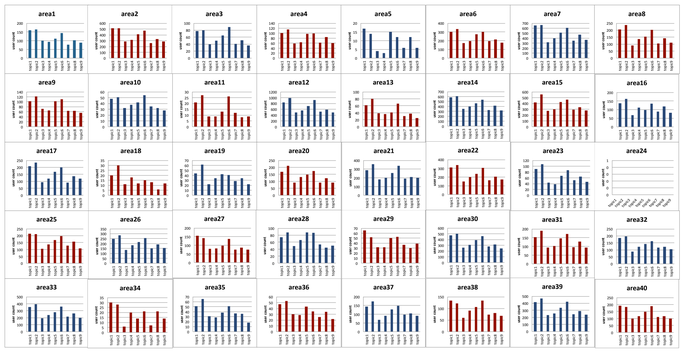
$$\Rightarrow \widehat{index}_k = \log\left(\frac{\hat{p}}{1-\hat{p}}\right) = -1.707 + 0.048x_{k1} - 0.042x_{k2} + 0.023x_{k3} - 0.204x_{k4} + 0.107x_{k5} + 0.102x_{k6} - 0.125x_{k7} + 0.059x_{k8} - 0.061x_{k9}$$

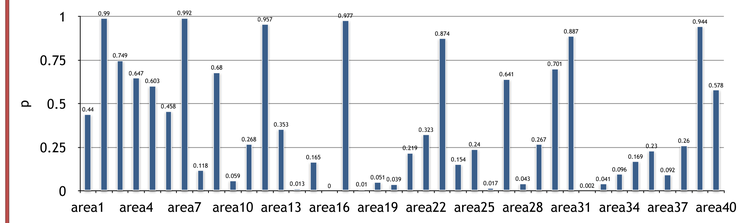
Index Prediction

Predict the index for each area to be gentrifying. For each area k:

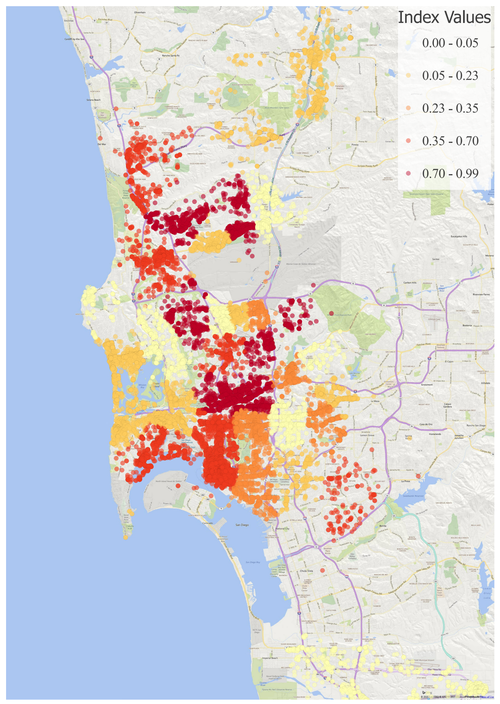
$Index_k = \beta_0 + \sum_{j=1}^{9} \beta_j x_{kj}$

$x_{kj} = \#\ of\ users\ from\ group\ j\ who\ reviewed\ businesses\ in\ area\ k, where\ k = 1,\dots,40; j = 1,\dots,9$





Index Prediction Map



Acknowledgements

We would like to thank professor Sayan Mukherjee for giving us helpful suggestions, especially the recommendation of using topic modeling for user classification. We would also like to thank Yelp for its employees being condescending and noncooperative in our attempt to purchase the data.