Chathura Gunasekara
cjgunase@mtu.edu

Using location related Yelp data to estimate and predict restaurant success

Key words: Yelp dataset challenge, location for business success.

Abstract

Restaurants are a multi billion-dollar industry in the US. The sizable economic impact of restaurants makes this industry a significant part of American life. There has been few research and surveys conducted based on smaller dataset limiting to smaller geographic areas. The research done by H.G.Parsa in [1] explains restaurant survivability based on the population in the neighborhood, income and educational level. But there has not been any research done using a big data source which includes intricate details of the restaurants such as precise geographic location, customer feedback and commercial value of the demographic trends. The Yelp academic dataset provides a window of opportunity to analyze the complex interactions of the restaurant industry using a multiple viewpoints spanning from a reviews of an individuals to finer details about the restaurant. Although the success of a restaurant is a complex process involving many variables such as location, menu, management skills, whether serving the correct people this study will focus on the location factor using large sample data set of 35000 restaurants across the United States.

Introduction

Location of restaurant, plays a major role in attracting customers. Easy access and close proximity to places where people frequently patron should be an ideal place for a restaurant business. Having located near a major people attraction such as department stores, schools, hotels should be a contributing factor to a restaurant success. But on the contrary to this common held belief issues related to commercial area such as high rent, not enough parking, high noise level, living cost of employees, competition from direct and indirect restaurants can be a limiting factor for a restaurant success. The availability of Yelp! Data has made possible this analysis which would have be not have been possible few years back. This project will focus on how much the location can account for the restaurant

success and to how extent we can predict a restaurant success based on only the location related data.

Data Preprocessing and Feature Extraction

The yelp dataset contains 5 files as large as 2Gb. Each file is JSON format and stores information about Business, review, user, tips. To do any type of machine learning the data has to be converted to a format which the algorithms can operate on.

A record of the business data file consists of the following information.

```
{
    'type': 'business',
    'business_id': (encrypted business id),
    'name': (business name),
    'neighborhoods': [(hood names)],
    'full_address': (localized address),
    'city': (city),
    'state': (state),
    'latitude': latitude,
    'longitude': longitude,
    'stars': (star rating, rounded to half-stars),
    'review_count': review count,
    'categories': [(localized category names)]
    'open': True / False (corresponds to closed, not business hours),
    'hours': {
        (day_of_week): {
            'open': (HH:MM),
            'close': (HH:MM)
        },
        ...
    },
    'attributes': {
        (attribute_name): (attribute_value),
        ...
    },
}
```

In this project, the problem that is going to be answered is whether the businesses in the vicinity affect the restaurant success or not and how much the success of a restaurant is accountable to its location in a commercial area. From background search we can assume the value of the real-estate by the amount of businesses near by and how successful they are in attracting more and more customers.

Identified large people attractions are,

1. Department stores such as Walmart, Kohl's, Target, Home Depot
2. Shopping stores such as Macy's or any other fashion outlets
3. Other restaurants nearby
4. Hotels or Motels
5. Arts and Entertainment such as cinemas, museums, theatres.

The data extraction algorithm will work as follows,

1. The Zip-code of the restaurant is extracted.
2. All the businesses 1 – 5 categories in the same Zip-code is extracted.
3. Cumulative scores for each categories is calculated and stored.
4. From the business file information about parking is extracted.

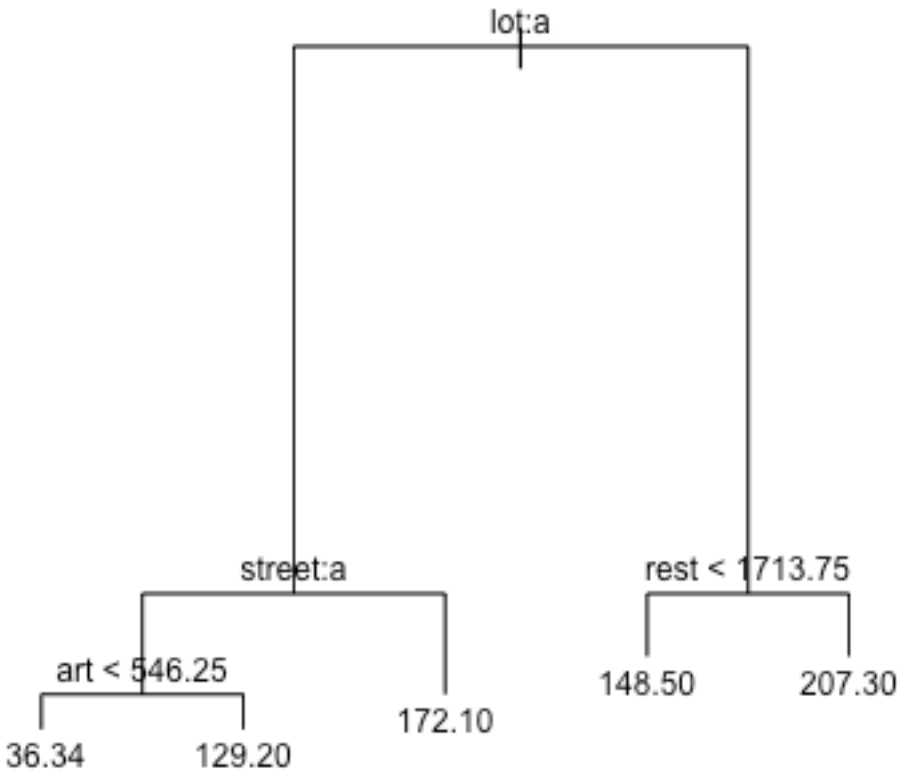
Finally following table is constructed for each restaurant.

street	lot	dep_stores	rest	shopping	hotel	art	score
FALSE	TRUE	0	7215.5	0.0	0.0	0.0	33.0
FALSE	FALSE	0	190.5	1084.5	0.0	192.0	17.5
TRUE	FALSE	0	14059.5	598.0	1414.0	1379.0	57.5
FALSE	TRUE	0	6144.5	157.5	52.5	0.0	1288.0
FALSE	FALSE	144	3329.5	87.0	476.0	623.0	668.5
FALSE	FALSE	0	8084.0	740.0	0.0	162.5	33.0
FALSE	FALSE	0	5534.5	182.5	52.5	0.0	17.5
FALSE	FALSE	0	13106.5	474.5	1618.0	196.5	9.0
FALSE	FALSE	0	807.0	340.5	0.0	162.0	28.0
TRUE	FALSE	0	12682.0	551.0	1414.0	1320.5	595.0
FALSE	FALSE	144	3953.0	87.0	476.0	648.0	45.0
FALSE	TRUE	0	5403.0	182.5	52.5	444.0	129.0

A Regression Tree is built to predict the score based on the values in other variables.

Analysis steps:

1. In the restaurant store there are outliers which can cause problems in the analyses are removed.
2. Data set was split to training (4/5) and testing set (1/5)
3. “tree” package in R is used to build a regression tree from the training set.



using the predict() function, score for the testing data is predicted and RMSE is calculated as follows.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(p_i - a_i)^2}{n}}$$

where p_i is the predicted score and a_i is the actuals score.

RMSE = 102.37

This RMSE value is not very low which implies the prediction of restaurant score solely based on nearby businesses is not a good method to measure a restaurants success as there can be many other factors such managerial issues, menu, food related issues.

Geo-statistical Analysis

Every restaurant has a star rating from 1 to 5. Which is calculated by averaging of the star ratings of user reviews. But this rating is inadequate for our analysis because a restaurant with one 5-star review and a restaurant with multiple 5-star reviews will be considered same because of averaging. So to counter this effect average star rating is multiplied by the review count, which gives a value which is better to explain the success because it explains number of customers visited and the rating each customer gave.

The following table is built using the business data for all the 35000 restaurants.

latitude	longitude	Stars * Review count

This dataset is a good candidate to see if all high scoring restaurants clustered together in a city or are they randomly distributed throughout the city. This can be described as Spatial Autocorrelation, from ESRI GIS definition, "A measure of the degree to which a set of spatial features and their associated data values tend to be clustered together in space (positive spatial autocorrelation) or dispersed (negative spatial autocorrelation)" [2].

Moran's Index is a measure of spatial autocorrelation, which is how much is the similar scores are clustered together. Using the "geoR" package in R, the Moran's Index gave approximately zero which indicates there is no spatial autocorrelation in the restaurants score. The restaurants have a random distribution regardless of the score of the restaurant. So it is not that high scoring restaurants are clustered together and low scoring restaurants cluster together. There is mix of high scoring and low scoring restaurants exist close by.

Moran's I	Expected	Std. Deviation	P-Value
09678886	-0.000245459	0.04677536	0.0380352

The following map of Las Vegas, NV shows the distribution of restaurants with respective colors for score.

Legend for Score

	(329,1.83e+04]
	(92.5,329]
	(30,92.5]
	[3,30]

Las Vegas, NV

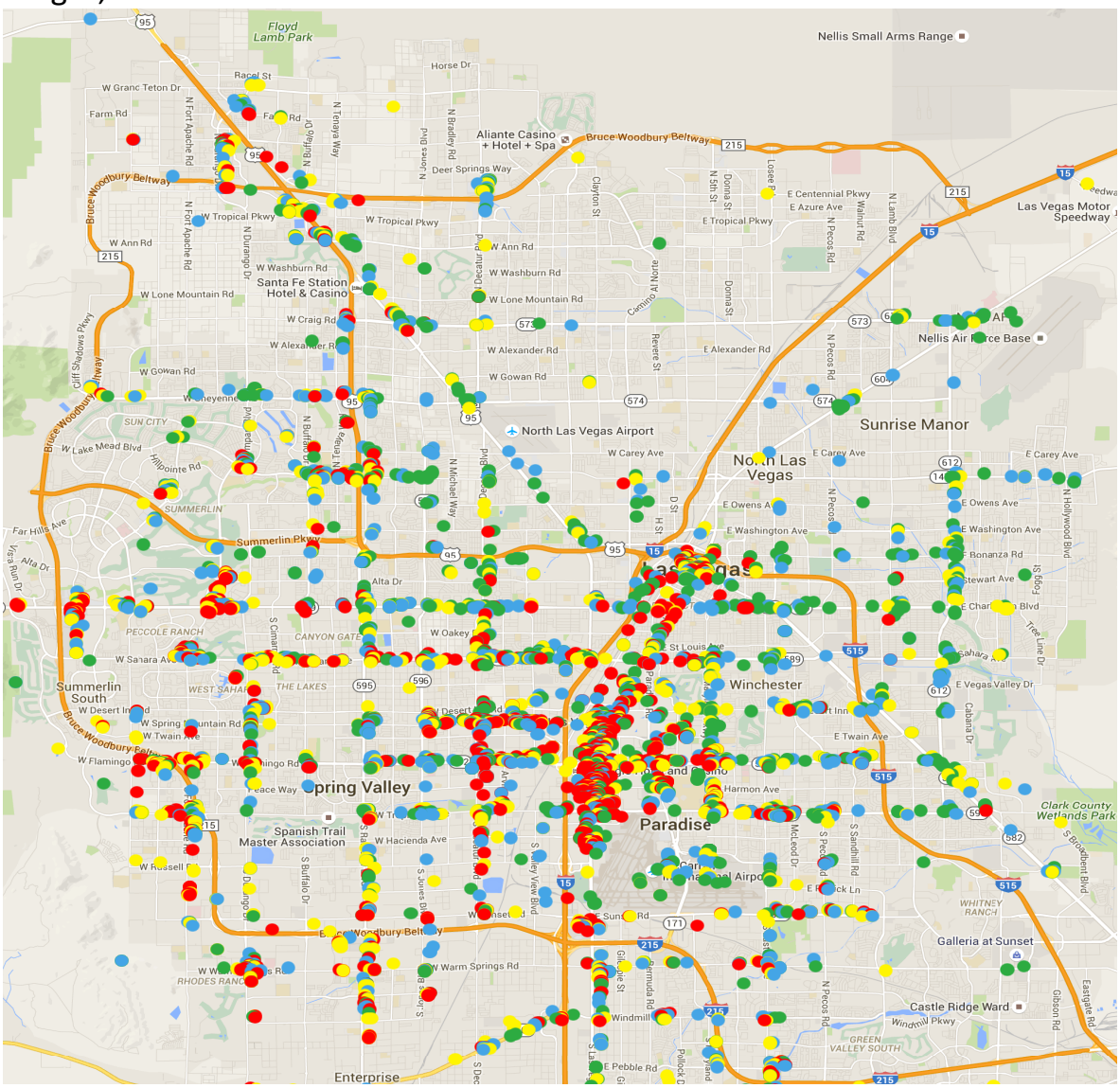


Figure 1 Distribution of Restaurants in Las Vegas, NV

The variogram in Figure 2 shows, in smaller distance there is very high variation in the score of restaurants, overall its randomly distributed throughout the city of Las Vegas.

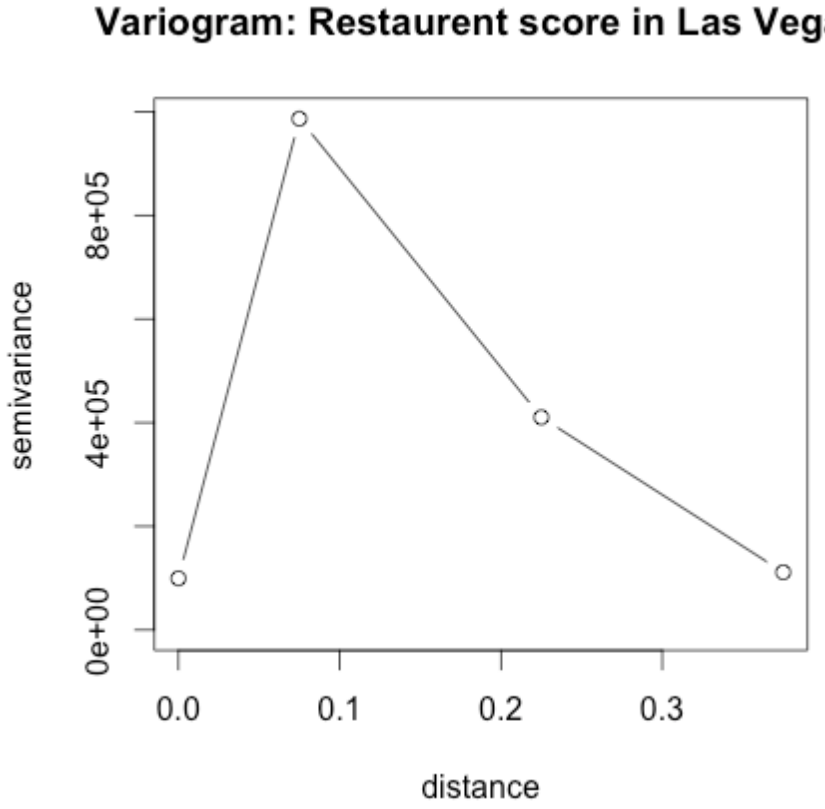


Figure 2 Variogram for Las Vagas Restaurent score

Both the Moran's I and Variogram could not capture the effect of high density of high scoring restaurants along the Strip of Las Vegas. But these restaurants do not represent the average restaurants which has far below score. So we can say there is are random distribution of high scoring restaurant throughout the city. It is noteworthy to emphasize the fact that a red dot represents restaurant score from 329 to 18300 where the average lies around 282.

Las Vegas is not an average city, so it would be interesting to see the difference between Las Vegas and a regular city. So the same analysis was done on the restaurant data of Pheonix, AZ with following results.

Moran's I	Expected	Std. Deviation	P-Value
0.1431702	**-0.0003827019**	**0.02327964**	**6.983667e-10**

Legend for score

	(329,15000]
	(92.5,329]
	(30,92.5]
	[3,30]

Pheonix, AZ

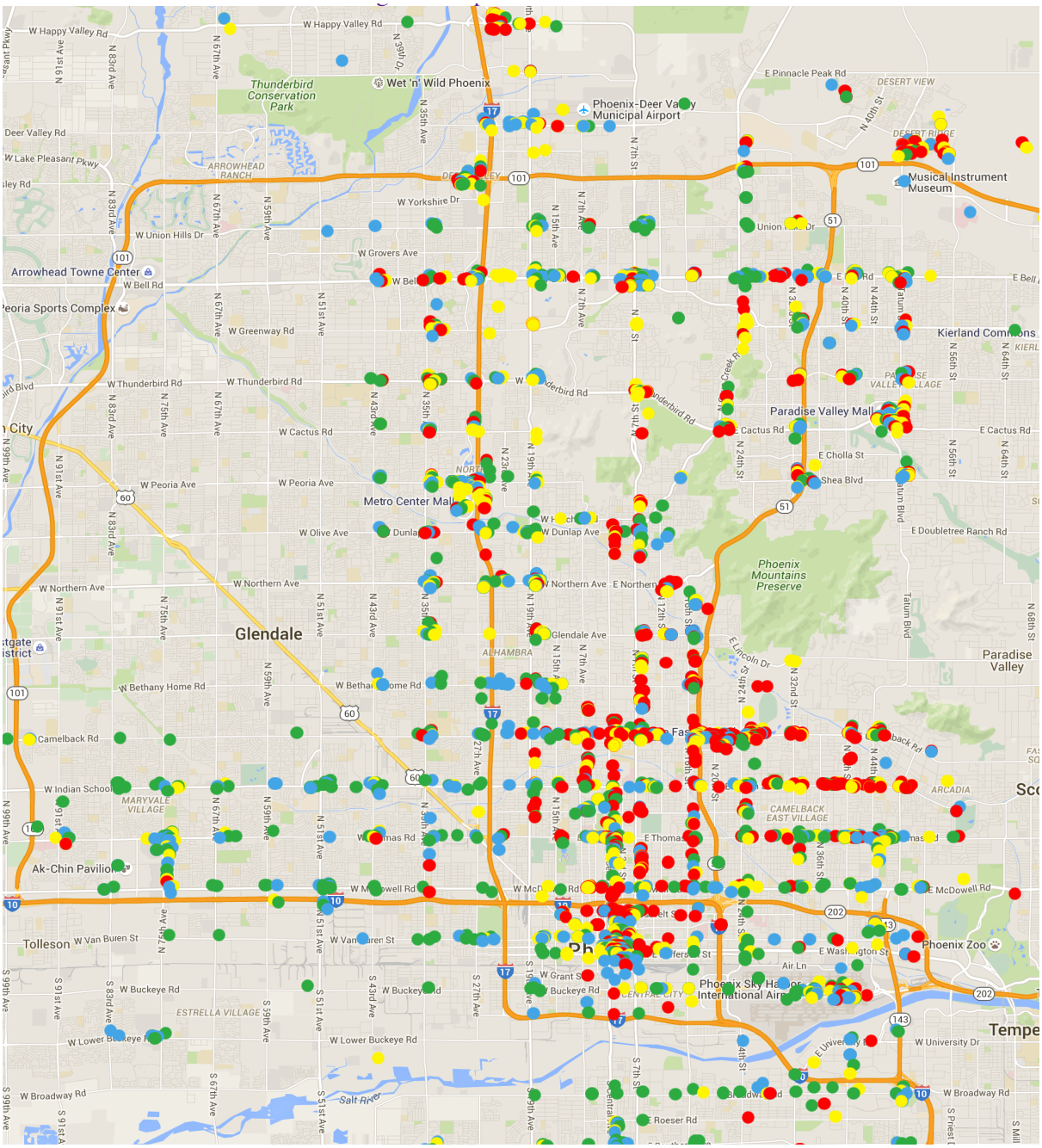
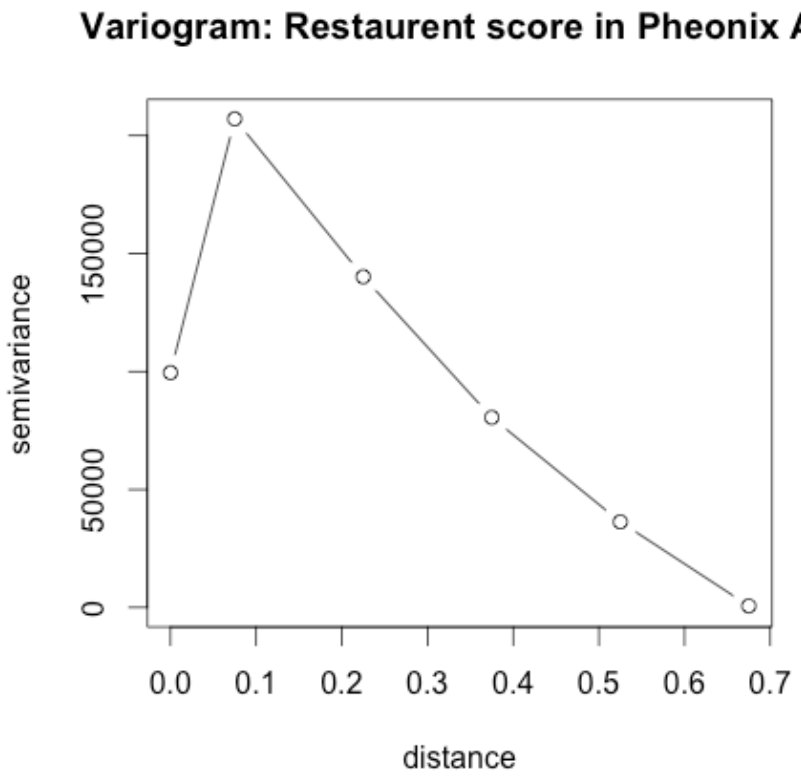


Figure 3 Restaurant score distribution in Pheonix AZ



The obvious factor is that there are lot less red dots in Phoenix than Las Vegas. Which is acceptable, Las Vegas being primarily a Hotel and Casino City. Also the Moran's Index is little higher in Phoenix which is reflected in the map as most red-dots are in the right side of the map but green dots randomly distributed throughout the map.

Also the variogram shows more restaurant score distribution over a much larger geographical area.

Conclusion

This research focused on predicting restaurant success based on the other businesses in the vicinity of the interested restaurant. The Regression Tree based model could predict the restaurant score with certain success, it was obvious from the geospatial analysis that follows, that there are many other factors contributes to restaurant success. From the plotted maps shows clusters of very high scoring restaurants and low scoring restaurants mixed together. So we can infer that low scoring restaurants did not benefit much from being closer to successful restaurants because of competition for limited resources such as parking or having

to pay a high rent which made those restaurants to suffer from lack of economic robustness to do improvement to their business.

Future Work

Future work could include a closer scrutiny on why exactly those restaurants located at a highly populated, commercial region could not succeed. The feature set could be improved to include information about the menu offered or do NLP analysis on user reviews to find out about individual restaurant pros and cons. Finally, this dataset truly reflects the power of big data analysis to improve an industry and make people's life better.

Reference:

[1]. http://daniels.du.edu/directory/haragopal-parsa/
[2]. http://support.esri.com/en/knowledgebase/GISDictionary
[3]. http://www.hamstermap.com/custommap.html