Predicting Food Desert in Brooklyn and Queens, New York

Xintian Li, Yujing Wu, Hanyong Xu 12/15/2019

Introduction

Access to healthy and affordable food can be difficult in many parts of the United States, even in big cities like New York. According to the United States Department of Agriculture (USDA), about 13.5 million Americans have limited access to supermarkets or grocery stores and 82% of them live in urban areas ("USDA ERS - Data Feature: Mapping Food Deserts in the U.S.," n.d.). In order to improve public health and general welfare, it is important to correctly identify food deserts, which is defined as low-income census tracts in which a large amount or percentage of residents have difficulty access to retail outlets selling healthy and affordable food ("USDA ERS - Data Feature: Mapping Food Deserts in the U.S.," n.d.). Through this project, we explored the issue of food deserts within Brooklyn and Queens in New York, United States.

Previously, Michael J.Widener and Wenwen Li explored the relationship between food deserts and references to healthy and unhealthy food in tweets in *Using geolocated Twitter data to monitor the prevalence of healthy and unhealthy food references across the US* (Widener & Li, 2014). Inspired by their work, we looked at food-related tweets generated in Queens and Brooklyn and used them for sentiment analysis. We also used demographic characteristics and distance to food-related amenities to predict the presence of food deserts in Brooklyn and Queens, New York using a Random Forest machine learning model.

Data collection & pre-processing

Food-related tweets

Our first step involves querying tweets generated in Queens and Brooklyn through the Twitter API. We found the place ID for these two locations and used them to query the API. Between December 8th and 11th, we queried the API for 1000 tweets every day and ended with 7717 tweets in total.

We obtained three lists of food names (healthy, unhealthy, fast food restaurants) by scraping websites containing food names. We also brainstormed a list of food-related terms that were neutral (e.g. meals, dinner, etc.). These lists were cleaned and then used to filter tweets in the study area.

Demographic features & distance to amenities

The demographic data came from the American Community Survey (ACS) 2017 and were acquired through census API. Variables queried included population density, racial construct, median household income, households without vehicles. In addition, we scraped public transit data from OpenStreetMaps' API using the osm2gpd package.

We downloaded street network data of Queens and Brooklyn using the Pandana package. Furthermore, we were interested in four types of food-related amenities: fast-food restaurants, market places (weekly or daily exchange of goods and services), convenience stores, and supermarkets. For these amenities, we downloaded the relevant data from the OpenStreetMap API. Using these data, we performed network analysis using the Pandana package to find the distance between each node (street intersection) and the closest amenity for each type of amenities.

Lastly, we acquired food desert data from the USDA website. We spatial joined the public transit data and the distance to amenity data to the census tracts. We summed the number of public transit stations and averaged the distance to amenities at the census tract

level. Finally, we joined the demographic, public transit and distance data to the food desert data.

Data Analysis / Prediction Modeling

Tweet sentiment analysis

We used Textblob to perform a sentiment analysis on the 129 food-related tweets to examine whether each tweet was positive, negative, or neutral and whether they were objective and subjective. There were roughly equal amounts of negative and positive tweets. Most of them were moderately positive or negative and only a few were quite extreme (polarity close to -1 or 1). As for subjectivity, there were a large number of tweets with subjectivity higher than 0.5, equally distributed between negative and positive sentiment categories. In addition, we grouped the tweets into three categories: healthy, unhealthy, and neutral according to the food-related terms included in the tweets. Overall, there were the most tweets in the unhealthy category, followed by the neutral category and lastly the healthy category. Among tweets with polarity below 0, there were the most tweets in the neutral category followed by the unhealthy category. Among tweets with polarity above 0, there were the most tweets in the unhealthy category while the numbers of tweets in the other two categories were roughly the same.

Finally, we isolated all the food-related terms from these tweets and calculated their frequency. We observed no obvious relationship between the category of the food-related terms and their frequency. For example, pizza is unhealthy food and have occurred quite frequently in the tweets. However, healthy foods such as orange and banana have also occurred quite frequently.

Random Forest Classification

We created a Random Forest model with predictors of demographic, public transit, and distance to amenities features to predict whether each census tract is a food desert. We use the population density, percent of different races, median income, number of public transit, population density, percent of non-vehicle households, etc. to predict food desert.

RandomForestClassifier from the library Scikit-learn was used to build a logistic model. Our data have a disproportionate ratio of observations in each class: Within a total of 1430 census tracts in Queens and Brooklyn, only 119 of them are defined as food deserts. To solve the problem of imbalanced data and improve the accuracy of machine learning classification, we tried three methods: oversampling minority class, undersampling majority class, and generating synthetic samples. After comparison, we chose imblearn's SMOTE package to generate new and synthetic data with the nearest neighbor algorithm. Additionally, we use GridSearchCV to do the hyper tuning and find the best parameters for the machine learning model.

The precision of true positive is 0.53, which means in our prediction of food deserts, 53% of the predictions are correct (Precision is the number of true positives divided by all positive predictions.) The recall value of true positive is 0.5, which means we predict 50% of the true food deserts as food deserts (Recall is the number of true positives divided by the number of positive values). In conclusion, our model can predict 95% of the not-food-desert census tracts and 53% of the food-desert census tracts correctly.

Data Visualizations

Tweet sentiment analysis

To visualize the results of the tweet sentiment analysis, we used the library of Altair. We created a scatter plot to assess the distribution of the polarity and subjectivity of

food-related tweets. The scatter plot is also linked to a bar plot that visualizes the count of tweets in each of the three categories: unhealthy, healthy, and neutral through a selection brush. To assess the frequency of food-related terms, we created a bar plot sorted by the frequency.

Demographic features & distance to amenities

We created a choropleth map as well as a histogram for the demographic features. Using the panel library, we allowed the user to choose the demographic features she or he is interested in viewing from a dropdown menu. The user can also switch between different color maps from another dropdown menu. The choropleth map and the histogram were both created with hvplot.

Street network and distance to amenities

We visualized the network data with datashader, hvplot, and altair to create a distribution map and a scatter plot. Again, using the panel library, we are able to allow a user to input the distance, the Nth closest amenity, and the amenity types. The user can also choose a preferred zoom level of the map to visualize a subset of the data in the scatter plot. The resulting map shows two layers of information: first, it shows the network density using a color map with datashader, which transforms the network node points into pixels; secondly, user inputs filter the node points that are within the inputted distance to the selected Nth closest amenity. The scatter plot shows the association of the distance of each node to the two selected amenities.

Prediction result and confusion matrix

In order to see the precision of our predicting model, we visualize the prediction result of all the census tracts next to the food desert map released by USDA with hvplot. Also, we plot a confusion matrix with matplotlib.

Works Cited

USDA ERS - Data Feature: Mapping Food Deserts in the U.S. (n.d.). Retrieved December 15, 2019, from https://www.ers.usda.gov/amber-waves/2011/december/data-feature-mapping-food-deserts-in-the-us/

Widener, M. J., & Li, W. (2014). Using geolocated Twitter data to monitor the prevalence of healthy and unhealthy food references across the US. *Applied Geography*, *54*, 189–197. https://doi.org/10.1016/j.apgeog.2014.07.017