Food Desert CNN : Week 7 Project Milestone

Team Members: Prerna Singh (prerna_singh@berkeley.edu), Yu-Sheng Lee (yushenglee@berkeley.edu), Yiwen Hou (yiwen_hou@berkeley.edu), Kandace Webber (kandywebberlove@berkeley.edu)

Motivation

A food desert is an area with limited access to affordable, nutritious food, often due to a lack of nearby grocery stores. We aim to build a machine learning model that uses satellite imagery to identify food deserts in the Southern U.S. Our goal is to determine whether visual features of neighborhoods provide better or complementary ways to classify areas as food deserts using current USDA guidelines?

This model offers key benefits: Traditional data sources like census surveys and store inventories are slow to update and often incomplete in rural or low-income areas. By detecting visual patterns from satellite imagery, our model can help identify or predict food deserts without relying on fresh local data. This is especially important given past efforts to restrict SNAP benefits, which could worsen food insecurity. The model could also support better city planning and guide nonprofits or policymakers in targeting areas for new grocery stores or food access programs.

Data Description and Preprocessing

1. **USDA Food Access Research Atlas :** https://www.ers.usda.gov/data-products/food-access-research-atlas
   We used this dataset to identify tracts classified as food deserts based on low income/access criteria. This dataset was originally 93,000 census tracts before preprocessing.
2. **US Census TIGER shapefiles :** https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.html
   This dataset includes latitude and longitude coordinates for each census tract, matched from shapefiles to provide precise geographic information.
3. **Satellite Imagery from the Google Maps API :** https://developers.google.com/maps/documentation/maps-static/start
   We used the Google Maps Static API to retrieve satellite images centered on each tract's coordinates, serving as inputs for our convolutional neural network.

We combined these datasets to create our final dataset for the project. After focusing on the Southeastern U.S. and removing missing data, we retained 7,121 census tracts across Tennessee, Georgia, Alabama, Kentucky, South Carolina, Louisiana, Arkansas, and Mississippi.

Since our model will be built using satellite imagery as input to a convolutional neural network (CNN), our feature selection for this phase was limited to variables required for image retrieval: StateName, latitude and longitude, and tract type (urban vs rural). We created a target variable 'IsFoodDesert' which we set as LILA_Urban1_Rural10 = 1.

For outlier detection, we visually inspected that no coordinates were outside of the Southeast US. No outliers were found. Images were checked to ensure that they were all a uniform resolution of 400x400. StateName and Tract_Type were one-hot encoded so that each state in the dataset was converted to a binary indicator column. This will enable our model to interpret state-level information.

To train the CNN, we split the dataset 60/20/20 into training, validation, and test sets. This resulted in 4,273 records in the training set and 1,424 records each in the validation and test sets.

Exploratory Data Analysis (EDA)

We visualized the distribution of StateName, LILA_Urban1_Rural10, and IsFoodDesert. Figure 1 shows that most tracts are urban, but food deserts are disproportionately concentrated in urban areas. The distribution of food deserts confirmed that there is a meaningful split for binary classification.

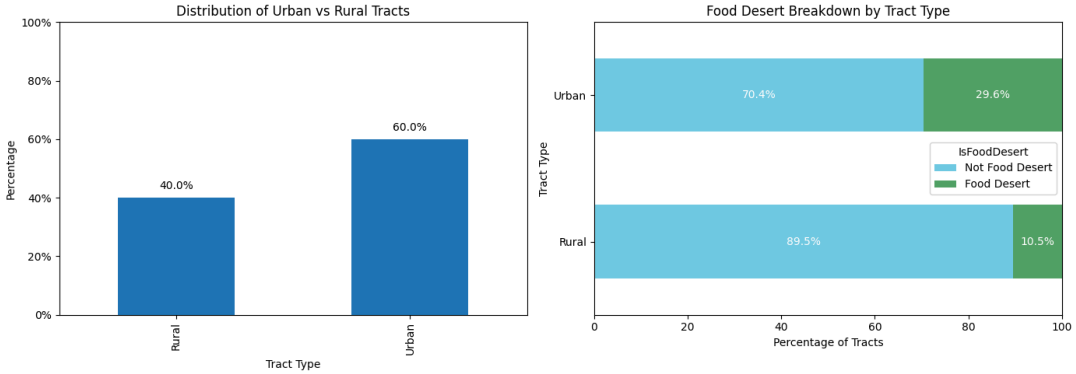

Figure 1: Distribution of Food Deserts by Tract Type

We conducted a correlation analysis of our 'IsFoodDesert' and other numeric features in the dataset. Figure 2 shows strong positive correlations with IsLowIncome and LILA_Urban1_Rural10, consistent with how the label was defined. Other related variables like PovertyRate or VehicleOr20Miles also show some correlations. These could be good candidates for potential future model iterations.

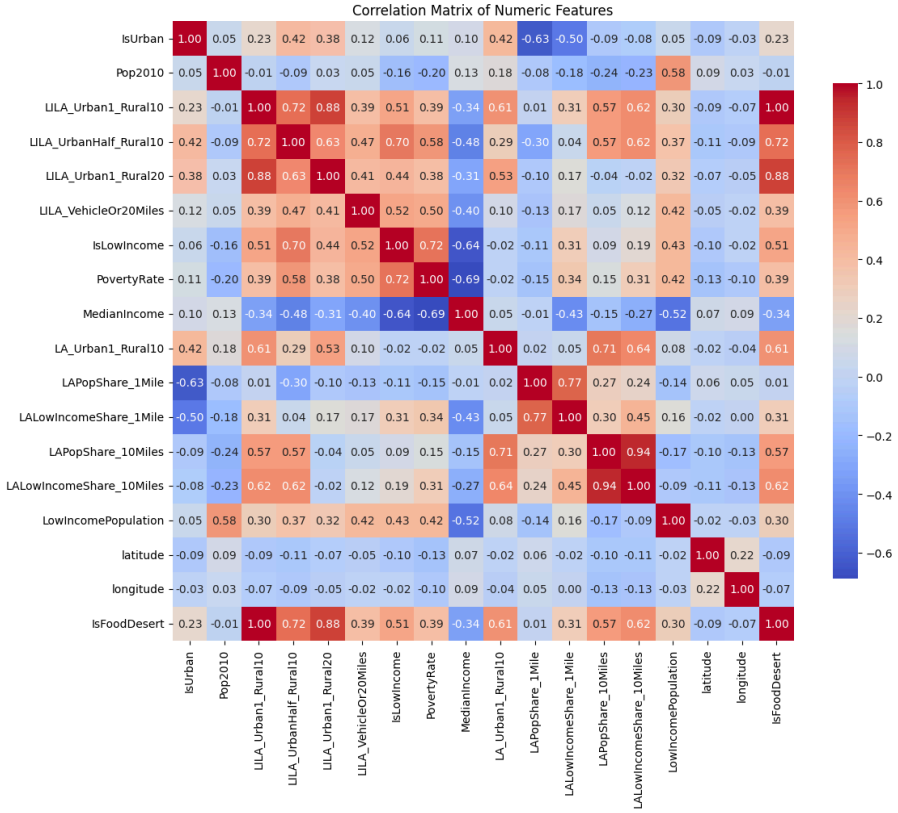

Figure 2: Correlation Matrix of Numeric Features

Food deserts tend to cluster around urban centers on the outskirts. The urban/rural breakdown showed that while most tracts are urban, they also contain a disproportionately high number of food deserts. Figure 3 highlights some spatial patterns we found which suggest geographic grouping, especially in the western parts of states. Additionally, some states (Arkansas, Mississippi) have fewer samples, which may affect model generalization.

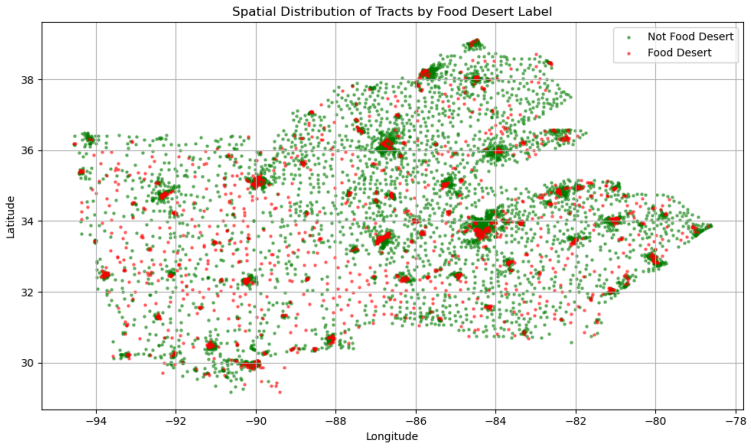

Figure 3: Geographic Spatial Distribution of Food Deserts

Data Challenges

1. ***Undefined Target Variable***: The USDA dataset lacks a predefined target for identifying food deserts. To resolve this, we reviewed academic & industry sources on food access and defined a binary target variable, **IsFoodDesert**, as tracts that are both low income and low access (i.e., LILA_Urban1_Rural10 = 1).
2. ***Temporal Mismatch Between Labels and Imagery:*** The USDA relies on decennial Census data, so some low-income and low-access designations may be outdated. This creates a temporal mismatch with the recent satellite images we pulled via the Google Maps API, potentially introducing label noise into our CNN model. We chose to proceed with the best available labels while acknowledging this limitation.

Methodology

Our data consist of metadata and derived images. For our baseline model, we will use principal component analysis (PCA) for feature selection and then apply logistic regression trained on the USDA metadata. We will exclude the low-income and low-access data, as these features are used to define our target variable. This will help us understand whether the metadata alone can predict whether a census tract is a food desert. We will then use this baseline to compare against our satellite imagery CNN models. Logistic regression was selected because it is simple, fast, and well within our understanding.

The first improvement that we will make will be to create a CNN using solely the satellite images. The only input will be the 400x400x3 RGB images that will result in a binary classification (food desert or not). We will compare these results to our baseline to determine if the visual characteristics alone can predict food access. A CNN was selected as they capture spatial and visual patterns that metadata may lack. We will use Keras to tune learning rate, epoch counts, batch size, and the number of layers to optimize the model.

Our final improved model will be combining our CNN with metadata. We will build off of our CNN and then add in additional features (potentially tract type, income, etc) to see if this performs better than the two previous models. This final model was selected as we believe that both types of data may provide details that the other lacks. We will use the same tuning strategy as our CNN model with additional tweaks to the learning rate and optimizer.

To evaluate which model performs better, we will split our data 60/20/20. The training and validation sets will be used to train and tune the model. The validation set will only be used to test at the end and provide a performance comparison across all three models. We will focus on standard model metrics with particular focus on accuracy and recall. Since the model could be used to determine food assistance eligibility, we want it to be highly accurate overall while also achieving high recall. We want it to correctly identify as many true food deserts as possible. Missing true positives could result in communities being overlooked, which makes recall especially important from a public health and equity standpoint.

Team Contributions

Github : https://github.com/singhprernap/ucb_mids_207_Final_Project_Food_Deserts/tree/main

Prerna Singh: Worked on individual EDA, Added visualization to Combined_EDA, Motivation Section of Writeup and Editted/Formatted Write-Up :
https://github.com/singhprernap/ucb_mids_207_Final_Project_Food_Deserts/blob/main/EDA_individual/Prerna's_eda.ipynb

Kandace Webber: EDA, Write-Up (Data Description, EDA, Challenges, and Methodology):
https://github.com/singhprernap/ucb_mids_207_Final_Project_Food_Deserts/blob/main/EDA_individual/webber_207_ml_food_desert_eda.py

Yu-Sheng Lee: EDA, EDA compilation, Write-Up :
https://github.com/singhprernap/ucb_mids_207_Final_Project_Food_Deserts/tree/clyde/Yusheng_initial_analysis.ipynb

Yiwen Hou: EDA, Write-Up :
https://github.com/singhprernap/ucb_mids_207_Final_Project_Food_Deserts/blob/main/EDA_individual/EDA_Yiwen.ipynb