Food Bank Allocation Using Integer Programming

Helena Aytenfisu, Eban Ebssa, Lucy Zimmerman

CS 221 — December 04, 2024

Project Video: https://youtu.be/pMlYpWeLuDg

Kaggle Dataset:
https://www.kaggle.com/datasets/tcrammond/food-access-and-food-deserts?resource=download

U.S. Census Bureau Dataset (map visualization data):
https://www.census.gov/cgi-bin/geo/shapefiles/index.php?year=2010&layergroup=Census%2BTracts

Code Repository: https://github.com/lucyzimmo/Food-Planning-Model

ArcGIS Interactive Map Visualization:
https://stanford.maps.arcgis.com/apps/mapviewer/index.html?webmap=6fbbc4ee8d4b401a9129378382ab82df

Introduction

Food insecurity, defined as the lack of sufficient access to food, is a serious global issue (Haskell). Food deserts are regions marked by a lack of access to healthful food options that are affordable (Haskell). In 2021, it was estimated that in the state of California alone, there are 4.1 million people who are food insecure (Data Commons). Additionally, between 2021 and 2023, it was estimated that 11.4% of the population of California was food insecure (Rabbitt et al., USDA). While in this same three year period the prevalence of food insecurity in California was lower than the national average, there were significant increases observed in the prevalence of food insecurity and very low food security in California (Rabbitt et al., USDA). Within the state of California, the highest prevalence of food insecurity (17% as of 2021) has been observed in Imperial County, located in California's southeast border (Data Commons). In order to help address the issue of food insecurity, our aim is to determine how to best distribute food banks among areas of high food insecurity to serve those in most need and to service as many people as we can. In order to do so, we will be modeling the task as a constraint satisfaction problem where we will be building a fixed number of food banks, the variables are the various locations (identified by their census tract), and their respective domains are the number of food banks that will be built in that location, which we elaborate on in the modeling section. We will be using data for Imperial County, CA, considering its relevance to the issue of food insecurity.

Literature Review

Two key works relevant to our approach are those by Mercado (2018) and Ruan et al. (2024), both of which focus on improving the allocation of food banks. In Calgary, Mercado (2018) applied a K-means clustering algorithm to improve the allocation of food bank depots. Mercado recorded food hamper collection in households for five months, then ran the K-means algorithm 100 times to determine the best (smallest variance) clusters. They compared the AI locations to actuals, noting that public transportation access may be less relevant. Ruan, Guo, and Lin (2024) built upon Mercado's work by developing a two-level optimization framework that combines a pseudo-weighted K-Medoids clustering algorithm with the Open-Source Routing Machine (OSRM) engine. Unlike the Euclidean distance metric used in K-means, their approach incorporates real road distances, which is particularly relevant for optimizing food bank locations in areas with complex road networks. They applied this method to 3 million household residence locations in Indiana, sourced from GIS data, and randomly sampled 6,293 households to calculate road distances via OSRM. The K-Medoids algorithm was applied twice: first to identify optimal food bank locations, and then within each cluster to determine the best pantry locations.

These approaches are complementary to ours but differ from ours in data and methodology. As described in the next section, we use census tract-level data, allowing for greater scalability to multiple counties, as well as incorporating population, addressing the assumption that all data points are of equal importance. We may draw on Ruan's work in future iterations of this project if we consider the road distance between tracts, or consider placing food pantries in addition to food banks. The theoretical foundation of our work is in integer programming, where we satisfy multiple constraints while also optimizing for a specific objective function (integers because the number of food banks in one tract is discrete). De Turck (2020) illustrates how integer programming can be applied to resource allocation problems, such as distributing software programs across a set of computers in an energy-efficient manner. Our study will build upon these methods by using census tract-level data and leveraging integer programming to ensure that food resources are allocated efficiently and effectively to areas of highest need.

Dataset

The dataset we are utilizing is a dataset sourced from Kaggle that has over 70,000 data points organized by census tract. For each census tract, there is information about the state and county that the region identified by the census tract falls in, as well as information about the demographics of the region, median household income, the poverty rate, the low-income population (aggregate as well as demographic breakdown), information about vehicle access, those receiving SNAP benefits, and more. The data is available in CSV format, so there is virtually no preprocessing needed (aside from isolating the counties/regions of interest).

Baseline

The baseline for this project involves a random allocation of food banks across census tracts in Imperial County, California. Rather than using demographic or geographic data to inform the placement of food banks, this baseline randomly distributes a set number of food banks among census tracts. We generate a random allocation of food banks using a Dirichlet distribution, which assigns random weights to each tract that sum to 1. These weights are scaled to the total number of food banks, which we set at 100.

To evaluate the baseline, we use three main metrics: low-income household coverage, population coverage, and geographic coverage.

Low-income household coverage measures the total population of the county eligible for SNAP benefits that is served by tracts that have food banks. This metric addresses the primary need for food access by ensuring that a larger share of people can easily reach food resources.

Low-Income Household Coverage = $\frac{SNAP-eligible\ population\ of\ tracts\ with\ at\ least\ one\ food\ bank}{Total\ SNAP-eligible\ population\ of\ county} \times 100$

Population coverage, on the other hand, measures the total population served by tracts that have food banks. This metric addresses the primary need for food access by ensuring that a larger share of people can easily reach food resources.

Population Coverage = $\frac{Population\ of\ tracts\ with\ at\ least\ one\ food\ bank}{Total\ population\ of\ county} \times 100$

Geographic coverage measures the percentage of census tracts that receive at least one food bank. This metric is important because it aligns with the goal of providing broad access across as many areas as possible, meaning more tracts with food banks result in better overall coverage.

Geographic Coverage = $\frac{Total\ tracts\ with\ at\ least\ one\ food\ bank}{Total\ number\ of\ census\ Tracts} \times 100$

Main approach

We will be using constraint satisfaction, specifically the integer programming model, where we define a set of discrete variables and corresponding constraints, as well as a linear objective function. The definitions of our integer programming model are given below.

Variables: Individual census tracts (numeric identifier for a geographic region provided by the U.S. census bureau) for all the regions falling within Imperial County, CA, of which there are 31 total in the data set. Thus,
$X = (6025010101, 6025010102, 6025010200, 6025010300, 6025010400, 6025010500, 6025010600, 6025010700, 6025010800, 6025010900, 6025011000, 6025011100, 6025011201, 6025011202, 6025011300, 6025011400, 6025011500, 6025011600, 6025011700, 6025011801, 6025011802, 6025011803, 6025011900, 6025012001, 6025012002, 6025012100, 6025012200, 6025012301, 6025012302, 6025012400, 6025940000)$

Domain: We have $Domain_i \in [0, 1, 2, ..., n]$, where $n = TotalNewFoodbanks$. That is, the assignment $6025010101 = 2$ indicates that there will be 2 new food banks built in the first census tract which has ID 6025010101. For the purposes of our main implementation, we set $n = 100$.

Constraints:

1. No more than Z food banks may be placed in two adjacent census tracts. Formalized, this is $f_1(X) = adjacent(tract_i, tract_j) => (sum(x_i, x_j) < Z)$. For the purposes of our main implementation, we set $Z = 4$.
2. Our variables must sum to n. That is, $f_2(X) = x_1 + x_2 + x_3 + ... + x_i = n,$ where again, $n = 100$ in our implementation.

Objective (function):
Formalized as:

$$F(x_1, ..., x_n) = \sum_{1}^{n} (count(lowIncome_{x_i}) + count(population_{x_i}) - variance(income_{x_i}))$$

. Aim:

1. Maximize the amount of low-income households served
2. Maximize the amount of people served overall
3. Minimize the variance in income among tracts

Evaluation Metrics

We plan to evaluate using a few metrics. We will use the population coverage metric from the baseline evaluation. In addition, we will be evaluating the coverage of low-income households through calculating the percentage of low-income households (L) in a census tract (i) served by a food bank within the specified distance:

$$\frac{\sum_{i=1}^{n} L_i}{L} \times 100$$

Furthermore, to measure the equity in food bank distribution across income levels with the aim to prioritize low-income areas, we will be measuring the variance of the median household income (I) in tracts (i) with food banks where N is the set of tracts with at least one food bank. Lower variance indicates a more balanced distribution.

$$= \frac{1}{|N|} \sum_{i \in N} (I_i - \bar{I}_N)^2$$

For qualitative metrics, we will look at whether food banks are placed in accessible locations (e.g., near public transit routes).

Results & Analysis

The baseline allocation achieved **57.87% low-income household coverage**, **77.79% population coverage**, and **83.87% geographic coverage** reflecting the limitations of a random allocation model in equitably serving the populations most in need for the sake of the general population. For our experimental model allocation, we used the Constraint Satisfaction Problem model along with integer programming, specifically the framework PULP. This approach significantly improved **low-income household coverage (99.23%)**, **population coverage (96.60%)**, and **geographic coverage (96.77%)**, demonstrating the effectiveness of a targeted, data-driven allocation strategy.

Table 1. Comparison of Baseline Random Allocation Model and CSP & IP Experimental Model

Metric	Baseline Model (Random Allocation)	Experimental Model (CSP + IP)
Low-Income Household Coverage	57.87%	99.23%
Population Coverage	77.79%	96.60%
Geographic Coverage	83.87%	96.77%

Our constraints were effective in that adjacency constraints ensured that food banks are allocated in ways that maximize access across neighboring tracts, while income-based constraints prioritized low-income areas. Geographic constraints ensured that food banks were located within a reasonable distance of underserved populations. The improvements in these sectors demonstrate the model's ability to effectively direct resources to tracts with the greatest need while still balancing distribution.

Error Analysis

In our experimentation, we also tested out a proportional allocation system whereby the number of food banks assigned to each tract was directly proportional to the population of that particular tract. In almost all cases, the proportional implementation did extremely well in terms of evaluation. It achieved 99.40%, 98.74%, and 93.55% success for low-income household coverage, population coverage, and geographic coverage, respectively. This could be a result of an underlying association between high population tracts and populations with high need. As observed in the diagrams below, the tracts with higher relative population also have more residents receiving SNAP benefits.

Figure 1. Imperial County Population 2010 by Tracts

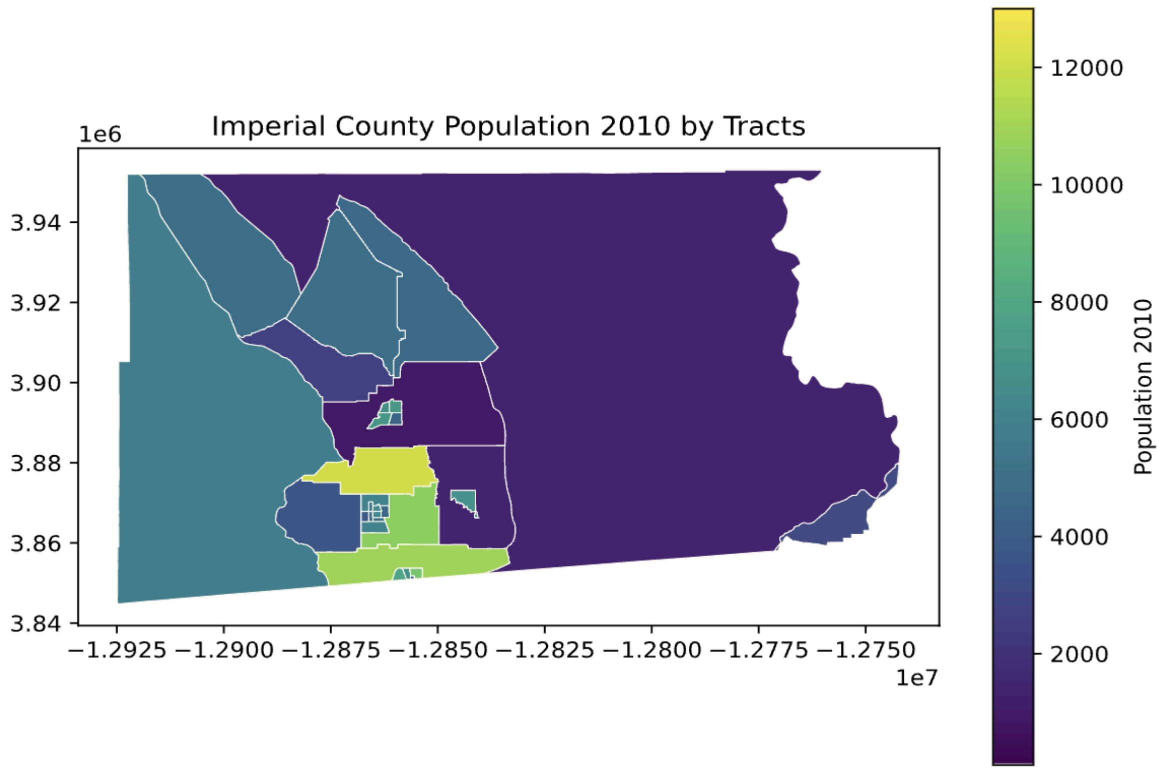
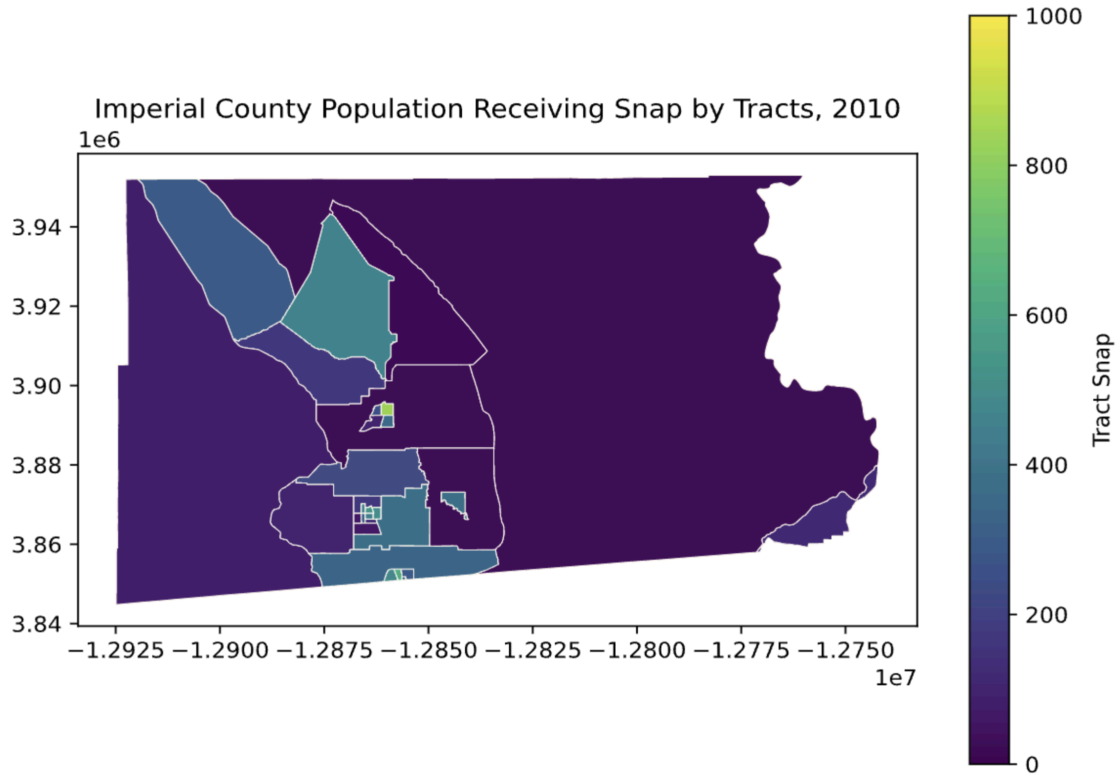


Figure 2. Imperial County Population Receiving SNAP by Tracts, 2010



The population receiving SNAP benefits is the metric we used in our objective function to determine the level of need. Thus, a proportional distribution model is able to address both the population objective and the low income objective simultaneously.

Future Work

In future implementations, we would like to work on expanding our coverage from solely Imperial County, CA, to additional counties in California and across the United States. Further, the task we are addressing is determining how many food distribution sites to place within each census tract in a particular county. However, there is further work to be done in deciding where these distribution sites would be placed in a county to be most impactful. This would require more data, since the dataset is organized by census tracts and we don't have information on a more granular level.

Ethical Considerations

One key ethical consideration is the ability for our model to make decisions solely based on data without sufficient human intervention.While data-driven models like ours can optimize food bank allocation efficiently, they lack the nuanced understanding of local contexts and human factors. For example, a tract with a high population density and low-income households might be prioritized by the model, but it may not account for other critical factors such as local infrastructure limitations, community preferences, or cultural dynamics. A consideration of on-the-ground realities is necessary, as a reliance on algorithmic decision-making is dangerous. Partnering with local community stakeholders to gain qualitative insights beyond census data would be imperative to understanding the problem and building community trust.

Additionally, there is an ethical principle of distributive justice which we try to mitigate inherently in our experimental model. While our baseline model allocates food banks to different census tracts randomly, our experimental model designs constraints and an objective in such a way that the model prioritizes individuals who are the worst off. The proportional model we experimented with, however, has an underlying principle of maximizing overall well-being.

References

Data Commons. (n.d.). *California • Food Security*. Home - Data Commons. https://datacommons.org/explore#q=food%20deserts%20in%20california&dc=

De Turck, F. (2020). Efficient Resource Allocation through Integer Linear Programming: a detailed example. arXiv preprint arXiv:2009.13178.

Haskell, S. (2021, February 11). *Food insecurity and food deserts: How are they related?*. Institute for Food Laws and Regulations. https://www.canr.msu.edu/news/food-insecurity-and-food-deserts-how-are-they-related

Mercado, J. (2018). "Using machine learning to find optimal locations for food bank depots". 13th Annual Students' Union Undergraduate Research Symposium, December 6, 2018. University of Calgary, Calgary, AB.

Rabbitt, M. P., Reed-Jones, M., Hales, L. J., & Burke, M. P. (2024, September). *Household Food Security in the United States in 2023*. USDA ERS - Publications. https://www.ers.usda.gov/webdocs/publications/109896/err-337.pdf?v=8208.3

Ruan, G., Guo, Z., & Lin, G. (2024). *Where to build food banks and pantries: A two-level machine learning approach*. arXiv. https://arxiv.org/abs/2410.15420

United States Government. (n.d.). *2010 tiger/line® shapefiles: Census+tracts*. United States Census Bureau. https://www.census.gov/cgi-bin/geo/shapefiles/index.php?year=2010&layergroup=Census%2BTracts