Food Deserts and k-Means Clustering

Garrett Kepler[†] and Maria Palomino [‡]

Project advisor: Andrea Arauza Rivera [§]

Abstract. Food deserts are regions where people lack access to healthy foods. In this article we use k-means clustering to cluster the food deserts in two Bay Area counties. The centroids (means) of these clusters are optimal locations for intervention sites (such as food pantries) since they minimize the distance that a person within a food desert cluster would need to travel to reach the resources they require. We present the results of both a standard and a weighted k-means clustering algorithm. The weighted algorithm takes into account the poverty levels in each food desert when determining the placement of a centroid. We find that this weighting can make significant changes to the proposed locations of intervention sites.

1. Introduction.

Access to healthy foods is a fundamental human right, yet healthy foods remain inaccessible to thousands of people throughout the country. Local and federal entities study the lack of access to healthy foods in various communities and seek both temporary and lasting solutions. Food banks, food pantries, and resource pantries are examples of temporary attempts at addressing the lack of access to healthy foods. More longterm solutions include opening new grocery store locations or community gardens. This work provides a framework for choosing the location(s) of either temporary or lasting intervention sites.

Regions with a lack of access to healthy foods are often known as *food deserts*, and entities such as the U.S. Department of Agriculture (USDA) have precise definitions for what it means for an area to be a food desert.

According to the USDA, food deserts are defined by a lack of access to supermarkets, grocery stores, or other sources of healthy and affordable foods [6]. To more precisely define what we mean by "access" we combine measures of income with measures of proximity to sources of healthy foods. The US census subdivides counties into what are known as *census tracts*. A census tract is then designated a food desert by the USDA based on it being identified as a low income and a low access area. A census tract is a low income area if it satisfies any of the following:

- The census tract's poverty rate is 20% or greater;
- The census tract's median family income is less than or equal to 80% of the state-wide median family income;
- The census tract is in a metropolitan area and has a median family income less than or equal to 80% of the metropolitan area's median family income.

We can combine the above with measures of proximity to sources of healthy foods to arrive at the definitions of three types of food deserts. These definitions are those used by the USDA and are described in [6].

Half Mile Food Deserts. Low-income census tracts where a significant number (at least

[†] California State University, East Bay (gkepler@horizon.csueastbay.edu).

[‡] California State University, East Bay (mpalomino2@horizon.csueastbay.edu).

[§] California State University, East Bay (andrea.arauzarivera@csueastbay.edu).

500 people) or share (at least 33 percent) of the population is greater than one-half mile to the nearest supermarket or grocery store for an urban area or greater than 10 miles for a rural area.

1 Mile Food Deserts. Low-income census tracts where a significant number (at least 500 people) or share (at least 33 percent) of the population is greater than 1 mile to the nearest supermarket or grocery store for an urban area or greater than 10 miles for a rural area.

1 and 20 Mile Food Deserts. Low-income census tracts where a significant number (at least 500 people) or share (at least 33 percent) of the population is greater than 1 to from the nearest supermarket or grocery store for an urban area or greater than 20 miles for a rural area.

Our focus is on the San Francisco Bay Area, and in particular on the counties that make up the East Bay. For this reason we will primarily focus on the half mile and 1 mile food deserts in urban areas defined above. The San Francisco Bay Area, is notorious for it's high housing prices and the large divide between the wealthiest and poorest families. For example, in Alameda county (the most populated county in the East Bay with nearly 1.7 million residents), 21.7% of households have an income of over $200,000 while 24.8% of households have an income under $50,000 [3].

In this article we look to leverage mathematical tools to address the following question:

Main Question. Given the location of food desert in a county, how can we optimally place intervention sites (e.g. food pantries) to benefit the most people?

The optimality of our approach is based in the use of an unsupervised learning algorithm known as k-means clustering. This algorithm takes as input a positive integer k and a set of data, and produces k clusters of that data along with a centroid for each cluster. In our analysis we perform k-means to cluster food deserts and compute the centroid of these clusters. A centroid serves as the ideal location for an intervention site (such as a food pantry) because it minimizes the distance from each food desert in the cluster to the centroid. This creates better access for those living far from a grocery store or other source of healthy food. The k-means algorithm includes an analysis of the proper choice of k; for this we use what are known as *elbow plots*. We describe these in detail in Section 2.

To address our main question, we must also quantify what we mean by "benefit the most people". Our initial k-means analysis gives equal weight to each food desert when generating the centroid of the cluster. Our subsequent analysis puts additional weight on food deserts with larger percentages of low-income households. In Figure 7 one can see the locations for intervention sites suggested by our unweighted and weighted k-means clustering for half mile food deserts in Alameda county.

The remainder of this paper is organized as follows:

- Section 2 gives an introduction to the k-means clustering algorithm as well as the use of elbow plots in determining the best value for k. In this section we give a number of important definitions and establish some needed notation.
- Section 3 explains the results of our k-means analysis (both with and without weightings) for two counties in the Bay Area–Alameda County and Contra Costa County. This section addresses the main question posed above.

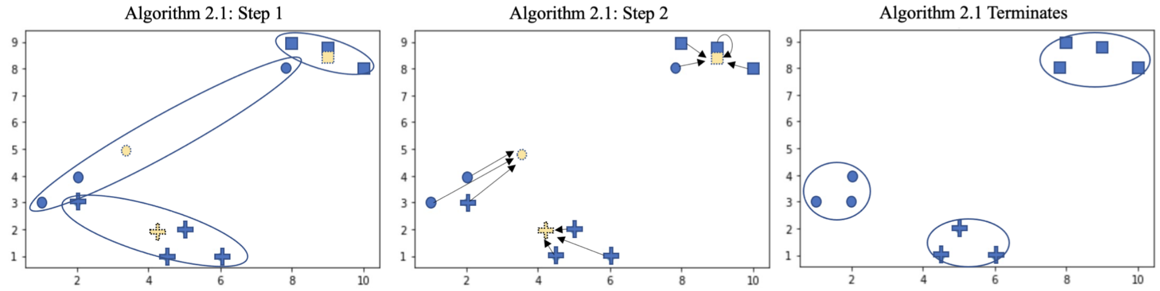

Figure 1. *(Left) A sample data set (blue) that has been randomly assigned to 3 clusters (circles, pluses, and squares) along with the centroids (yellow) for each cluster. (Middle) Each data point indicates which centroid is closest to it. This reconstructs the clusters. (Right) The algorithm terminates when the clusters no longer change.*

- We conclude in section 4 with some future directions for this work.

2. The k-means algorithm.

There are many techniques in mathematics and statistics used to understand (or "learn") data, and they are often classified as *supervised* or *unsupervised*. Supervised learning algorithms/techniques often involve building a model to predict or estimate a quantity based on some set of inputs. For example, when we address a problem by starting with a set of data, plotting the line which most closely approximates that data, and use the line to predict future values or fill in missing values, we are performing a supervised learning technique. The word supervised refers to the fact that we have a clearly specified goal/output (predicting, estimating, filling in values).

Unsupervised learning on the other hand involves algorithms/techniques that aim to understand a set of data, but have no predetermined output. Unsupervised learning techniques include for example clustering algorithms or dimensionality reduction techniques like principal component analysis. Some helpful references on both supervised and unsupervised learning algorithms include [5] and [4]. In this paper we will use the k-means clustering algorithm which is an example of an unsupervised learning algorithm.

2.1. Description of the k-means algorithm.

We begin with a set of data,

$$X = \{\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}\}.$$

For our purposes it will be sufficient to consider $X \subseteq \mathbb{R}^2$; however, this is not a necessary restriction when apply the k-means algorithm. In this work each $\mathbf{x_i} \in \mathbb{R}^2$ encodes the location of a food desert.

Next, consider indexing subsets $C_i \subseteq \{1, 2, \ldots, n\}$ where $i \in \{1, 2, \ldots, k\}$. The sets $C_1, C_2, \ldots, C_k$ tell us how to cluster the points in X. We require that

1. $C_1 \cup C_2 \cup \cdots \cup C_k = \{1, 2, \ldots, n\}$ so that every data point is in at least one cluster, and
2. for any $i \neq j$, $C_i \cap C_j = \emptyset$ so that every data point is in exactly one cluster.

We want clusters that minimize the distances between the points in the cluster. For a given cluster C_r, the average variation between the points in the cluster is given by

$$AV(C_r) = \frac{1}{|C_r|} \sum_{i \in C_r} \sum_{j \in C_r} d^2(\mathbf{x_i}, \mathbf{x_j})$$

where $|A|$ denotes the number of elements in the set A and $d^2((a_1, a_2), (b_1, b_2)) = (a_1 - b_1)^2 + (a_2 - b_2)^2$ is the square of the Euclidean distance between points in $\mathbb{R}^2$.

We now want to minimize the sum of the quantities $AV(C_r)$ over all possible choices of $C_1, C_2, \ldots, C_k$. In other words, we want to solve the following minimization problem

$$\underset{C_1, C_2, \ldots, C_k}{\text{minimize}} \sum_{r=1}^{k} AV(C_r). \tag{2.1}$$

Next, we describe how we choose the sets $C_1, C_2, \ldots, C_k$ to solve this minimization problem.

We will write $\mathbf{x_i} = (x_{i1}, x_{i2})$ and the mean, or **centroid**, for the cluster C_r is given by

$$\mathbf{c_r} = (c_{r1}, c_{r2}) = \frac{1}{|C_r|} \sum_{i \in C_r} \mathbf{x_i}$$

Notice

$$\begin{aligned}
AV(C_r) &= \frac{1}{|C_r|} \sum_{i \in C_r} \sum_{j \in C_r} d^2(\mathbf{x_i}, \mathbf{x_j}) \\
&= \frac{1}{|C_r|} \sum_{i \in C_r} \sum_{j \in C_r} \left((x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2\right) \\
&= \frac{1}{|C_r|} \sum_{i \in C_r} \sum_{j \in C_r} \left((x_{i1} - c_{r1} + c_{r1} - x_{j1})^2 + (x_{i2} - c_{r2} + c_{r2} - x_{j2})^2\right) \\
&= \frac{1}{|C_r|} \sum_{i \in C_r} \sum_{j \in C_r} \left((x_{i1} - c_{r1})^2 + (x_{j1} - c_{r1})^2 + (x_{i2} - c_{r2})^2 + (x_{j2} - c_{r2})^2\right) \qquad (2.2) \\
&= \sum_{i \in C_r} \left((x_{i1} - c_{r1})^2 + (x_{i2} - c_{r2})^2\right) + \sum_{j \in C_r} \left((x_{j1} - c_{r1})^2 + (x_{j2} - c_{r2})^2\right) \\
&= 2 \sum_{i \in C_r} \left((x_{i1} - c_{r1})^2 + (x_{i2} - c_{r2})^2\right) \\
&= 2 \sum_{i \in C_r} d^2(\mathbf{x_i}, \mathbf{c_r})
\end{aligned}$$

where we use in line 2.2 that

$$\begin{aligned}
\frac{1}{|C_r|} \sum_{i \in C_r} \sum_{j \in C_r} 2(x_{i1} - c_{r1})(c_{r1} - x_{j1}) &= \sum_{i \in C_r} 2(x_{i1} - c_{r1}) \frac{1}{|C_r|} \sum_{j \in C_r} (c_{r1} - x_{j1}) \\
&= \sum_{i \in C_r} 2(x_{i1} - c_{r1})(c_{r1} - c_{r1}) \\
&= 0
\end{aligned}$$

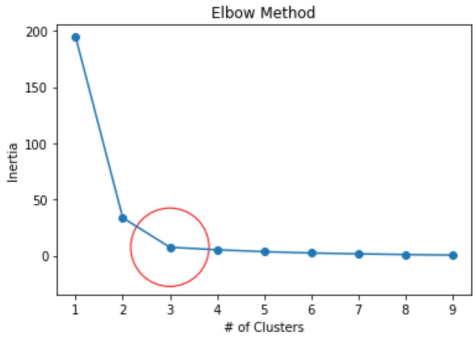

Figure 2. *For the example in Figure 1, we graph the inertia of the clusters vs. the number of clusters. Observe a plateau in the inertia following 3 clusters.*

and similarly

$$\frac{1}{|C_r|}\sum_{i\in C_r}\sum_{j\in C_r} 2(x_{i2}-c_{r2})(c_{r2}-x_{j2}) = 0.$$

This tells us that the clusters $C_1, C_2, \ldots, C_k$ which minimize the average variation over all clusters are the same as those that minimize the quantity below which is often called the **inertia** of the clusters:

$$\text{Inertia} = \sum_{r=1}^{k}\sum_{i\in C_r} d^2(\mathbf{x_i}, \mathbf{c_r}) \tag{2.3}$$

We can now describe an algorithm which give us a (local) solution to this minimization problem.

Algorithm 2.1.
Input: a data set X and an integer $k \geq 2$.
Output: sets $C_1, C_2, \ldots, C_k$ which minimize the average variation over all clusters.

1. First, randomly assign an integer from 1 to k to each point in X. This is our initial clustering of the data. For each cluster, C_r, compute the centroid $\mathbf{c_r}$. This gives centroids $\mathbf{c_1}, \mathbf{c_2}, \ldots, \mathbf{c_k}$.
2. Reassign points in X so that a point $x \in X$ is in cluster C_r if out of all centroids, x is closest to $\mathbf{c_r}$.
3. For each cluster recompute the cluster centroid, then repeat step 2. Termination will occur when the configuration of the clusters no longer changes.

The algorithm above will cluster the data in X in a way which minimizes the average variation over all clusters; see Figure 1 for an illustration of the steps in the k-means algorithm. Note that this process gives a *local* solution in the sense that different initial random assignments in step Alg 1. may yield different clusters with different values for the quantity (2.3). In practice, one can run the k-means algorithm a number of times and choose the clustering which gives the smallest value for (2.3).

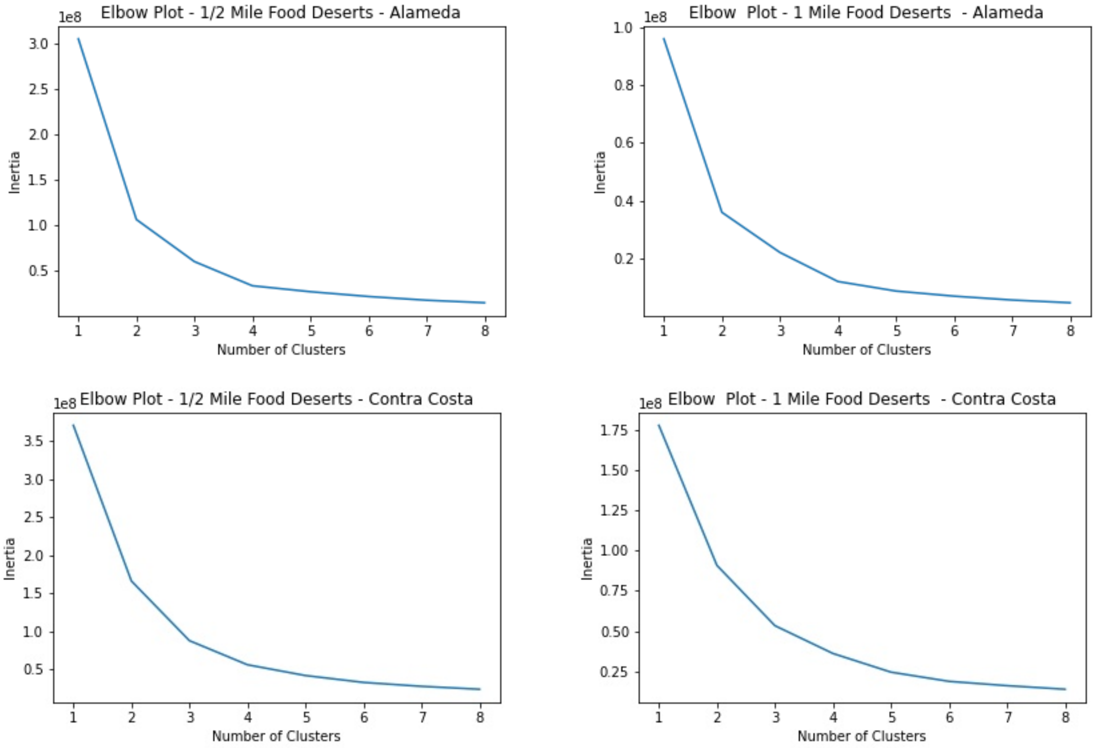

Figure 3. *Elbow plots for half mile (left) and 1 mile food deserts (right) in Alameda (top) and Contra Costa (bottom) county. Plots are generated with the standard unweighted k-means algorithm.*

What we have described above is the standard k-means algorithm. In our analysis of food desert data, we will also use a *weighted* k-means algorithm. This means that we will add weights to individual data points. The k-means algorithm will then prefer a higher weight data point over a lower weight data point and adjust the calculations of the centroid in favor of the heavier data points. This is described in greater detail in Section 3.

The final critical question to answer before we apply the k-means algorithm is about how we choose the value for k.

2.2. Elbow plots. One of our objectives is to carefully choose the number of clusters k. Using larger and larger values for k will create smaller average variation (and inertia) within the clusters; however, in our food deserts application we may have to limit the number of intervention sites we can place in a county. Resources for a food pantry may be limited. (Note that when $k = n = |X|$, each cluster contains exactly one point from X and so the average variation within the cluster is 0.)

We want to carefully choose the number of intervention sites to sustainably alleviate food insecurity, so we will use what is known as an elbow plot. An **elbow plot** shows the relationship between different values of k and the inertia in (2.3) which we wish to minimize. As noted above, when $k = n = |X|$, the value of the elbow plot will be 0. In our analysis we will produce an elbow plot for a given data set X, and use the "elbow" k_0 value in our

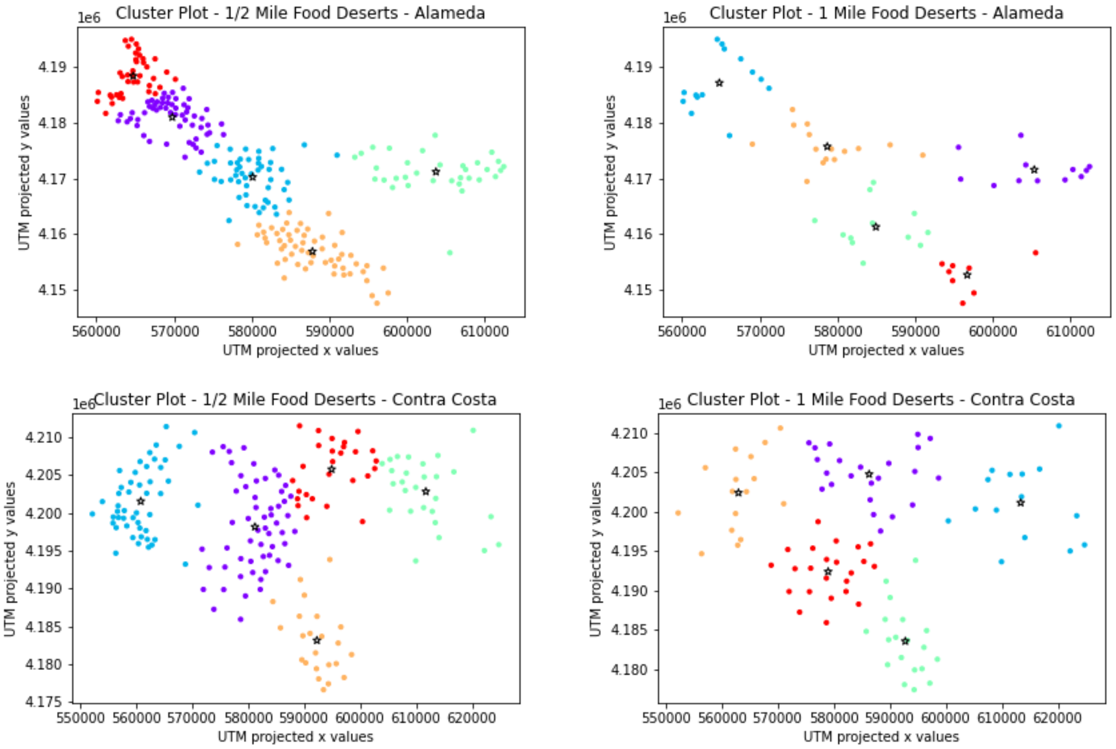

Figure 4. *Cluster plots for half mile (left) and 1 mile food deserts (right) in Alameda (top) and Contra Costa (bottom) county. Plots are generated with the standard unweighted k-means algorithm. Stars indicate the location of the cluster centroids.*

clustering. Elbow plots are also sometimes called *inertia plots*. See Figure 2 for the elbow plot associated to the clustering example in Figure 1.

In some cases, it is unclear what the correct k_0 value should be. There are various methods used to address this such as calculating the *silhouette* value for each cluster. The silhouette value measures "tightness and separation" within a cluster and can be applied to identify k_0; see [7] for further details. This technique is especially useful when clustering is being used to classify data points. For our application to food desert clusters we found that the inertia with $k = 5$ would be small across multiple counties, so we choose to use $k = 5$ throughout our analysis.

3. Analysis of food deserts in the Bay Area. To begin our analysis we use the publicly accessible food deserts data provided by the USDA in [6]. We will focus on two counties in the San Francisco Bay Area: Alameda County and Contra Costa County.

Our plots are generated by using Universal Transverse Mercator (UTM) coordinates. UTM coordinates take latitude and longitude coordinates and project them onto a 2D coordinate system that preserves distances for locations that are relatively close to one another. Because our analysis is contained to a county, any error in distances caused by using UTM coordinates is negligible. Additionally, we note that UTM distances are in meters. One can find latitude

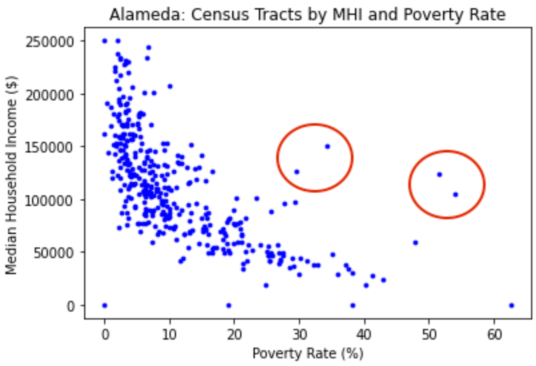

Figure 5. *Median household income (MHI) vs. poverty rate for census tracts in Alameda county. Circled census tracts are an example of high MHI and high poverty rates.*

and longitude coordinates for census tracts via the US Census [2] and we use the Bidirectional UTM converter (Python) found in [8]. Note that the latitude and longitude coordinates for a census tract provided by the US Census point to the center of the census tract.

3.1. Classical k-means clustering. We now compute elbow plots for Alameda and Contra Costa counties, using both half mile and 1 mile food desert data. These elbow plots inform our choice of k as we compute clusters and cluster centroids. In Figure 3 one can find the elbow plots which suggest using $k = 5$ for our clustering for both counties; Figure 4 shows the clustering and centroids produced by the k-means algorithm. The latitude and longitude coordinates for the centroids of each cluster can be found in Table 1. These centroids determine where a food pantry could be placed to close the gap of food access within these neighborhoods.

3.2. Weighted k-means clustering. Following our initial analysis, the authors (who are residents of the Bay Area) wanted a better idea of how the disparities in wealth that we notice every day manifest in this discussion on food deserts. This led to an analysis of one key component in the definition of a food desert—low income. In Figure 5 one can see the prime motivator for adjusting the way in which we perform our clustering. When plotting median household income (MHI) versus poverty rate, in addition to observing the expected trend, one also notices some startling outliers with a MHI of over $100,000 and a poverty rate above 30% (and in some cases above 50%).

We believe this is an indicator of the harsh disparities in wealth within the Bay Area. Relatively wealthy households are found around the corner from communities of substantial size living below the poverty line. To incorporate this into our clustering analysis we turn to a weighted version of the k-means clustering algorithm.

For a data set X with $|X| = n$, one can assign a weight w_i to each data point $\mathbf{x_i} \in X$, so that

1. $0 \leq w_i \leq 1$ for each $i \in \{1, 2, \ldots, n\}$, and
2. $W = \sum_{i=1}^{n} w_i = 1$.

For a cluster C_r define

$$W_r = \sum_{i \in C_r} w_i.$$

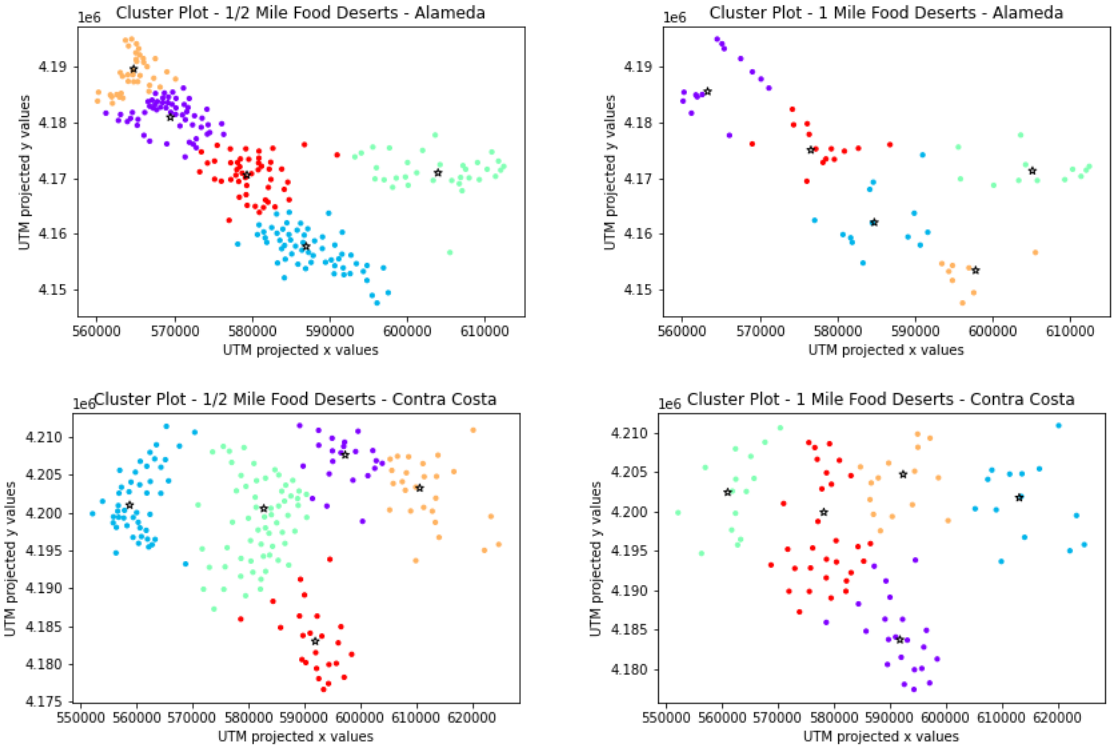

Figure 6. ***Weighted****: Cluster plots for half mile (left) and 1 mile food deserts (right) in Alameda (top) and Contra Costa (bottom) county. Plots are generated with the weighted k-means algorithm which prioritize census tracts with a larger population of people living below the poverty line. Stars indicate the location of the cluster centroids.*

The centroid of the cluster C_r is then defined by

$$\mathbf{c_r^w} = \frac{1}{W_r} \sum_{i \in C_r} w_i \mathbf{x_i}.$$

Furthermore, the inertia we wish to minimize is given by

$$\text{Weighted Inertia} = \sum_{r=1}^{k} \sum_{i \in C_r} \frac{w_i \; d^2(\mathbf{x_i}, \mathbf{c_r^w})}{W_r} \tag{3.1}$$

The algorithm used to compute the clusters which minimize the weighted inertia is the same as for the unweighted k-means algorithm.

For our analysis we assign weights as follows. First, we identify the food deserts within a county (half mile or 1 mile); this is our data set X. Suppose $|X| = n$. Now, for each food desert $\mathbf{x_j} \in X$ we determine the number of people in the food desert who are living below the poverty line, call this number p_j. We then assign a weight of

$$w_j = \frac{p_j}{\sum_{i=1}^{n} p_i}$$

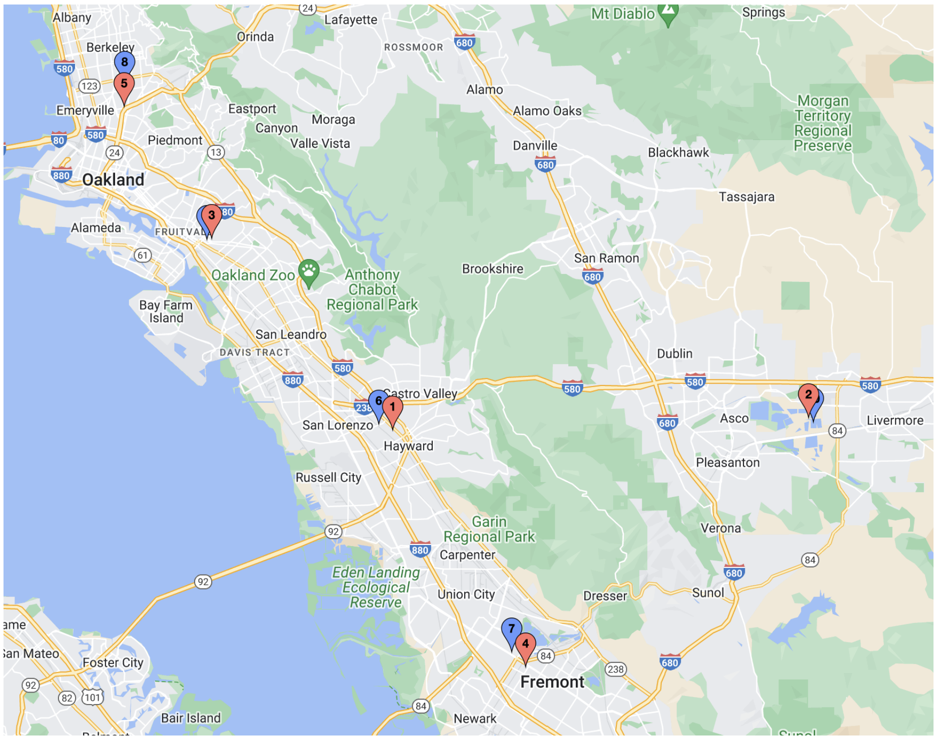

Figure 7. *Intervention sites suggested by our unweighted (red) and weighted (blue) k-means clustering for half mile food deserts in Alameda county. Created using [1].*

to the food desert $\mathbf{x_j}$. Our calculation for each p_j is based on the poverty rate and total population within a census tract. This data is provided by the USDA in [6].

In Figure 6 one can see the results of our poverty rate weighted clustering for half mile and one mile food deserts in Alameda county and Contra Costa county. The latitude and longitude coordinates for the centroids of these clusters can be found in Table 1. In Table 1, one can also see the distances between the centroids produced by the unweighted and the weighted k-means algorithm. Because we used the same $k = 5$ for all of our clustering, we were able to match centroids for clusters that are in a similar location in the unweighted and weighted cases. This gives us a way of comparing the results of the unweighted and weighted clustering.

The centroids in Table 1 can be interpreted as optimal locations for food pantries or other intervention sites since they minimize the distance that a person within a food desert cluster would need to travel to reach the resources they require. In some cases we observe significant changes in the locations of the centroids produced by the unweighted and the weighted clustering methods. Careful collaboration with community leaders to inform our weighted clustering may mean that the centroids resulting from the weighted clustering are more suitable locations for intervention sites.

4. Conclusion.

Food deserts and inequities in access to healthy foods continue to affect millions of Americans. A mathematically driven approach to launching interventions may be useful as local organizers and larger institutions build plans to address the food insecurities in their communities. While our k-means clustering approach to finding optimal intervention locations can be a useful tool, it does not replace the careful considerations of those familiar with the community being served.

Limitations to our approach include for example, the realizability of an intervention site produced by k-means when taking into account park lands, bodies of water, freeway locations, etc. Additionally, k-means is built to minimize the "straight line" (Euclidean) distance between points rather than the true distance (which accounts for buildings, roads, etc.) when traversing from one location to another which may be significantly longer than the "straight line" distance.

In the future, we hope to further incorporate relevant elements into our weighted k-means algorithm. Relevant elements may include access to a vehicle, proportion of the population which is over the age of 65, or the number of households receiving government food vouchers. Additionally, in the future we wish to deeply analyze the wealth gaps in the Bay Area, especially when the gaps are large within a small region like a census tract.

Acknowledgments. This goes out to all our friends, family, professors and to the universe. The authors would like to acknowledge the Muwekma Ohlone people on whose land Cal State East Bay is built. A special thank you also goes to MAA Tensor-SUMMA for funding this research experience.

REFERENCES

[1] BatchGeo, *Batchgeo mapping tool - suggested intervention sites, half mile food deserts alameda county*, https://batchgeo.com/map/256700ac86ce80686331d3014d77fcf4 (accessed 05/2022).

[2] U. Census, *U.s. gazetteer files - national census tracts gazetteer file*, 2020, https://www.census.gov/geographies/reference-files/time-series/geo/gazetteer-files.html (accessed 05/2022).

[3] U. Census, *Income in past 12 months (in 2020 inflation-adjusted dollars) alameda county*, 2022, https://data.census.gov/cedsci/table?q=income%20Alameda%20county&tid=ACSST5Y2020.S1901 (accessed 05/2022).

[4] L. M. Chihara and T. C. Hesterberg, *Mathematical Statistics with Resampling and R*, John Wiley & Sons, 2018.

[5] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, vol. 112, Springer, 2013.

[6] U. D. of Agriculture, *Usda food access research atlas - documentation*, 2020, https://www.ers.usda.gov/data-products/food-access-research-atlas/documentation/ (accessed 05/2022).

[7] P. J. Rousseeuw, *Silhouettes: a graphical aid to the interpretation and validation of cluster analysis*, Journal of computational and applied mathematics, 20 (1987), pp. 53–65.

[8] Tobias Bieniek, *Bidirectional utm-wgs84 converter for python*, 2021, https://github.com/Turbo87/utm (accessed 05/2022).

Alameda County - Half mile		
Unweighted (UW)	Weighted (W)	Distance - UW and W
(37.67591308, -122.0914838)	(37.67887313, -122.1001886)	0.52 miles
(37.55517098, -122.0057895)	(37.56294137, -122.0149782)	0.74 miles
(37.68210033, -121.8237974)	(37.67996982, -121.8208087)	0.22 miles
(37.84029789, -122.2642464)	(37.85118401, -122.2636467)	0.75 miles
(37.77337406, -122.2077322)	(37.77286856, -122.2107932)	0.17 miles
Contra Costa County - Half mile		
Unweighted (UW)	Weighted (W)	Distance - UW and W
(37.95942639, -122.3064405)	(37.95470429, -122.3294357)	1.29 miles
(37.92735069, -122.0750883)	(37.94872479, -122.0570464)	1.77 miles
(37.99464495, -121.9183478)	(38.01125601, -121.8907408)	1.89 miles
(37.96624864, -121.727528)	(37.97017592, -121.7399519)	0.73 miles
(37.79084227, -121.9511829)	(37.78952763, -121.9547526)	0.22 miles
Alameda County - One mile		
Unweighted (UW)	Weighted (W)	Distance - UW and W
(37.6852069, -121.8053787)	(37.68305658, -121.8077943)	0.20 miles
(37.82908148, -122.2637951)	(37.81493418, -122.2809348)	1.35 miles
(37.59501937, -122.0377943)	(37.60191583, -122.039759)	0.49 miles
(37.51642384, -121.9057412)	(37.52323853, -121.893133)	0.84 miles
(37.7252664, -122.1071411)	(37.71948896, -122.1310454)	1.37 miles
Contra Costa County - One mile		
Unweighted (UW)	Weighted (W)	Distance - UW and W
(37.96724195, -122.2821287)	(37.96761876, -122.3037876)	1.18 miles
(37.87566509, -122.1012026)	(37.94343747, -122.1088063)	4.70 miles
(37.98660562, -122.0165476)	(37.98549825, -121.9471073)	3.78 miles
(37.79609181, -121.9562659)	(37.79464452, -121.9458961)	0.57 miles
(37.95093693, -121.7093565)	(37.95650395, -121.7117546)	0.41 miles

Table 1

Latitude and Longitude Coordinates for the centroids corresponding to the clusters shown in Figure 4 (unweighted) and Figure 6 (weighted). Additionally, we give the distances between centroids corresponding to similarly located clusters.