CS 4774 MACHINE LEARNING PROJECT

TECHNICAL REPORT

Andrew Balch
Department of Computer Science
University of Virginia
Charlottesville, VA 22904
xxv2zh@virginia.edu

Eric Hamilton
Department of Computer Science
University of Virginia
Charlottesville, VA 22904
zay6tw@virginia.edu

Peter Tessier
Department of Computer Science
University of Virginia
Charlottesville, VA 22904
fpv5gr@virginia.edu

1 Abstract

This project sought to address the problem of food insecurity in Virginia by classifying regions to help inform policy. The approach involved combining a dataset of Virginia food bank data with USDA's Food Access Research Atlas, which contains data on food security. Early on, three distinct clusters of Virginia counties were discovered. The researchers found little success in linearly differentiating these clusters with a single metric, but were able to clearly distinguish them using a more holistic approach and feature analysis. These three clusters were called "food oasis", "food swamp", and "food desert", which corroborated previous literature's characterizations of such regions. Finally, the researchers were able to hand-select 9 features from the over 100 features, and use only these to predict which category a new region belonged to with 100% testing accuracy using a Deep Neural Network. With such a model, the researchers hope to better inform policy for both existing counties which have already been classified and new regions - whether in other states or in census tracts more local than counties - since food oases, food swamps, and food deserts should employ separate protocols to better suit their needs.

2 Introduction

According to 2021 data from the Federation of Virginia Food Banks, the food insecurity rate in Virginia is 8.1%. This means that over 704,000 people, of which over 164,000 are children, had to worry about where their next meal would come from on any given day [**?**]. Food banks are a vital resource in addressing the issue of food insecurity; however, access to food banks is tightly related to geography, leaving many Virginians - especially in rural areas - unable to access and receive benefits from these institutions. Some related obstacles that prevent many Virginians from accessing food banks include: finding transportation to and from food banks, navigating social safety nets, and dealing with stigma when relying on resources like food banks. For this project, we will focus on identifying under-served regions of Virginia and suggesting policies that would increase the access to food banks and the capacity of food banks to serve people specifically in these under-served regions.

There have been some established efforts in literature applying machine learning to the understanding and identification of food deserts. Amin et al. [2021] applied random forest, XGBoost, and LASSO algorithms to predict the modified Retail Food Environment Index (mRFEI) in a census tract, then classify it as a food desert or a food swamp by a range of mREFI values. They found that food deserts are commonly rural, lightly populated, and with low ethnic diversity while swamps are more urban, highly populated, and non-white with little vehicle access. However, their model only achieved a classification accuracy of $\tilde{7}2\%$.

Sucharitha and Lee [2019] applied clustering approaches instead of prediction to diagnose so-called food assistance deserts in Ohio and identify areas for improvement. These food assistance deserts are places where access to food banks or pantries is relatively low. Clusters yielded by their Gaussian Mixture Model were largely characterized by the distance of a family from a food bank or pantry. Clusters further from food assistance had a larger high-income population, leaving the low-income families underserved. They conclude that their observations cannot be expanded to other regions, and similar analyses are necessary to improve food equity across the country.

From the related work, it is clear that using machine learning to analyze food assistance as well as food availability is a valuable endeavor. However, such analyses are region-specific and are rarely conducted within the context of each other. With Virginia in mind, we aim to combine data on food banks together with county-by-county food access metrics to uncover unseen shared characteristics between Virginia counties, with respect to their food assistance programs and documented food insecurity.

3 Data and Methods

To reach our goal, we found two data sets that are readily accessible - the Federation of Virginia Food Bank (FVFB) and Food Access Research Atlas (FA) datasets. The FVFB contains data on each Virginia food bank, including how much food was distributed and how many households, individuals and children were served [noa, a]. This data is available for each month from 2019 to 2021. The second data set is the USDA's Food Access Research Atlas [noa, b]. It looks at each census tract and how much of its population has low access to healthy food. The population with low access to healthy food is further broken up over several key demographics such as race, age group, access to a car, income level, and more.

The two separate data sets were combined into one data set, with each row representing a Virginia county with data from 2019. FVFB was modified to only contain data from 2019, remove columns "Year" and "Month" (now that we only have data from 2019), remove the redundant columns "FIPS" (this county code is accounted for with the "Locality" column) and "Geopoint" (latitude and longitude already existed), and capitalize all county names to match with those from FA. The "Locality" column also became the index of the merged dataset. On the other hand, FA was modified to only contain data from Virginia, remove the now-unneeded column "State", remove the "CensusTract" column (all Census Tracts from the same county would have their data merged into the same county), and again capitalize all county names. Like in FVFB, the "County" column became the index column of the merged dataset. For both datasets, each feature was determined whether it should be summed or averaged when combining all data points into one county (for instance, all population values within a given county were summed together, while median family income was averaged). This resulted in a final count of 104 features for 130 counties, with only 5 counties being left out due to not being in both datasets.

After merging the tables, the data was cleaned in a three-step process. First columns with only null values or with one unique value were dropped from the dataset. This allowed us to reduce the dimensionality of the data by getting rid of any columns that could not provide us with any distinguishing information. Since after that process, missing values in the dataset were sparse (only two examples contained missing values), any examples with missing values were dropped. Finally, extreme outliers in each feature column were handled by clipping any data outside 1.5 times the inner-quartile range for that feature column down to the minimum and maximum of this scaled range, respectively. This allowed us to achieve a much less skewed distribution of the data across all features while only altering a small proportion of the examples. Through this process, many features of the distribution of the data that were previously lost due to the differences in scale between outliers were recovered.

With the data cleaned and combined, we used a semi-supervised learning technique to analyze it. This analysis involved clustering the data, providing each sample with a label based on the observed characteristics of each cluster, and finally classifying those labels with a neural network classifier. Dimensionality reduction was also a key consideration for our approach, since our final data set held 103 features. We also wanted to see if we could use feature engineering to derive a metric that represented the differences between food availability and assistance in Virginia counties. Such a metric would be useful both in this study as a comparison to clustering methods and in future work as an empirical reference point for the state of food assistance and insecurity in an area. We decided to evaluate our clustering experiments using two metrics: Silhouette Score and V-Measure. The Silhouette Score does not require ground a ground-truth label, instead assessing how well the clusters themselves are defined by comparing average intra-cluster and nearest-cluster cosine distances noa [c]. This score ranges from -1 to 1, where 1 is a dense and well-separated cluster. V-Measure requires ground truth labels, and is equivalent to Normalized Mutual Information. It is the harmonic mean of homogeneity (each cluster has the only one label) and completeness (the same labels are in the same cluster). The scale here is 0 to 1, where higher is better.

4 Experiments

First, we created a new feature of the data called Bank Score, which attempts to assess how well the food banks of a given region are meeting the needs of their population. Similar metrics have been proposed to assess the availability of meals provided by food assistance programs, such as food banks noa [2021]. The raw is calculated as follows.

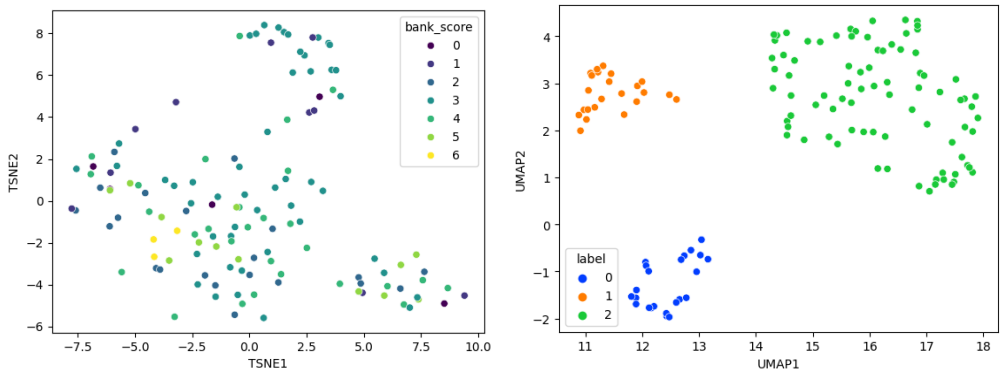
$$s^i_{\text{raw}} = \log \frac{\text{lbs. distributed}}{\text{low-income population}}$$

Outliers with a value of 0 for $\frac{\text{lbs. distributed}}{\text{low-income population}}$ (n=3) were imputed with the minimum observed value. The log was taken to make this score approximately normal and the samples were split into 7 bins across this distribution to create labels for our data. While these labels are flawed in their formulation, no single metric can holistically evaluate a food assistance program, they provide a useful point of comparison for clustering experiments.

We applied the following techniques to cluster the data, allowing us to make comparisons between the clusters and our derived Bank Score. We chose K-Means (KM), DBSCAN, Gaussian Mixture Modeling (GMM) as our clustering techniques. These three techniques were chosen to provide a reasonable baseline for how well common clustering approaches perform on our data. The input data were scaled prior to fitting each model.

KM clustering was performed with the same number of clusters as Bank Score labels, 7. The resulting clusters had a silhouette score of 0.213 and a v-measure of 0.153. Next, DBSCAN clustering was conducted with $\epsilon = 8$. This hyperparameter was determined by tuning with respect to v-measure. DBSCAN yielded a silhouette score of 0.219 and a v-measure of 0.158. Finally, GMM was fitted with the number of mixture components being equal to the number of Bank Score labels as well. The highest silhouette score of 0.35 and a v-measure of 0.166 were found with this approach. These clustering results indicate that dimensionality reduction is necessary to improve performance and increase interpretability. As it stands, with over 100 features, these metrics are our only reasonable means of interpreting the results of these approaches. As for the Bank Score, v-measure for each clustering method is still relatively low. This would normally indicate that this label is ill-suited to illustrate differences between food banks. However, the poor silhouette scores lead us to believe that it is worthwhile to consider dimensionality reduction and/or feature extraction before arriving at such a conclusion.

Lastly, we experimented with dimensionality reduction to visualize the data in two dimensions. The data were run through an encoding neural network to reduce the number of features from 104 to 10. The encoded data was then further reduced to two features using a t-Distributed Stochastic Neighbor Encoding (t-SNE) algorithm. The two dimensions from t-SNE were used to produce a scatter plot, which was colored with the aforementioned Bank Score to produce Figure 1a.



(a) t-SNE of Autoencoded Data with Bank Score (b) t-SNE of Autoencoded Data with Bank Score

Figure 1: Preliminary Clustering

These initial experiments left much to be desired, and we knew there had to be a more effective dimensionality reduction technique for our data than autoencoders and t-SNE. We found these qualities in an approach called Uniform Manifold Approximation and Projection (UMAP). We found that it performed well when clustering our data and created three distinct clusters that were easy to visualize because of its dimensionality reduction capabilities (Fig. 1b). The resulting silhouette score was 0.301, which was slightly lower than GMM. However, since we could visualize the clusters, we determined that the separation between data points was satisfactory and the score was likely due to the fact that the density of our clusters is relatively low. Therefore, we decided to use this approach moving forward to experiment with more metrics, analyze the derived clusters, and, finally, create a classifier based on them.

5 Results

The GitHub repository for the code used to conduct the above experiments and produce these results can be found at https://github.com/ericwhamilton1/gradient-descent-into-madness

5.1 Metric Creation

Our approach for finding a single metric that linearly differentiated the clusters follows noa [2021], which rates counties based on metrics like accessibility, availability, and stability. Bank Score, the first metric we designed, was distributed roughly equally between clusters (Fig. 1a). This indicated that the clusters perform similarly at distributing food based on the size of their low-income population. This result was corroborated by a host of similar metrics we explored (Fig. 2).

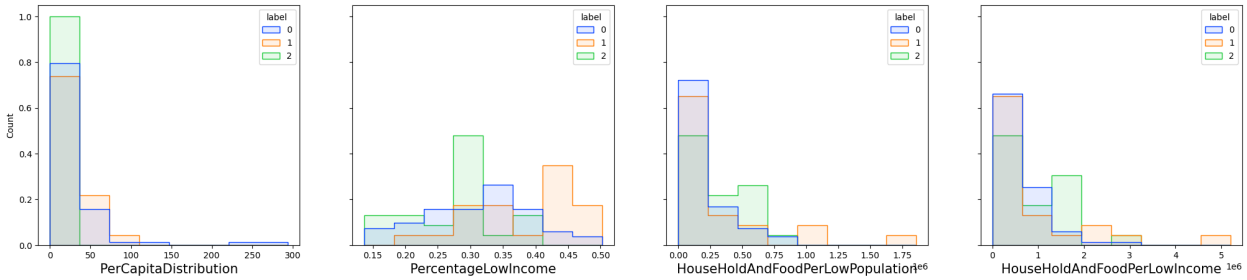


Figure 2: Metric Experimentation

5.2 Clustering Analysis

Our first step in clustering analysis was to see how the results of UMAP compared to KNN, DBSCAN, and GMM.

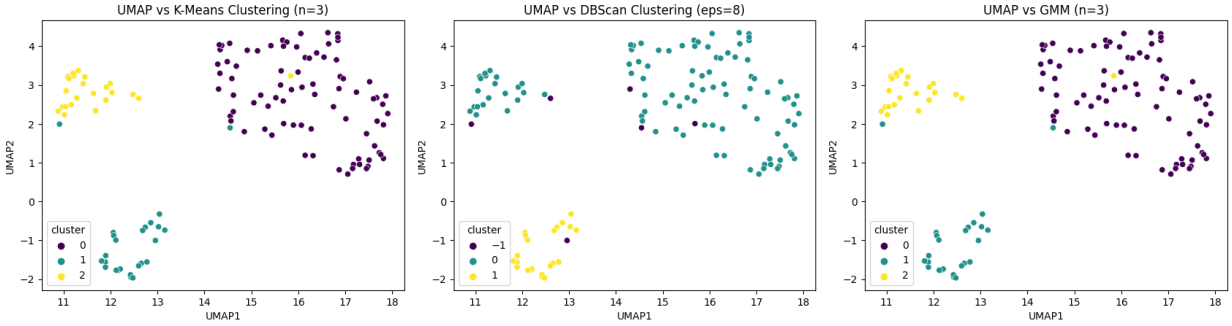


Figure 3: Traditional Clustering Results vs UMAP

Figure 3 shows that the three clusters derived by UMAP largely agree with the labels given by other clustering algorithms, indicating UMAP provides a relatively natural separation of the data. This observation is empirically confirmed by the v-measure scores for each baseline method against the UMAP labels. KM and GMM clustering showed the highest v-measure of 0.883. This drops to 0.588 for DBS. We suspect this result is likely due to how DBS clusters based on density because our data does not form very dense groupings (Fig. 1b). Silhouette scores for each baseline approach were also reevaluated and were more in line with earlier experiments. KM and GMM again had the highest scores of 0.31, very close to UMAP itself, while DBS was the lowest at 0.238.

Next, we examined the distributions of individual features across the data in each UMAP clusters in order to interpret how to interpret each cluster. We identified 6 key features that show relatively high separation between clusters. (Fig. 4). We found that cluster 0 populations were more rural, more likely to be receiving SNAP benefits and or to not have access to a vehicle. By contrast, clusters 1 and 2 are more urban, with a fewer proportion of low-access households or households receiving SNAP benefits. That being said, cluster 1 tended to have the lowest median family income and cluster 2 tended to have the highest.

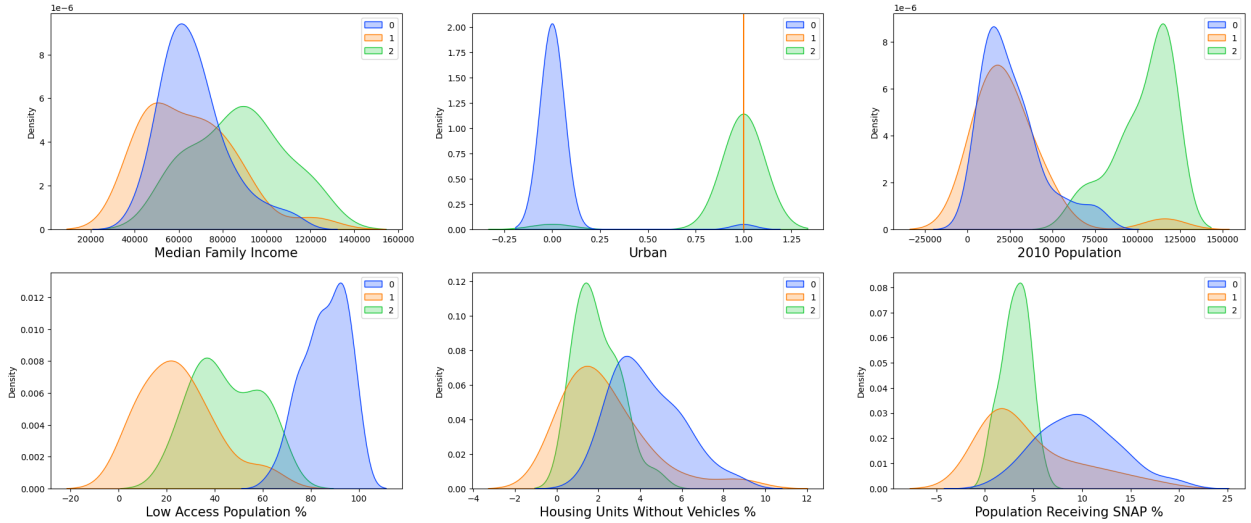


Figure 4: Key Feature Distributions based on UMAP cluster

Figure 5: ANN Classifier Training History

5.3 Classification

Finally, we trained a simple Artificial Neural Network classifier to predict the assigned clusters using the following subset of our original features:

`['Households Served', 'Pounds of Food Distributed', 'Pop2010', 'TractLOWI', 'PovertyRate', 'Urban', 'MedianFamilyIncome', 'lapophalf', 'lapop1', 'lapop10']`

We were able to reach 100% validation and testing accuracy after 100 epochs of training (Fig. 5)

6 Discussion and Conclusions

Our preliminary experiments with metric analysis proved unable to differentiate between the derived UMAP clusters. However, when we turned to analyzing the distribution of features already present within the database, we did find that each cluster represented valuable information about food availability and assistance in each Virginia county. Based on these distributions, we decided that counties in cluster 0 fit the criteria to be classified as food deserts and cluster 1 counties as food swamps. For cluster 2, we decided on the label "food oasis" reflecting the tendency of these counties to be higher income and more likely to have higher access to nutritious foods. It is interesting to note that Amin et al.

[2021] also classified regions into the same groups: "food desert," "food swamp," and "healthful," (analogous to "food oasis"). However, they had success with metric analysis by instead looking at the relative proportion of healthy food retailers in each census tract. Their metric "modified Retail Food Environment Index" or mRFEI is defined as follows:

$$mRFEI = 100 \times \frac{\text{number of healthy retailers in tract}}{\text{number of healthful retailers in tract} + \text{number of unhealthful retailers in tract}}$$

They found that their groups were linearly separable according to the metric. Future research could investigate to see how much overlap there is between both clustering results, and whether we could try a similar metric to classify our data.

One concern about the clusters was that there were many similar features describing the geographical distribution of different population groups in our dataset. We worried that these features would be highly cross-correlated, leading to UMAP focusing disproportionately on this local structure and neglecting more important global features. However, since our ANN classifier has success while only being trained on 9 handpicked features in the dataset, this indicates to us UMAP was learning more than just the local topology of similar features. Additionally, it meant that classifying additional counties would likely only require data across these 9 features instead of the original 100+ dimensional feature space.

Our work was unable to gain a lot of specific insights into the effectiveness of the food banks themselves, or provide policy suggestions, due to the limited nature of our food bank-related data. Based on Ginsburg et al. [2019], our food bank data is missing a lot of information that would be useful to form a deeper representation of each countys food assistance, such as hours of operation, the quality of the food, and more. Future work should attempt to gather and analyze more detailed data about individual food banks within Virginia counties. When placed within the context of the results of this project, such research could provide a more holistic view of the three clusters we have identified, or even allow for the discovery of new levels of food access within Virginia.

7 Member Contributions

All authors contributed equally to this paper. Peter Tessier combined the two datasets and experimented with metric creation, and edited the video. Eric Hamilton cleaned the merged data, experimented with dimensionality reduction techniques, and implemented the DNN classifier. Andrew Balch implemented and evaluated the described clustering approaches, as well as interpreted the resulting clusters. All members contributed to creating visualizations and utility code.

References

Modhurima Dey Amin, Syed Badruddoza, and Jill J. McCluskey. Predicting access to healthful food retailers with machine learning. *Food Policy*, 99:101985, February 2021. ISSN 0306-9192. doi:10.1016/j.foodpol.2020.101985. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7564312/.

Rahul Srinivas Sucharitha and Seokcheon Lee. Application of Clustering Analysis for Investigation of Food Accessibility. *Procedia Manufacturing*, 39:1809–1816, January 2019. ISSN 2351-9789. doi:10.1016/j.promfg.2020.01.258. URL https://www.sciencedirect.com/science/article/pii/S235197892030322X.

Food Bank Data | Virginia Open Data Portal, a. URL https://data.virginia.gov/Economy/Food-Bank-Data/xvir-sctz.

USDA ERS - Food Access Research Atlas, b. URL https://www.ers.usda.gov/data-products/food-access-research-atlas/.

2.3. Clustering, c. URL https://scikit-learn/stable/modules/clustering.html.

A Playbook for Localities: Inclusion of Food Security Metrics into Strategic Planning, January 2021. URL https://www.capitalareafoodbank.org/wp-content/uploads/2021/01/Food-Security-Playbook_Localities_FINAL.pdf.

Zoë A. Ginsburg, Alexander Bryan, Ellen B. Rubinstein, Hilary J. Frankel, Andrew R. Maroko, Clyde B. Schechter, Kristen Cooksey Stowers, and Sean C. Lucan. Unreliable and Difficult-to-Access Food for Those in Need: A Qualitative and Quantitative Study of Urban Food Pantries. *Journal of community health*, 44(1):16–31, February 2019. ISSN 0094-5145. doi:10.1007/s10900-018-0549-2. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6330151/.