Predicting the Emergence of Food Deserts

Nigele X McCoy (nmccoy9@gatech.edu), Anthony Philip Lee (alee657@gatech.edu)

Problem Statement

Food deserts, defined as geographic areas with limited access to affordable and nutritious food, have become a significant concern in both urban and rural communities. According to the USDA's Economic Research Service, over 6,500 food desert tracts were identified in the United States based on Census data and information about grocery/supermarket locations. These tracts represent areas where residents have minimal access to fresh, healthy food options. Specifically, 23.5 million people live in low-income areas more than a mile from a supermarket or large grocery store in urban areas, and over 10 miles in rural areas, further highlighting the disparities in food accessibility. [1]

The issue of food deserts is of critical importance due to its profound effects on public health, economic stability, and overall quality of life. The FAO defines food security as a state in which all individuals have access to sufficient, safe, and nutritious food to lead active and healthy lives. Food deserts undermine this goal by restricting access to essential resources, often leading to poorer dietary outcomes and increased rates of diet-related diseases such as obesity and diabetes. In times of economic and supply chain disruptions, ensuring communities have the necessary resources for food security is essential for resilience and well-being.[4]

Furthermore, addressing food deserts aligns with global initiatives, such as the United Nations' Sustainable Development Goal 2, which aims to eradicate hunger by 2030. With the UN projecting that over 600 million people worldwide will still face hunger by 2030, understanding and mitigating the development of food deserts is essential in this broader fight against hunger. [2]

Different studies have previously looked at characteristics and predictors of food deserts. This research has been instrumental in helping shape government policy. These studies have looked at demographic, socio-economic and mobility/transportation factors that are significant predictors of food deserts. However, most of these studies have not incorporated crime rate data into their analysis and may not be able to capture the recent trends of higher level of petty thefts and vandalisms that have driven business away from potentially high risk food desert areas.

This projects sought to address two key questions: **What demographic, socio-economic, and crime rate factors contribute to the formation of food deserts?** and **How can we predict their emergence?** By leveraging data from the USDA, census reports, and crime statistics, this study will fit classification and clustering models to identify the key drivers of food deserts and potentially forecast their development at the county level.

Related Literature

One of the most relevant studies in this space is a sanctioned study by the Department of Agriculture – *Characteristics and Influential Factors of Food Deserts* by Dutko, Ver Ploeg and Farrigan. In this study, they evaluated US Census data and 2006 data on locations of supermarket. They found that the there are different factors that are common and different between designated urban and rural areas. For both urban and rural areas, they found that minority population, poverty rates and region of the country are significant predictors of food deserts. In rural areas, additional factors like percent of vacant housing and the inverse of unemployment rates. Furthermore, they found different set of predictors for highly dense and less dense urban areas. [3]

Dutko, et. al. used two approaches to their analysis. First, they creatively used Descriptive Analysis to look at the data and the difference between the categories. Second, they used a logit multivariate regression to identify significant predictors for food deserts. They had set up different models that covered the different cases and used a Chow test to determine if the factors are significant.

This study is different from the cited paper by evaluating the impact of crime to predicting food deserts. That is answering the question – is the presence of high crime rates correlated to the emergence of food deserts? In addition, this study will use different classification and clustering models to see if we can identify the same or a different set of significant predictors.

Data Sources

In this study, we utilize several key data sources to investigate and predict the emergence of food deserts by county in the United States. These data sets provide insights into food access, crime rates, and environmental factors that contribute to food insecurity.

1. **Food Access Research Atlas**
   This dataset offers a comprehensive overview of food access indicators for low-income and other census tracts across the U.S. It measures supermarket accessibility through different metrics, helping to assess the availability of affordable and nutritious food in various regions. This data is vital for identifying areas where food access is particularly limited. Source: USDA Economic Research Service

2. **United States Crime Rates by County**
   Crime rates at the county level are an important factor in understanding the broader context of food deserts, as high crime areas may deter grocery store development or contribute to diminished food access. This dataset provides crime statistics across U.S. counties, offering critical insight into how safety concerns interact with food availability. Source: United States Crime Rates by County

3. **Food Environment Atlas**
   This dataset highlights various food environment factors, such as the proximity of stores and restaurants, food prices, nutrition assistance programs, and community characteristics. These factors directly influence food choices and diet quality, which are essential for identifying the

causes of food deserts and exploring potential policy interventions. Source: USDA Economic Research Service

By combining these data sources, we aim to develop a more comprehensive understanding of the factors that contribute to food deserts and use this information to predict their emergence and inform potential solutions.

Methodology

Data Wrangling and Preparation

The data wrangling process involved data acquisition, cleaning, feature engineering, and model development, leveraging socio-economic, crime, and environmental data. Data acquisition and cleaning were conducted using dictionary comprehensions of Pandas DataFrames, ensuring the data was easily readable and accessible for merging and modification. Cleaning techniques included removing invalid data and performing imputation and scaling of data. Data analysis and visualization encompassed analyzing data distributions, creating KDE plots to explore relationships, and developing visualizations to better understand the dataset and highlight food deserts.

Definitions

1. Crime Feature definitions and meaning can be found in the reference - https://www.icpsr.umich.edu/web/ICPSR/studies/37059/variables.
2. Food Environment Atlas feature definitions and meanings are laid out in Appendix A.

Feature Selection

Variance Inflation Factor (VIF) is a measure used to detect multicollinearity which occurs when predictor variables are highly correlated. VIF quantifies how much the variance of a regression coefficient is inflated due to multicollinearity; a higher VIF indicates a stronger correlation. A VIF value of 1 suggests no correlation between a variable and others, while values exceeding 5 or 10 are often considered problematic, depending on the context. By identifying variables with high VIFs, analysts can decide whether to remove or combine those variables to improve model performance and interpretation.

Table 1 shows the VIF report for the features of the entire data set. From these results, we could see how certain variables like PCF_NHWHITE10, PCT_NHBLACK10, PCT_HISP10, PCT_NHASIAN10 have a significantly high VIF indicating strongly linear correlation between each.

Table 1 Variance Inflation Factor for features of entire data set

Variable	VIF
PCT_NHWHITE10	482.147768
PCT_NHBLACK10	245.516887
PCT_HISP10	221.629207
PCT_NHASIAN10	12.684934

PCT_NHNA10	64.843583
PCT_NHPI10	3.728196
PCT_65OLDER10	3.505275
PCT_18YOUNGER10	2.180616
MEDHHINC10	5.942611
POVRATE10	18.238307
PERPOV10	2.163169
CHILDPOVRATE10	15.324779
PERCHLDPOV10	2.350806
METRO13	1.670099
POPLOSS00	1.262392
2010 Census population	886.686714
crime_rate_per_100000	1.82596
CPOPARST	1625.614755
CPOPCRIM	1615.535533
AG_ARRST	22.532485
AG_OFF	22.158557
MURDER	11.883829
RAPE	8.494323
ROBBERY	12.310923
AGASSLT	15.715638
BURGLRY	23.530015
LARCENY	25.436302
MVTHEFT	7.662684
ARSON	4.89334

Based on the VIF report, variables were recursively eliminated to see the reduction in the VIFs. For instance, the variables *'PCT_NHBLACK10',* and *'PCT_HISP10'* were eliminated and the variables *'PCT_NHWHITE10'* and *"PCT_NHASIAN10"* were retained. This is indicative that race has a potential for a significant impact to areas that are food deserts.

The variables *'CPOPCRIM'* and *'CPOPARST''* were also eliminated. These variables are aggregates of the other crime metrics and should be removed from our feature set.

The variable *'crime_rate_per_100000'* was also removed as it was strongly correlated with the *'2010 Census population'* and the sum of all the crimes reports.

The final features eliminated were *'CPOPCRIM'*, *'CPOPARST', 'crime_rate_per_100000', 'PCT_NHBLACK10',* and *'PCT_HISP10'*. A more detailed discussion of the features removed and its qualitative significance to the study follows in the Evaluation and Results section.

Descriptive Analysis

With the curated data set, we look at some of the core characteristics of the data. In total, there are 3064 counties with data in the study. 24% of these counties are considered low food access areas in that they meet the threshold defined that at least 30% of the population have low access to food.

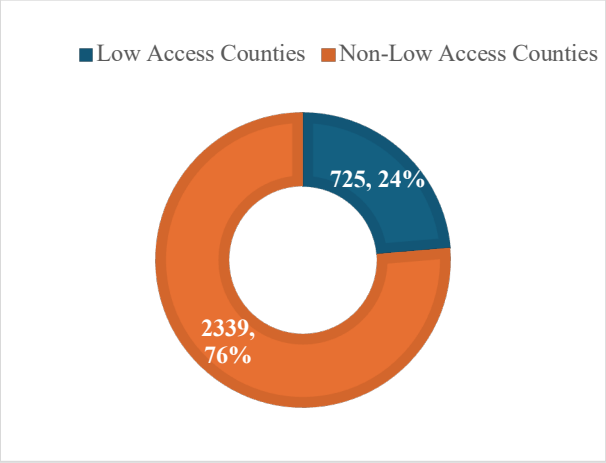

Figure 1 Ratio of low food access counties

Figure 2 shows the distribution of the data specifically for the percentage of the population with low access to data by county. From here, we can see that most of the data is centered around 20% low food access. This data allowed us to select 30% as the threshold to define a low food access county or what is otherwise defined as a food desert.

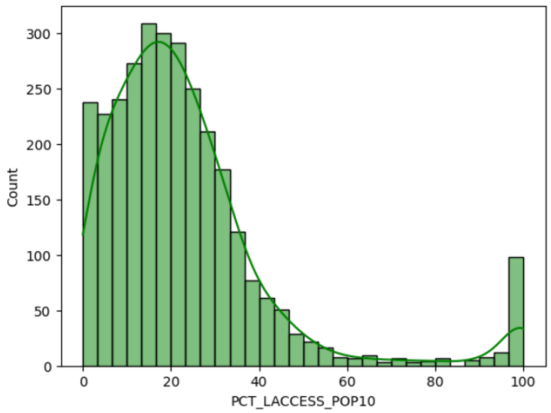

Figure 2 Distribution of percent low access counties

Figure 3 shows a map of the continental United States identifying the counties that were categorized as low food access counties. 20% of counties designated as urban or metro counties are low food access areas while 25% of counties designated as rural counties are low food access areas. As the map strongly suggest, rural areas in the western United States have far greater land area which makes food access potentially more difficult (i.e. longer distances between groceries and markets). *[4]*

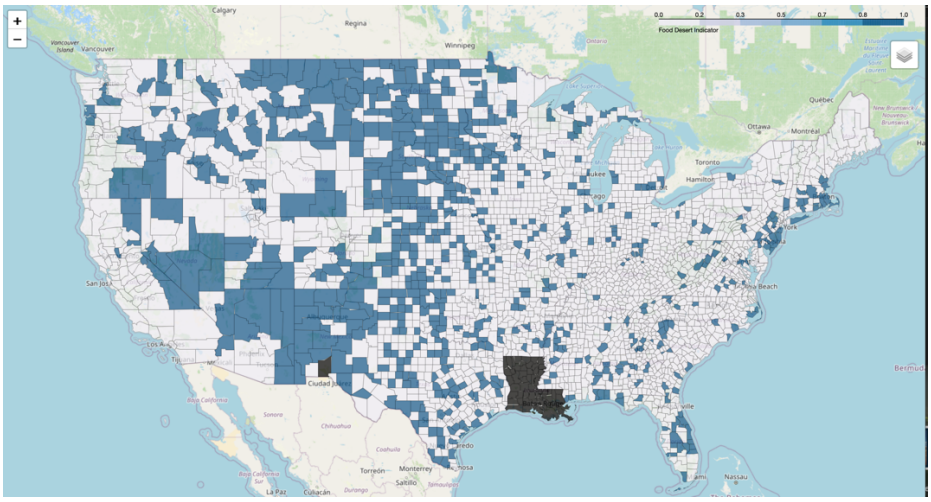

Figure 3 Identifying low food access counties across the continental USA

Figure 4 shows a different view of the map that colors each county based on the percentage of the population with low access to food. Counties that have a higher percentage of the population with low access to food are highlighted with a more intense red color.

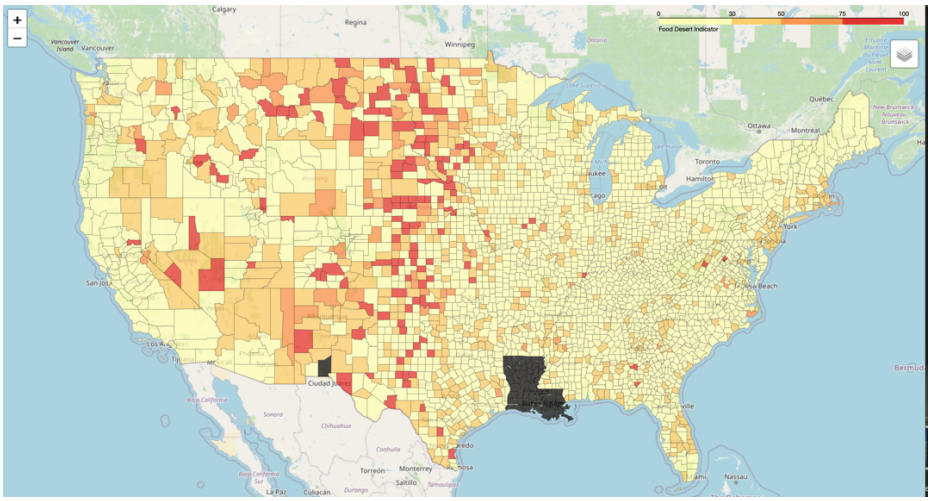

Figure 4 Map of counties by percentage of population with low access to food

Evaluation and Results

Model Evaluation and Selection

As part of the analysis, different models were evaluated to see if any particular model performs better than the rest. A 5-fold cross validation method was used to determine the accuracy of each model. These results are reported in *Table 2 Accuracy of different models evaluated*. The accuracy performance of the different models were found to be relatively comparable with logistic regression being marginally better than the other models.

Table 2 Accuracy of different models evaluated

Model	Accuracy	Observations
Logistic Regression	77.43%	
Support Vector Machines (Svm) - Linear	76.82%	
Support Vector Machines (Svm) – Rbf	76.46%	
K-Nearest Neighbors (Knn)	76.33%	
Gradient Boosted Classifier	75.97%	Runtime performance was 10x slower than other models

Logistic Regression was selected as the primary model. This is due to two reasons. Firstly, the model offered the best accuracy performance. Secondly, the logistic regression model creates results that are easily explainable and allows us to identify significant factors that can be rationalized in the social context. This allows us to provide stronger recommendations consistent with the goal of the study in helping drive socio-political priorities and programs that could help address the problem of food access.

Identifying Significant Factors

Once the model was selected as logistic regression, two packages were used to build the model – *scikit-learn* and *statsmodel*. The *statsmodel* package offered an advantage in that it creates a report on the factors that allows to easily identify their significance.

Table 3 Significant factors for logistic regression model shows the summary report for the statsmodel logistic regression. From the previous section, this model reported a 77.43% accuracy performance based on the testing data split.

Table 3 Significant factors for logistic regression model

Variable	coef	std err	z	P>|z|	[0.025	0.975]
PCT_NHWHITE10	-0.5319	0.080	-6.671	0.000	-0.688	-0.376
PCT_NHASIAN10	0.1238	0.082	1.509	0.131	-0.037	0.285
PCT_NHNA10	0.2096	0.051	4.108	0.000	0.110	0.310
PCT_NHPI10	-0.0678	0.054	-1.266	0.206	-0.173	0.037
PCT_65OLDER10	0.6628	0.090	7.335	0.000	0.486	0.840
PCT_18YOUNGER10	0.1524	0.073	2.081	0.037	0.009	0.296
MEDHHINC10	0.0997	0.120	0.834	0.404	-0.135	0.334
POVRATE10	0.2577	0.219	1.176	0.240	-0.172	0.687

PERPOV10	0.1460	0.077	1.904	0.057	-0.004	0.296
CHILDPOVRATE10	0.6545	0.199	-3.283	0.001	-1.045	-0.264
PERCHLDPOV10	-0.1372	0.082	-1.666	0.096	-0.299	0.024
METRO13	-0.0402	0.068	-0.590	0.555	-0.174	0.093
POPLOSS00	0.2657	0.052	5.084	0.000	0.163	0.368
2010 Census Population	0.0549	0.364	0.151	0.880	-0.659	0.769
AG_ARRST	-0.3552	0.267	-1.329	0.184	-0.879	0.168
AG_OFF	0.1890	0.263	0.718	0.473	-0.327	0.705
MURDER	-0.2997	0.419	-0.716	0.474	-1.121	0.521
RAPE	0.2069	0.219	0.946	0.344	-0.222	0.636
ROBBERY	1.2532	0.591	-2.120	0.034	-2.412	-0.095
AGASSLT	0.1075	0.363	0.296	0.767	-0.605	0.820
BURGLRY	1.5929	0.456	3.495	0.000	0.700	2.486
LARCENY	0.3342	0.420	0.795	0.426	-0.490	1.158
MVTHEFT	1.8777	0.555	-3.381	0.001	-2.966	-0.789
ARSON	-0.3915	0.243	-1.613	0.107	-0.867	0.084

From Table 3 Significant factors for logistic regression model, the factors with p-values < 0.05 were selected as statistically significant in predicting food deserts. These significant variables are reported below and ordered respectively by the absolute value of their coefficients indicating which variables most strongly impact our predictor.

1. “MVTHEFT” – total count of motor vehicle thefts reported
2. “BURGLRY” – total count of burglaries reported
3. “ROBBERY” – total count of robberies reported
4. “PCT_65OLDER10” - % Population 65 or older in 2010
5. “CHILD_POVRATE10” - % Child poverty rate in 2010***
6. “PCT_NHWHITE10” - % White population in 2010
7. “POPLOSS00” – categorical variable indicating population loss in 2010
8. “PCT_NHNA10” - % American Indian or Alaska Native in 2010

Note that “PERPOVRATE10” or persistent poverty rate in 2010 narrowly missed our significance test.

Interpreting Results

The significant factors impacting food deserts identified were eye-opening and alarming. These factors can be categorized into the following groups.

1. Occurrences of crime – The prevalence of crime has a significant correlation or impact to food deserts. More specifically though, motor vehicle theft, burglary and robbery were the most significant and other types of crimes like murder, rape, assault, larceny and arson do not have impacts to food deserts. This is consistent with current news trends of retailers such as Walmart, Target, and Walgreens, closing stores in locations where there is rampant theft. [5] At the onset of this study, we had asked the question if the existence of higher crimes rates are related to food deserts and the study strongly indicates that the crimes are the heaviest factors for food deserts as demonstrated by the highest coefficients in our regression model.

2. Age and poverty demographics – One of the worst affected demographic cohorts by food access are some of the most vulnerable in our society. Specifically, a heavy concentration of 65 and older members of the population is a strong indicator of food access problems. In addition, the higher rates of poverty in children under the age of 18 also is indicative of food access issues.

3. Race composition – Race also significant impacts areas that are associated as food deserts. More specifically, it disproportionately impacts non-white communities (i.e. black and Hispanic) give the high significance and negative coefficient of the "PCT_NHWHITE10" variable. This indicates that counties that have a smaller White population are more likely to be food deserts. In addition, the model also identified Native American and Alaska Native people as significantly impacted by food deserts.

4. Population loss – The last significant factor identified is population loss in 2010. This is a binary variable indicating if a particular county decreased in population from 2005 to 2010. The model identified this as highly correlated to low food access counties consistent with the findings of Dutko et al.

Conclusion

When we started with the project, we set out to answer two questions – 1.) **what demographic, socio-economic, and crime rate factors contribute to the formation of food deserts?** 2.) **how can we predict their emergence?**

After performing a logistic regression analysis, we have found that there are several socio-economic and crime factors that are highly correlated to low food access areas or food deserts. This includes the occurrences of specific types of crime, age and poverty demographics, race composition and population loss. It was also found that crime, especially crimes against property (theft, robbery, burglary) is the most significant factor associated with low food access areas.

These factors identify core priorities that our policy makers and community leaders must monitor and develop programs that help address. For instance, a higher degree of attention must be focused on communities that are predominantly non-white. Investments must be made to help address criminality and poverty in these areas and programs must be put in place to help these communities hold or grow their population levels.

Appendix A – Definitions of Feature Variables

Feature Variable	Long Name	Description
FIPS	Federal Information Processing Standard Code	A unique identifier for geographic regions.
State	State	State name or abbreviation.
County	County	County name.
LACCESS_POP10	Low Access Population 2010	The number of people with low access to grocery stores in 2010.
PCT_LACCESS_POP10	Percentage Low Access Population 2010	The percentage of the population with low access to grocery stores in 2010.
LACCESS_LOWI10	Low Income Low Access 2010	The number of low-income individuals with low access to grocery stores in 2010.
PCT_LACCESS_LOWI10	Percentage Low Income Low Access 2010	The percentage of low-income individuals with low access to grocery stores in 2010.
LACCESS_CHILD10	Low Access Children 2010	The number of children with low access to grocery stores in 2010.
PCT_LACCESS_CHILD10	Percentage Low Access Children 2010	The percentage of children with low access to grocery stores in 2010.
LACCESS_SENIORS10	Low Access Seniors 2010	The number of seniors with low access to grocery stores in 2010.
PCT_LACCESS_SENIORS10	Percentage Low Access Seniors 2010	The percentage of seniors with low access to grocery stores in 2010.
LACCESS_HHNV10	Low Access Households Without Vehicles 2010	The number of households without vehicles with low access to grocery stores in 2010.
PCT_LACCESS_HHNV10	Percentage Low Access Households Without Vehicles 2010	The percentage of households without vehicles with low access to grocery stores in 2010.
PCT_NHWHITE10	Percentage Non-Hispanic White 2010	The percentage of the population that is non-Hispanic White in 2010.
PCT_NHBLACK10	Percentage Non-Hispanic Black 2010	The percentage of the population that is non-Hispanic Black in 2010.
PCT_HISP10	Percentage Hispanic 2010	The percentage of the population that is Hispanic in 2010.
PCT_NHASIAN10	Percentage Non-Hispanic Asian 2010	The percentage of the population that is non-Hispanic Asian in 2010.
PCT_NHNA10	Percentage Non-Hispanic Native American 2010	The percentage of the population that is non-Hispanic Native American in 2010.
PCT_NHPI10	Percentage Non-Hispanic Pacific Islander 2010	The percentage of the population that is non-Hispanic Pacific Islander in 2010.
PCT_65OLDER10	Percentage 65 and Older 2010	The percentage of the population that is 65 years or older in 2010.
PCT_18YOUNGER10	Percentage 18 and Younger 2010	The percentage of the population that is 18 years or younger in 2010.
MEDHHINC10	Median Household Income 2010	The median household income in 2010.
POVRATE10	Poverty Rate 2010	The overall poverty rate in 2010.
PERPOV10	Percentage of People in Poverty 2010	The percentage of individuals living in poverty in 2010.
CHILDPOVRATE10	Child Poverty Rate 2010	The poverty rate among children in 2010.
PERCHLDPOV10	Percentage of Children in Poverty 2010	The percentage of children living in poverty in 2010.
METRO13	Metro Status 2013	Indicates if the county is part of a metropolitan area in 2013.

POPLOSS00	Population Loss Since 2000	Population loss in the county since 2000.
2010 Census Population	2010 Census Population	Total population recorded during the 2010 Census.
LACCESS_POP10_FLAG	Flag for Low Access Population 2010	Indicator if data is missing or incomplete.
LACCESS_LOWI10_FLAG	Flag for Low Income Low Access 2010	Indicator if data is missing or incomplete.
LACCESS_CHILD10_FLAG	Flag for Low Access Children 2010	Indicator if data is missing or incomplete.
LACCESS_SENIORS10_FLAG	Flag for Low Access Seniors 2010	Indicator if data is missing or incomplete.
LACCESS_HHNV10_FLAG	Flag for Low Access Households Without Vehicles 2010	Indicator if data is missing or incomplete.
crime_rate_per_100000	Crime Rate Per 100,000	Number of crimes per 100,000 population.
CPOPARST	County Police Arrests	Total arrests made by county police.
CPOPCRIM	County Police Crimes	Total crimes reported in the county.
AG_ARRST	Aggregate Arrests	Total arrests across all crime types.
AG_OFF	Aggregate Offenses	Total number of offenses recorded.
MURDER	Murder	Total murders reported.
RAPE	Rape	Total rapes reported.
ROBBERY	Robbery	Total robberies reported.
AGASSLT	Aggravated Assault	Total aggravated assaults reported.
BURGLRY	Burglary	Total burglaries reported.
LARCENY	Larceny	Total larcenies reported.
MVTHEFT	Motor Vehicle Theft	Total motor vehicle thefts reported.
ARSON	Arson	Total arson cases reported.

Bibliography

[1] USDA, "Access to Affordable and Nutritious Food: Measuring and Understanding Food Deserts and Their Consequences," *USDA,* 2009.

[2] U. Nations, "Sustainable Development Goals," [Online]. Available: https://www.un.org/sustainabledevelopment/hunger/. [Accessed 16 October 2024].

[3] P. M. V. P. a. T. F. Dutko, "Characteristics and Influential Factors of Food Deserts," U.S. Department of Agriculture, Economic Research Service, August, 2012.

[4] "GeoJSON Map of United States Counties by FIPS Code," Opendatasoft.com, [Online]. Available: https://raw.githubusercontent.com/plotly/datasets/master/geojson-counties-fips.json. [Accessed 6 December 2024].

[5] I. Ivanova, "Retailers Like Walmart and Starbucks Are Closing in Big Cities. Some Cite Crime, But Changing Habits May Be More Likely," *WTTW New,* 12 May 2023.