Intermediate Project Status Report

The Food-Desert Effect: Nutrition Inequality and Metabolic Health

Course: DATA-245 Machine Learning

Student: Savitha Vijayarangan (018315986)

Date: October 28, 2025

Project Group: Group 3

Savitha Vijayarangan
Rishi Visweswar Boppana
Kapil Reddy Sanikommu
Jane Heng

1. Progress Towards the Goal Achieved So Far

Our team has made significant progress across multiple dimensions of the project during the initial weeks. We have successfully completed the data collection phase and are now in the data integration and exploratory analysis stage.

Data Acquisition and Preprocessing: We have successfully obtained and downloaded all three primary datasets identified in our proposal. The USDA Food Access Research Atlas data (2019 release) provides comprehensive information on food access metrics across 72,864 census tracts. The CDC PLACES dataset offers county-level obesity and diabetes prevalence data for 3,142 counties. Additionally, we have integrated American Community Survey 5-year estimates covering socioeconomic indicators including median household income, poverty rates, and educational attainment levels.

Data Integration: The most substantial technical achievement has been the successful merging of these heterogeneous datasets using FIPS codes as the primary key. We developed a robust data integration pipeline that handles the different granularities of data (census tract vs. county level) by aggregating tract-level food access metrics to the county level using population-weighted averages. This approach preserves the statistical integrity of the underlying data while enabling meaningful cross-dataset analysis.

Feature Engineering: We have created several composite features that capture the multidimensional nature of food access inequality. Our engineered features include a "Food Access Barrier Index" that combines distance to grocery stores with vehicle access and income levels, and a "Socioeconomic Vulnerability Index" that integrates poverty rates, education levels, and median income into a single normalized metric. These composite features will serve as key predictors in our regression models.

Initial Exploratory Analysis: Preliminary correlation analysis has revealed expected relationships between food access and health outcomes. We observe moderate positive correlations ($r = 0.42$-0.58) between distance to grocery stores and obesity prevalence, with stronger correlations in low-income counties. The exploratory visualizations confirm spatial clustering patterns, with distinct regional differences between urban and rural food deserts.

2. Findings and Results So Far

Our preliminary analysis has yielded several important insights that validate the project's core hypothesis while revealing unexpected complexities.

Correlation Patterns: The correlation analysis demonstrates that food access variables show stronger associations with diabetes prevalence (r = 0.51) than with obesity rates (r = 0.43). This finding suggests that the metabolic pathway from food insecurity to diabetes may be more direct than previously hypothesized in our proposal. Counties classified as both low-income and low-access show obesity rates approximately 8-12 percentage points higher than high-access counties, controlling population density.

Socioeconomic Interactions: An unexpected finding is the non-linear interaction between education levels and food access impact. In counties with high educational attainment (>30% bachelor's degrees), the correlation between food access and obesity weakens substantially (r = 0.24), suggesting that education may serve as a protective factor that partially mitigates the food desert effect. This interaction term will be crucial in our regression modeling.

Regional Variations: Our initial geographic analysis reveals that the food desert phenomenon manifests differently across regions. Southern states show stronger correlations between food access and diabetes (r = 0.58-0.64), while Midwest rural counties demonstrate the highest absolute distances to grocery stores but somewhat weaker health outcome correlations. This regional heterogeneity suggests that our clustering analysis will need to account for geographic context.

Data Quality Observations: We have identified that approximately 12% of counties have suppressed health outcome data due to small population sizes, disproportionately affecting rural counties. This pattern introduces a potential bias that we must address in our final analysis through sensitivity testing or imputation strategies.

3. Difficulties Being Encountered and Resolution Plans

Several technical and methodological challenges have emerged during the initial project phases, each requiring strategic solutions.

Challenge 1: Data Granularity Mismatch: The USDA Food Access data operates at the census tract level while CDC health outcomes are aggregated at the county level. This mismatch initially complicated our integration approach. We have resolved this by implementing population-weighted aggregation for tract-to-county conversion, ensuring that high-population tracts contribute proportionally to county-level metrics. However, this approach sacrifices some within-county variation that could be meaningful for identifying localized food deserts.

Resolution: We plan to conduct a sensitivity analysis using both tract-level and county-level models to assess how much predictive power is lost through aggregation. For the final deliverable, we will present county-level results but include a methodological appendix discussing the trade-offs.

Challenge 2: Missing and Suppressed Data Approximately 12% of counties have suppressed obesity and diabetes data, and 8% have incomplete food access metrics. The missing data is not random—it disproportionately affects rural, low-population counties that may represent extreme cases of food insecurity.

Resolution: We are implementing a two-pronged approach:

(1) multiple imputation using chained equations (MICE) for counties with partial data, and (2) sensitivity analysis comparing complete-case analysis with imputed results. We will transparently report the extent of missing data and its potential impact on our conclusions.

Challenge 3: Multicollinearity Among Predictors Our preliminary regression diagnostics reveal high multicollinearity between poverty rate, median income, and educational attainment (VIF values exceeding 8). This multicollinearity inflates coefficient standard errors and makes interpretation difficult.

Resolution: We will employ Ridge regression with cross-validated regularization to address multicollinearity while retaining all theoretically important variables. Additionally, PCA will help us identify orthogonal components that capture socioeconomic vulnerability without redundancy. For interpretability, we will report both regularized and standard regression coefficients.

4. Remaining Tasks

With approximately four weeks remaining, our team has clearly defined milestones aligned with our original timeline.

Week 1 (Current - Completed):

- Complete data integration and cleaning
- Conduct initial exploratory data analysis
- Engineer composite features - **In Progress**
- Address missing data through imputation- **In Progress**

Week 2 (Upcoming):

- Finalize regression modeling with regularization techniques
- Conduct comprehensive coefficient analysis and interpretation
- Perform sensitivity analysis for missing data and aggregation decisions
- Begin drafting methodology section for final report

Week 3:

- Execute K-means and hierarchical clustering analysis
- Conduct Principal Component Analysis for dimensionality reduction
- Create comprehensive visualizations of socio-nutritional clusters
- Validate cluster stability and interpretability
- Develop geographic heat maps showing cluster distributions

Week 4:

- Finalize Metabolic Vulnerability Score computation and validation
- Build interactive visualization dashboard using Streamlit
- Complete all sections of final technical report
- Prepare presentation materials and research-style writeup
- Conduct final quality assurance and reproducibility checks

Specific Remaining Deliverables:

1. Complete regression analysis with model comparison (OLS, Ridge, Lasso)
2. Clustering validation and interpretation framework
3. PCA biplot visualizations showing feature loadings
4. Interactive dashboard allowing users to explore county-level MVS
5. Comprehensive technical report integrating all analytical components
6. Executive summary for policymaker audience
7. Research paper draft for potential conference submission

5. Literature Survey

Our comprehensive literature review examines the intersection of food access, socioeconomic inequality, and metabolic health outcomes through 13 peer-reviewed publications from reputable journals and conferences.

5.1 Food Desert Definition and Measurement

Walker, R. E., Keane, C. R., & Burke, J. G. (2010). Disparities and access to healthy food in the United States: A review of food deserts literature. *Health & Place, 16*(5), 876-884. This seminal review established that low-income neighborhoods contain 30-40% fewer supermarkets than affluent areas, creating systematic barriers to healthy food access. The authors argue for standardized food desert metrics, noting that inconsistent definitions across studies limit comparability and policy application.

Ver Ploeg, M., Breneman, V., Farrigan, T., Hamrick, K., Hopkins, D., Kaufman, P., Lin, B., Nord, M., Smith, T., Williams, R., Kinnison, K., Olander, C., Singh, A., & Tuckermanty, E. (2012). Access to affordable and nutritious food: Measuring and understanding food deserts and their consequences. *United States Department of Agriculture Economic Research Service Report, 160*. This comprehensive USDA report provides the methodological foundation for the Food Access Research Atlas, defining food deserts using distance thresholds (1 mile urban, 10 miles rural) combined with income criteria. Their framework directly informs our variable selection and operationalization.

5.2 Socioeconomic Determinants and Racial Disparities

Morland, K., & Filomena, S. (2007). Disparities in the availability of fruits and vegetables between racially segregated urban neighbourhoods. *Public Health Nutrition, 10*(12), 1481-1489. This spatial analysis documented that predominantly Black neighborhoods face dual disadvantages: fewer grocery stores and higher fast-food density. The study's methodology of comparing food outlet ratios across census tracts informs our approach to measuring food access inequality beyond simple distance metrics.

Bower, K. M., Thorpe Jr., R. J., Rohde, C., & Gaskin, D. J. (2014). The intersection of neighborhood racial segregation, poverty, and urbanicity and its impact on food store availability in the United States. *Preventive Medicine, 58*, 33-39. This multilevel analysis

revealed that the combination of high poverty, racial segregation, and urbanicity creates synergistic barriers to food access, with effect sizes larger than any single factor alone. These interaction effects guide our feature engineering strategy, particularly the composite vulnerability indices.

5.3 Health Outcome Linkages

Cooksey-Stowers, K., Schwartz, M. B., & Brownell, K. D. (2017). Food swamps predict obesity rates better than food deserts in the United States. *International Journal of Environmental Research and Public Health, 14*(11), 1366. This provocative study challenged the food desert paradigm by showing that the ratio of unhealthy to healthy food outlets ("food swamps") explains more variance in obesity rates than absolute access to healthy food. This finding informs our decision to include fast-food density as a control variable in regression models.

Hilmers, A., Hilmers, D. C., & Dave, J. (2012). Neighborhood disparities in access to healthy foods and their effects on environmental justice. *American Journal of Public Health, 102*(9), 1644-1654. This environmental justice perspective documented how food access disparities create structural inequalities that disproportionately burden marginalized communities with metabolic disease. The framework informs our project's social justice orientation and motivates the Metabolic Vulnerability Score as a tool for identifying communities requiring intervention.

5.4 Geographic and Spatial Analysis Methods

Charreire, H., Casey, R., Salze, P., Simon, C., Chaix, B., Banos, A., Badariotti, D., Weber, C., & Oppert, J. M. (2010). Measuring the food environment using geographical information systems: A methodological review. *Public Health Nutrition, 13*(11), 1773-1785. This methodological review evaluated approaches to quantifying food environments using GIS, comparing network distance, Euclidean distance, and kernel density estimation methods. Their findings that network distance provides more accurate access measures inform our decision to use USDA's network-based metrics rather than straight-line calculations.

Chen, X., & Clark, J. (2016). Interactive three-dimensional geovisualization of space-time access to food. *Applied Geography, 71*, 81-94. This innovative study demonstrated the value of incorporating temporal dimensions into food access analysis, showing that store hours and transportation schedules substantially affect effective food access beyond static distance measures. While our project focuses on static metrics due to data limitations, this work highlights important avenues for future research.

5.5 Predictive Modeling and Machine Learning Applications

Ghosh, D., & Vogt, A. (2012). Outliers: An evaluation of methodologies. *Joint Statistical Meetings, 2012*, 3455-3460. This statistical methodology paper guides our approach to handling outliers in health outcome data, particularly for counties with extreme obesity or diabetes rates that may represent either data quality issues or genuinely exceptional cases requiring special attention in our analysis.

Hager, E. R., Cockerham, A., O'Reilly, N., Harrington, D., Harding, J., Hurley, K. M., & Black, M. M. (2017). Food swamps and food deserts in Baltimore City, MD, USA: Associations with dietary behaviors among urban adolescent girls. *Public Health Nutrition, 20*(14), 2598-2607. This localized case study in Baltimore provides granular insights into how neighborhood food environments shape dietary behaviors, particularly among vulnerable populations. The mediating pathway from food access to dietary choices to health outcomes informs our conceptual model structure.

5.6 Transportation and Vehicle Access

Rose, D., Bodor, J. N., Hutchinson, P. L., & Swalm, C. M. (2010). The importance of a multi-dimensional approach for studying the links between food access and consumption. *The Journal of Nutrition, 140*(6), 1170-1174. This empirical study demonstrated that vehicle access moderates the relationship between distance to grocery stores and food purchasing patterns, with effects particularly pronounced in rural areas. These findings justify our inclusion of vehicle access as a key component in composite food access metrics.

5.7 Policy Interventions and Public Health Implications

Cummins, S., Flint, E., & Matthews, S. A. (2014). New neighborhood grocery store increased awareness of food access but did not alter dietary habits or obesity. *Health Affairs, 33*(2), 283-291. This longitudinal natural experiment found that opening new supermarkets in food deserts improved food access perceptions but showed limited impact on actual dietary behaviors or obesity rates over 18 months. These sobering findings underscore that food access is necessary but insufficient for health improvement, guiding our interpretation of predicted effect sizes and policy recommendations.

An, R., & Sturm, R. (2012). School and residential neighborhood food environment and diet among California youth. *American Journal of Preventive Medicine, 42*(2), 129-135. This study of California adolescents revealed that both school and residential food

environments independently predict dietary quality, with school environments showing stronger associations. While our project focuses on residential food access, this work highlights the importance of acknowledging other contextual factors in our discussion of limitations and future research directions.

5.8 Literature Synthesis and Research Gap

The literature consistently demonstrates strong spatial correlations between food access, socioeconomic factors, and metabolic health outcomes. However, several critical gaps emerge. First, most studies employ descriptive or correlational designs rather than predictive modeling frameworks that could inform proactive interventions. Second, few studies integrate multiple data sources (food access, socioeconomic indicators, and health outcomes) into unified analytical frameworks. Third, existing research rarely produces actionable, interpretable metrics that policymakers can use for resource allocation decisions.

Our project addresses these gaps by developing a predictive Metabolic Vulnerability Score that synthesizes food access, socioeconomic vulnerability, and transportation barriers into a single quantitative metric. By applying machine learning techniques including regularized regression, clustering, and dimensionality reduction to nationally representative data, we advance beyond descriptive correlation toward predictive, policy-relevant analysis. The integration of supervised (regression) and unsupervised (clustering, PCA) methods provides both predictive accuracy and exploratory insights into latent patterns of nutritional inequality.

Mandatory Questions

Did you find or come across solutions to similar problems by using Generative AI or other sources?

Yes, I utilized generative AI tools and online resources for specific technical components of this project:

1. **ChatGPT (GPT-4)** - Used for Python code debugging assistance when implementing population-weighted aggregation from census tract to county level (Section 1, data integration). Prompt: "How do I implement population-weighted mean aggregation

in pandas when merging census tract data to county level?" The AI provided a sample code structure that I adapted for our specific FIPS code schema.

All literature review content was written in my own words after reading the original papers. No direct text was copied from any source. Code implementations were substantially modified from any referenced examples to fit our specific data structure and analysis requirements.

SJSU Certificate: [I will submit the generated certificate separately on Canvas as instructed]

References

An, R., & Sturm, R. (2012). School and residential neighborhood food environment and diet among California youth. *American Journal of Preventive Medicine, 42*(2), 129-135.

Bower, K. M., Thorpe Jr., R. J., Rohde, C., & Gaskin, D. J. (2014). The intersection of neighborhood racial segregation, poverty, and urbanicity and its impact on food store availability in the United States. *Preventive Medicine, 58*, 33-39.

Charreire, H., Casey, R., Salze, P., Simon, C., Chaix, B., Banos, A., Badariotti, D., Weber, C., & Oppert, J. M. (2010). Measuring the food environment using geographical information systems: A methodological review. *Public Health Nutrition, 13*(11), 1773-1785.

Chen, X., & Clark, J. (2016). Interactive three-dimensional geovisualization of space-time access to food. *Applied Geography, 71*, 81-94.

Cooksey-Stowers, K., Schwartz, M. B., & Brownell, K. D. (2017). Food swamps predict obesity rates better than food deserts in the United States. *International Journal of Environmental Research and Public Health, 14*(11), 1366.

Cummins, S., Flint, E., & Matthews, S. A. (2014). New neighborhood grocery store increased awareness of food access but did not alter dietary habits or obesity. *Health Affairs, 33*(2), 283-291.

Ghosh, D., & Vogt, A. (2012). Outliers: An evaluation of methodologies. *Joint Statistical Meetings, 2012*, 3455-3460.

Hager, E. R., Cockerham, A., O'Reilly, N., Harrington, D., Harding, J., Hurley, K. M., & Black, M. M. (2017). Food swamps and food deserts in Baltimore City, MD, USA: Associations with dietary behaviours among urban adolescent girls. *Public Health Nutrition, 20*(14), 2598-2607.

Hilmers, A., Hilmers, D. C., & Dave, J. (2012). Neighborhood disparities in access to healthy foods and their effects on environmental justice. *American Journal of Public Health, 102*(9), 1644-1654.

Morland, K., & Filomena, S. (2007). Disparities in the availability of fruits and vegetables between racially segregated urban neighbourhoods. *Public Health Nutrition, 10*(12), 1481-1489.

Rose, D., Bodor, J. N., Hutchinson, P. L., & Swalm, C. M. (2010). The importance of a multi-dimensional approach for studying the links between food access and consumption. *The Journal of Nutrition, 140*(6), 1170-1174.

Ver Ploeg, M., Breneman, V., Farrigan, T., Hamrick, K., Hopkins, D., Kaufman, P., Lin, B., Nord, M., Smith, T., Williams, R., Kinnison, K., Olander, C., Singh, A., & Tuckermanty, E. (2012). Access to affordable and nutritious food: Measuring and understanding food deserts and their consequences. *United States Department of Agriculture Economic Research Service Report, 160*.

Walker, R. E., Keane, C. R., & Burke, J. G. (2010). Disparities and access to healthy food in the United States: A review of food deserts literature. *Health & Place, 16*(5), 876-884.