Overcoming computational hurdles to quantify the impact of food access on health: a statistical approach

Abstract Healthy foods are essential for a healthy life, but accessing healthy food can be more challenging for some people than others. With this disparity in food access comes disparities in well-being, leading to disproportionate rates of diseases in communities that face more challenges in accessing healthy food (i.e., low-access communities). Identifying low-access, high-risk communities for targeted interventions is a public health priority, but current methods to quantify food access are either computationally simple or accurate, but not both. We propose a hybrid statistical approach to combine these methods, allowing researchers to harness the computational ease of one with the accuracy of the other.

A. Objectives Consuming healthy foods is critical for lifelong health, leading to proper development in childhood and lowering the risk of chronic disease in adulthood. However, social and physical barriers, such as those impacting low-income and minority communities, can hamper access to healthy foods. This disparity in food access leads to disparities in health, as people with fewer healthy options see higher burdens of diseases, including coronary heart disease. To reduce these burdens, it is critical to identify neighborhoods or communities with low food access.

To identify low-access, high-risk neighborhoods, food access is often quantified using the distance to grocery stores (i.e., how far they are away). When calculating distance there currently exists a trade-off between (i) computationally simple methods that are less accurate and (ii) more accurate methods that are computationally complex. Computationally simple options draw a straight line between the two points "as the crow flies." However, this process ignores infrastructure like roads and topography like rivers, which underestimates true distances to grocery stores and overestimates food access. Map-based distances incorporate real-world obstacles, like roads and rivers, but free options like Google Maps, or even paid options like ArcGIS, can quickly become time-intensive. In Forsyth County, North Carolina (NC) there are 250 neighborhoods and 855 grocery stores; this would require 250 x 855 = 213,750 calculations to get the distances from every neighborhood to every store!

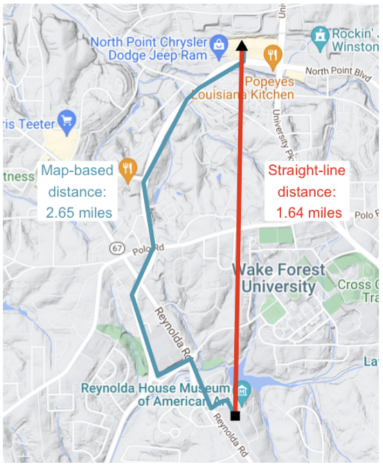

Fig 1. Straight-line and map-based distances from Reynolda House to a nearby Food Lion

An ideal solution would offer the accuracy of the map-based distances for all neighborhoods without the added computational complexity. Unfortunately, on a fixed budget and barring software developments from Google, this solution is not currently feasible. However, a hybrid one is possible: to obtain map-based distances for a subset of neighborhoods and use straight-line distances for the rest. Then, treating the straight-line distances as error-prone versions of the map-based ones we create a manageable measurement error problem instead of an unmanageable computational one. Problems like this have been the primary research focus of Dr. Sarah Lotspeich, an early career investigator and recipient of the David P. Byar Award for her prior work on statistical methods for measurement error. Collaborating with food access and health disparities experts, Dr. Lotspeich will tackle the following specific aims.

Aim 1: Accurately quantify food access using map-based distance for all neighborhoods and estimate the impact of food access on rates of adverse health outcomes.

Aim 2: Illustrate how badly error-prone straight-line distances can bias estimates of the impact of food access.

Aim 3: Combine food access calculated using straight-line distance for all neighborhoods and map-based distance for a subset of neighborhoods to obtain accurate estimates at reduced computational cost.

Upon completion, we will have demonstrated that the accuracy of the map-based distance calculations can be captured without querying all neighborhoods through a measurement error framework. This project will provide preliminary results for a future NIH R21 to develop new statistical methods to design and analyze food access and health disparities studies across NC.

Intellectual merit: This project will propose a new statistical approach to accurately quantify food access and model its impact on health with less computational strain. Through the adoption of a measurement error framework, the proposed approach will offer improved accuracy with computational ease.

Broader impacts: By overcoming the computational hurdles, the proposed methods will make large scale collection of accurate distance-based measures feasible. Therefore, the geographic scope of studies of access (to food, medical care, etc.) can expand to answer questions and drive decision-making for larger communities. All data collected and code written will also be made publicly available to propel further research.

B. Background and Significance

B1. Healthy eating is important to good health, yet not accessible to everyone. Eating healthy foods is critical to development in childhood and prevention of illnesses in adulthood, including cardiovascular disease, diabetes, and obesity.[1-3] Not just preference for healthy foods, but access, goes into what people eat. For some people, predominantly those in low-income or minority communities and with disabilities, healthy foods are not always accessible.[4-5] Access can be hampered by physical factors, like geographic proximity to stores and the availability of public transit,[6] and social factors, like structural racism and discrimination.[7-8] Thus, not just consumption of healthy foods but access to them can impact health,[9-12] and disparities in access perpetuate disparities in health. For public health officials seeking to reduce the burden of a disease, understanding its connection to food access and the landscape of food access in a community can be informative. Quantifying food access (e.g., how far people travel to buy healthy foods) is an important place to start.

B2. Current metrics of how close neighborhoods are to healthy foods have limitations. A neighborhood's access to healthy food can be quantified by the number of grocery stores within some proximity, but this metric relies on calculating the distances between the neighborhood and possible stores. Current distance calculations are either computationally simple but less accurate (e.g., they draw a straight line through the shortest path) or more accurate but computationally complex (e.g., they draw a path following roadways). Computationally simple methods can underestimate actual distance to the grocery store by ignoring lack of roadways or natural obstacles like rivers. The United States Department of Agriculture (USDA) uses a more sophisticated grid-based calculation,[13] but this method also ignores many realistic obstacles. These problems are likely exacerbated in rural areas, where there are generally fewer stores and fewer available straight-line routes. Preliminary data using straight-line distance suggest that many neighborhoods in Forsyth County do not have many food options (**Fig. 2**), and that's the best-case scenario. Using map-based distance to quantify these neighborhoods' food access would be more realistic.[6] The Google Maps API is an incredibly powerful for calculations like this, and software can integrate the API into a statistical workflow.[14-15] These tools can be used to calculate the map-based travel distance between two locations, offering a more accurate snapshot of a neighborhood's access. Still, making all of the necessary calculations (i.e., the distances between all neighborhoods and grocery stores) would not be feasible due to time-intensive computations and monthly limits on API usage. With map-based distances for only some neighborhoods, a challenge emerges: what about neighborhoods with only straight-line distances available? How can they be included in analyses?

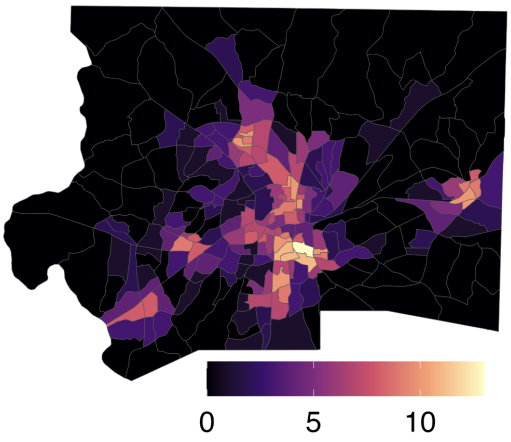

Fig 2. Map of the number of grocery stores within 1 mile (using straight-line distance) by neighborhood in Forsyth County, North Carolina

B3. Measuring access to healthy food is a measurement error problem. The straight-line and map-based methods seek to measure the same thing: distance between neighborhoods and grocery stores. Assuming that straight-line distance underestimates the more realistic map-based one, a measurement error model is a natural way to relate them. Consider estimating the neighborhood-level rate of coronary heart disease Y as a function of the number of grocery stores within one mile X (food access). Then assume that X is the number of stores with map-based distances less than one mile to the neighborhood, while X^*, the number of stores with straight-line distances less than one mile to the neighborhood, is an error-prone version of X. Since straight-line distances are computationally simple, X^* is available for all neighborhoods, but due to computational limits X is only available for some. This is the two-phase study design used in our prior

measurement error work,[16-17] and many methods can be used to analyze the resulting data with no missing X^* but some missing X. Replacing missing X values with predictions based on X^* (i.e., imputation) is a promising option,[18] because (i) it offers nice statistical precision and (ii) after replacing the missing Xs we can fit standard statistical models.[19]

B4. In Forsyth County, NC, 68%–87% of neighborhoods have low access to food. There are 95 census tracts in Forsyth County, NC. As of 2019, the USDA identified 63 (68%) and 81 (87%) of them as having low access to food, due to the proportion of people living more than 1 mile and 1/2 mile from the nearest supermarket, respectively. These high proportions of low-access tracts put Forsyth among the hardest-hit counties in NC.[20] Additional work is needed to determine if neighborhood variability in health (**Fig. 3**) can be explained by this variability in food access (**Fig. 2**). Dr. Lotspeich previously collaborated with public health experts to investigate neighborhood-level determinants of children being placed in state custody, where she employed health disparities methods.[21]

B5. To assess the impact of food access on health in Forsyth County, NC, data will be combined from three sources for analysis. As neighborhoods**,** the 250 census block groups in Forsyth County, NC will be used, which are the smallest geographic units with additional data, like demographics, available.[22] Their centroids will be used as the origin for distance calculations. Grocery store addresses were collected using the googleway package in the statistical computing language R,[15] and the ggmap package was used to geocode them to latitude and longitude coordinates.[14] Data for 855 grocery stores were collected this way from Forsyth and its bordering counties. Prevalences of health outcomes are available from the Centers for Disease Control and Prevention in the PLACES dataset.[23] Based on prior studies, coronary heart disease, diabetes, high blood pressure, and obesity will be considered.[9,11]

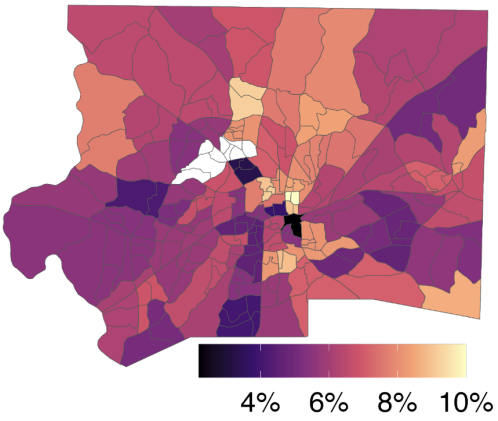

Fig. 3. Map of crude prevalence of coronary heart disease among adults by neighborhood in Forsyth County, North Carolina

C. Methods

C1. Setup and Notation For the health outcomes, let E denote the number of events (e.g., the number of people diagnosed with coronary heart disease) and and P denote the population in a neighborhood, such that E/P is the "rate" of that outcome in the neighborhood. To measure food access, let X be the number of grocery stores within one mile of the neighborhood. The data collected include M = 250 neighborhoods and n = 855 grocery stores. Assume that true food access, X, can only be obtained using the map-based distance calculations, while the straight-line distance calculations yield an error-prone measure of food access, X^*.

C2. Statistical Methodology Straight-line distances are calculated using the Haversine Formula, and map-based distances are calculated using the ggmap package to query the Google Maps API for the driving route.[14] Due to computational strain, X is only observed for m neighborhoods (m ≤ M), while X^* is observed for all. This setup creates three analytical datasets: (i) unqueried (i.e., m = 0), with X missing for all neighborhoods, (ii) fully-queried (i.e., m = 250), with X measured for all neighborhoods, and (iii) partially-queried (i.e., m = 125), with X measured for half of the neighborhoods and missing for the rest (**Fig. 4**). Constructing the unqueried dataset is computationally simple, as the geosphere package can be used to calculate all 250 x 855 = 213,750 calculations in 0.03 seconds,[24] but using this error-prone X^* in place of X in our analysis will lead to incorrect results.[25] Still, constructing the fully-queried dataset with X for all neighborhoods is computationally expensive. Calculating the map-based distance between one neighborhood and one grocery store takes longer than calculating all of the straight-line distances for all neighborhoods and

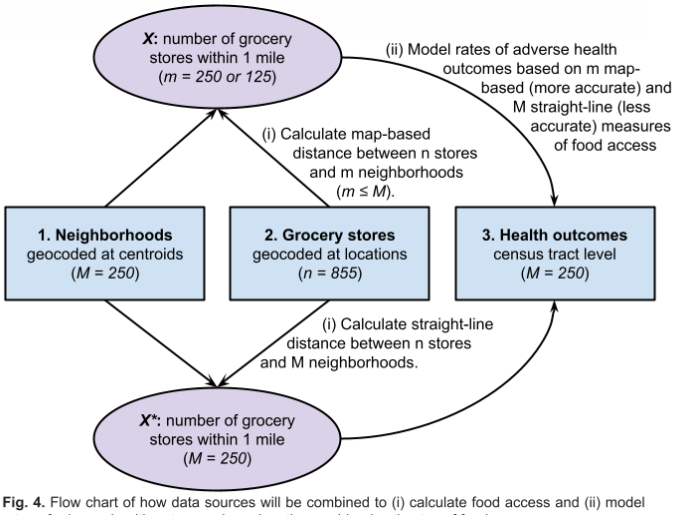

Fig. 4. Flow chart of how data sources will be combined to (i) calculate food access and (ii) model rates of adverse health outcomes based on the combined estimates of food access

stores (e.g., 0.42 seconds vs. 0.03 seconds). Creating the fully-queried dataset is a key challenge for this project, as resources are needed to overcome the barriers of the map-based calculations (e.g., slow computation and monthly query limits), but these data are critical to show (i) the extent to which food access impacts health (**Aim 1**), (ii) how badly error-prone measures to food access can bias the impacts on health (**Aim 2**), and (iii) how well the proposed approach captures the true impacts of food access on health while reducing computation (**Aim 3**). The partially-queried dataset will include X^* from the unqueried dataset for all neighborhoods and a random sample of X from the queried dataset for a subset of neighborhoods.

Poisson regression will be used to assess the impact of food access on the neighborhood-level rate of various adverse health outcomes. Inference and conclusions will be based on estimates of the model $\log(E/P) = \beta_0 + \beta_1 X$, testing the hypothesis that $\beta_1 = 0$ (i.e., that food access has no impact on health). In the partially-queried data, some Xs are missing. Dr. Lotspeich has developed various methods for partially-queried data like these, including new statistical models[16-17,26-27] and "optimal" designs to decide which data are most informative to query.[28-29] The specific aims will be addressed as follows.

Aim 1: The fully-queried data will be used to fit the model using E, P, and X from all neighborhoods. Results from this analysis will be used as a "gold standard" for Aims 2 and 3.

Aim 2: The unqueried data will be used to fit a "naive" model using E, P, and X^* (in place of X) from all neighborhoods. Comparisons between this model and the gold standard will describe the magnitude of bias due to using straight-line rather than map-based measures of food access.

Aim 3: For queried neighborhoods, let $\widehat{X} = X$. For unqueried neighborhoods, food access will be imputed as $\widehat{X} = \alpha_0 + \alpha_1 X^*$, with α_0 and α_1 estimated using an "imputation model." The imputation model is a linear regression of $X = \alpha_0 + \alpha_1 X^*$ fit to the subset of neighborhoods with both X and X^* measured. Then, the partially-queried data will be used to fit the model using E, P, and imputed $\widehat{X}$ from all neighborhoods. Multiple imputed datasets will be created, the same model fit to each, and then these models pooled to obtain final results in a multiple imputation framework.[18] Dr. Lotspeich has previously worked to implement and improve imputation methods like these.[30-32]

C3. Alternative Strategies If the imputation model for $\widehat{X}$ is inadequate, we will consider more complex options. For example, the model can be more flexible (e.g., using splines to allow a nonlinear relationship between X and X^*) or incorporate more information (e.g., a neighborhood indicator of rural/urban status). Depending on the performance of the imputed analysis, the size of the partially queried dataset could be modified (e.g., if it performs well with m = 125, how well would it perform with a smaller m = 100 or 75?). A proposed timeline can be found in **Fig. 5**.

Fig. 5. Proposed project timeline (June – December 2023)

Conclusions Healthy food is a critical determinant of healthy living, and yet many communities suffer from inadequate access to fresh, nutritious food. Building upon preliminary data from Forsyth County, we will develop new statistical methods for food access and health disparity studies on statewide scale in a future NIH R21. We will build on our prior measurement error work[16-17] to develop a more robust analytical approach that makes fewer assumptions about the relationship between straight-line and map-based distances for broader generalizability. We will derive new optimal designs to decide the most informative neighborhoods to query using Google Maps to achieve the best statistical precision, modifying our prior work to tackle a new data challenge.[27-28] With map-based travel time as an another metric of food access that is expensive to obtain,[33] we will consider imputing travel time from straight-line distance. Additionally, the computational gains from the proposed methods can be used to more easily expand the study area and quantify the impact of food access on health beyond our community here in Forsyth County to the entire state of North Carolina.

References

[1] Liu, S., Manson JE, Lee IM, Cole, SR, Hennekens CH, Willett WC and Buring, JE. (2000). Fruit and Vegetable Intake and Risk of Cardiovascular Disease: the Women's Health Study. *The American Journal of Clinical Nutrition*, 72(4), 922–928.

[2] Harding AH, Wareham NJ, Bingham SA, Khaw K, Luben R, Welch A and Forouhi NG. (2008). Plasma Vitamin C Level, Fruit and Vegetable Consumption, and the Risk of New-Onset Type 2 Diabetes Mellitus: the European Prospective Investigation of Cancer–Norfolk Prospective Study. *Archives of Internal Medicine*, 168(14), 1493–1499.

[3] California Center for Public Health Advocacy, PolicyLink, and the UCLA Center for Health Policy Research. (2008) *Designed for Disease: The Link Between Local Food Environments and Obesity and Diabetes.*

[4] Brucker DL and Coleman-Jensen A. (2017) Food Insecurity Across the Adult Lifespan for Persons with Disabilities, *Journal of Disability Policy Studies*, 28(2): 109–118.

[5] Bower KM, Thorpe RJ, Rohde C, Gaskin DJ. (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.

[6] Bennion N, Redelfs AH, Spruance L, Benally S and Sloan-Aagard C. (2022) Driving Distance and Food Accessibility: a Geospatial Analysis of the Food Environment in the Navajo Nation and Border Towns, *Frontiers in Nutrition*, 9: 904119.

[7] Burke MP, Jones SJ, Frongillo EA, Fram MS, Blake CE and Freedman DA. (2018). Severity of Household Food Insecurity and Lifetime Racial Discrimination Among African-American Households in South Carolina. *Ethnicity & Health*, *23*(3), 276–292.

[8] Odoms-Young A. and Bruce MA. (2018). Examining the Impact of Structural Racism on Food Insecurity: Implications for Addressing Racial/Ethnic Disparities. *Family & community health*, *41 Suppl 2 Suppl, Food Insecurity and Obesity*(Suppl 2 FOOD INSECURITY AND OBESITY), S3–S6.

[9] Ahern M., Brown C. and Dukas S. (2011), A National Study of the Association Between Food Environments and County-Level Health Outcomes. *The Journal of Rural Health*, 27: 367-379.

[10] Gallegos D, Eivers A, Sondergeld P and Pattinson C. (2021). Food Insecurity and Child Development: A State-of-the-Art Review. *International Journal of Environmental Research and Public Health*, 18(17), 8990.

[11] Kanchi R, Lopez P, Rummo PE, Lee DC, Adhikari S, Schwartz MD, Avramovic S, Siegel KR, Rolka DB, Imperatore G, Elbel B and Thorpe LE. (2021). Longitudinal Analysis of Neighborhood Food Environment and Diabetes Risk in the Veterans Administration Diabetes Risk Cohort. *JAMA Network Open*, 4(10), e2130789.

[12] Li Y, Wang S, Cao G, Li D and Ng BP. (2021). Disentangling Racial/Ethnic and Income Disparities of Food Retail Environments: Impacts on Adult Obesity Prevalence. *Applied Geography*, 137, 102607.

[13] Rhone A, Ver Ploeg M, Williams R and Breneman V. Understanding Low-Income and Low-Access Census Tracts Across the Nation: Subnational and Subpopulation Estimates of Access to Healthy Food, EIB-209, U.S. Department of Agriculture, Economic Research Service, May 2019

[14] Kahle D and Wickham H. (2013) ggmap: Spatial Visualization with ggplot2. *The R Journal*, 5(1), 144–161.

[15] Cooley D (2023). *googleway: Accesses Google Maps APIs to Retrieve Data and Plot Maps*. R package version 2.7.7, https://CRAN.R-project.org/package=googleway.

[16] Tao R, **Lotspeich SC**, Amorim G, Shaw PA and Shepherd BE. (2021) Efficient Semiparametric Inference for Two-Phase Studies with Outcome and Covariate Measurement Errors. *Statistics in Medicine*. 40(3): 725–738.

[17] **Lotspeich SC**, Shepherd BE, Amorim GC, Shaw PA and Tao R. (2022) Efficient Odds Ratio Estimation under Two-Phase Sampling Using Error-Prone Data from a Multi-National HIV Research Cohort. *Biometrics*. 78(4): 1674–1685.

[18] Cole SR, Chu H, Greenland S. (2006) Multiple-Imputation for Measurement-Error Correction, *International Journal of Epidemiology*, 35(4): 1074–1081.

[19] Little RJA. (1992) Regression with Missing X's: a Review. *Journal of the American Statistical Association*, 87(420), 1227–1237.

[20] U.S. Department of Agriculture, Economic Research Service. (n.d.). *Food Access Research Atlas*. Retrieved April 25, 2023, from https://www.ers.usda.gov/data-products/food-access-research-atlas/.

[21] **Lotspeich SC**, Jarrett RT, Epstein R, Schaffer A, Gracey K, Cull M and Raman R. (2020) Incidence of Children Placed in State Custody and Neighborhood-Level Risk Factors, *Child Abuse & Neglect,* 109, 104767.

[22] U. S. Bureau of the Census. (1994). *Geographic Areas Reference Manual*. U.S. Dept. of Commerce, Economics and Statistics Administration, Bureau of the Census.

[23] *PLACES*. Centers for Disease Control and Prevention. (2022). Public-use data file and documentation. https://chronicdata.cdc.gov/500-Cities-Places/PLACES-Local-Data-for-Better-Health-Census-Tract-D/cwsq-ngmh.

[24] Hijmans R (2022). *geosphere: Spherical Trigonometry*. R package version 1.5-18, https://CRAN.R-project.org/package=geosphere.

[25] Duan R, Cao M, Wu Y, Huang J, Denny JC, Xu H and Chen Y. (2017). An Empirical Study for Impacts of Measurement Errors on EHR Based Association Studies. *AMIA Annual Symposium proceedings. AMIA Symposium*, 1764–1773.

[26] Amorim G, Tao R, **Lotspeich S**, Shaw PA, Lumley T, Patel,RC and Shepherd BE. (2022). Three-Phase Generalized Raking and Multiple Imputation Estimators to Address Error-Prone Data. *arXiv preprint arXiv:2205.01743*.

[27] **Lotspeich SC**, Richardson BD, Baldoni PL, Enders KP and Hudgens MG. (2023). Quantifying the HIV Reservoir with Dilution Assays and Deep Viral Sequencing. *arXiv preprint arXiv:2302.00516*.

[28] Amorim G, Tao R, **Lotspeich S**, Shaw PA, Lumley T and Shepherd BE. (2021) Two-Phase Sampling Designs for Data Validation in Settings with Covariate Measurement Error and Continuous Outcome, *Journal of the Royal Statistical Society Series A: Statistics in Society*, 184(4): 1368–1389.

[29] **Lotspeich SC**, Shepherd BE, Amorim GC, Shaw PA and Tao R. (2023) Optimal Multiwave Validation of Secondary Use Data with Outcome and Exposure Misclassification. *Canadian Journal of Statistics*. https://doi.org/10.1002/cjs.11772.

[30] **Lotspeich SC**, Grosser KF and Garcia TP. (2022). Correcting Conditional Mean Imputation for Censored Covariates and Improving Usability. *Biometrical Journal*, 64, 858–862.

[31] **Lotspeich SC** and Garcia TP. (2022). Escaping the Trap: Replacing the Trapezoidal Rule to Better Impute Censored Covariates with Their Conditional Means. *arXiv preprint arXiv:2209.04716*.

[32] Grosser KF, **Lotspeich SC** and Garcia TP. (2023). Mission Imputable: Correcting for Berkson Error When Imputing a Censored Covariate. *arXiv preprint arXiv:2303.01602*.

[33] Swayne MRE and Lowery Bryce. (2021). Integrating Transit Data and Travel Time into Food Security Analysis: A Case Study of San Diego, California. *Applied Geography*, 131: 102461.

Budget and Justification

Student Research Assistant: To assist in data collection, statistical analysis, and manuscript preparation.
$20 per hour for 10 hours per week for 2 months (June and July) for a **total of $1600 research support**.

Computational Resources: To purchase additional queries through the Google Maps Distance Matrix API.
Approximately $0.004 per query for 250,000 queries for a **total of $1100 computational resources**.

Conference Travel: To present findings at the 2024 Women in Statistics and Data Science Conference (location to be announced in late 2023).
$400 for conference registration, $300 for airfare, $500 for hotel (2–3 nights), $100 for food for a **total of $1300 domestic travel**.

Outcome of Previous CEES Awards

None

Wake Forest University Collaborators

Dr. Lucy D'Agostino McGowan, Department of Statistical Sciences and Department of Biostatistics and Data Science