Supporting Information for "Combining straight-line and map-based distances to investigate the connection between proximity to healthy foods and disease"

Sarah C. Lotspeich, Ashley E. Mullan,
Lucy D'Agostino McGowan, and Staci Hepler
Department of Statistical Sciences, Wake Forest University

May 25, 2024

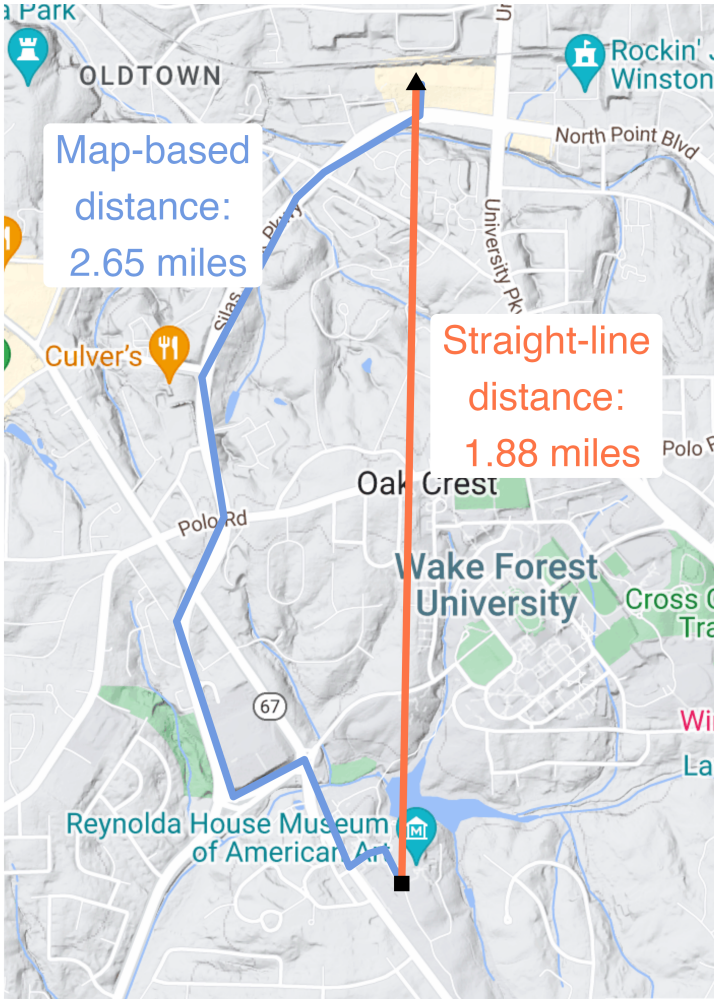

Figure S1: Straight-line and map-based distances from Reynolda House (square symbol) to a nearby Food Lion grocery store (triangle symbol) in Winston-Salem, North Carolina. It can be seen that the estimated driving distance (2.65 miles) was 1.6 times the straight-line distance between them (1.64 miles).

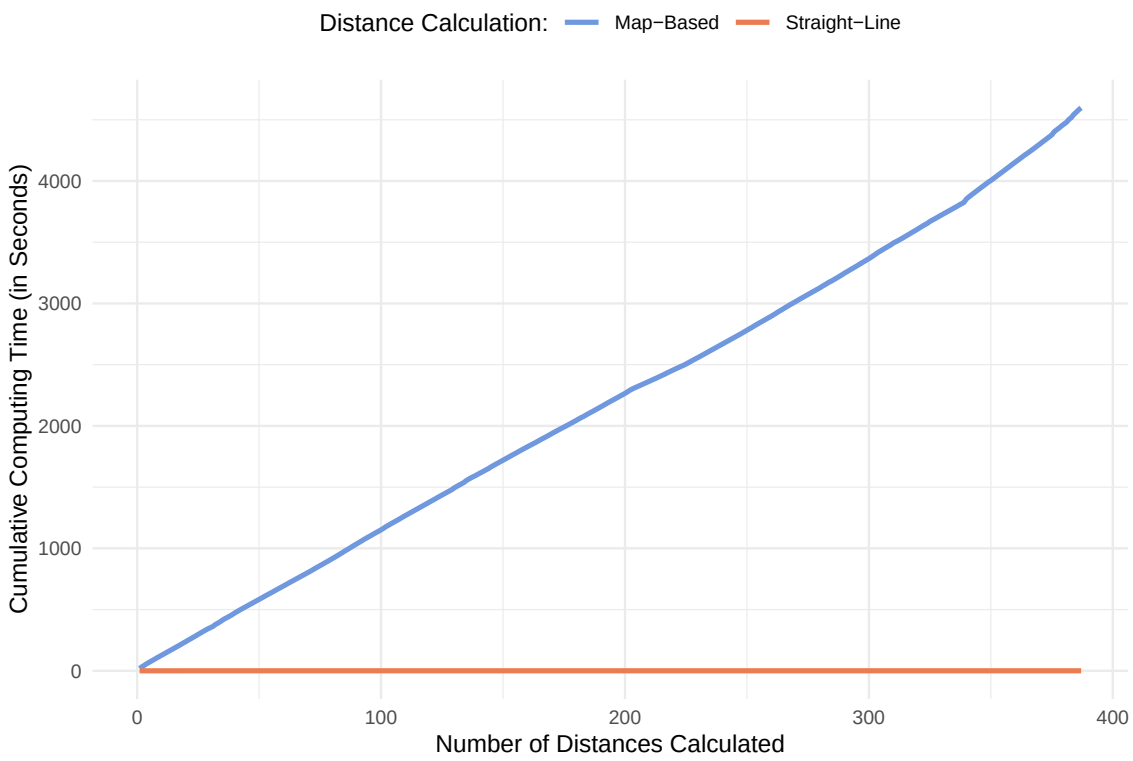

Figure S2: Line graph of the cumulative computing time (in seconds) for the map-based versus straight-line distance calculations in the Piedmont Triad data.

S.1 Including the Outcome in the Imputation Model

As recommended in Moons et al. (2006) and others, the analysis model outcome must be included in the predictor's imputation model for the imputation model to be congenial. Since the outcome in a Poisson regression model is transformed:

$$\log\{\mathrm{E}(Y|X)\} = \beta_0 + \beta_1 X + \log(Pop)$$
$$\rightarrow \log\left\{\frac{\mathrm{E}(Y|X)}{Pop}\right\} = \beta_0 + \beta_1 X. \tag{S.1}$$

Specifically, the analysis model outcome Y is being (i) divided by the offset Pop and then (ii) log-transformed. Thus, there were many possible ways of including Y in the imputation model for X considered.

Recall that the imputation model for map-based access X (described in Section 2.3) is a linear regression with X as the outcome. In a brief simulation study, five possible ways to include the health outcomes from the analysis model in the imputation model are considered:

1. $\mathrm{E}(X|X^*)$, ignoring the outcome and offset,
2. $\mathrm{E}(X|X^*, Y)$, including the untransformed outcome,
3. $\mathrm{E}\{X|X^*, \log(Y)\}$, including the transformed outcome,
4. $\mathrm{E}\{X|X^*, \log(Y/Pop)\}$, including the transformed prevalence, and
5. $\mathrm{E}\{X|X^*, \log(Y), \log(Pop)\}$, including the transformed outcome and offset.

Data were simulated following the data generating mechanism outlined in Section 3.1 for samples of $N = 390$ and 2200 neighborhoods, with a proportion of $q = 0.1$ that had map-based food access X available. An error mean and standard deviation of $\mu_U = -0.7$ and

$\sigma_U = 0.8$, respectively, were assumed, and the outcome prevalence and prevalence ratio for X were approximately 7% and 1.01, respectively. Results from these simulations are displayed in Figure S3.

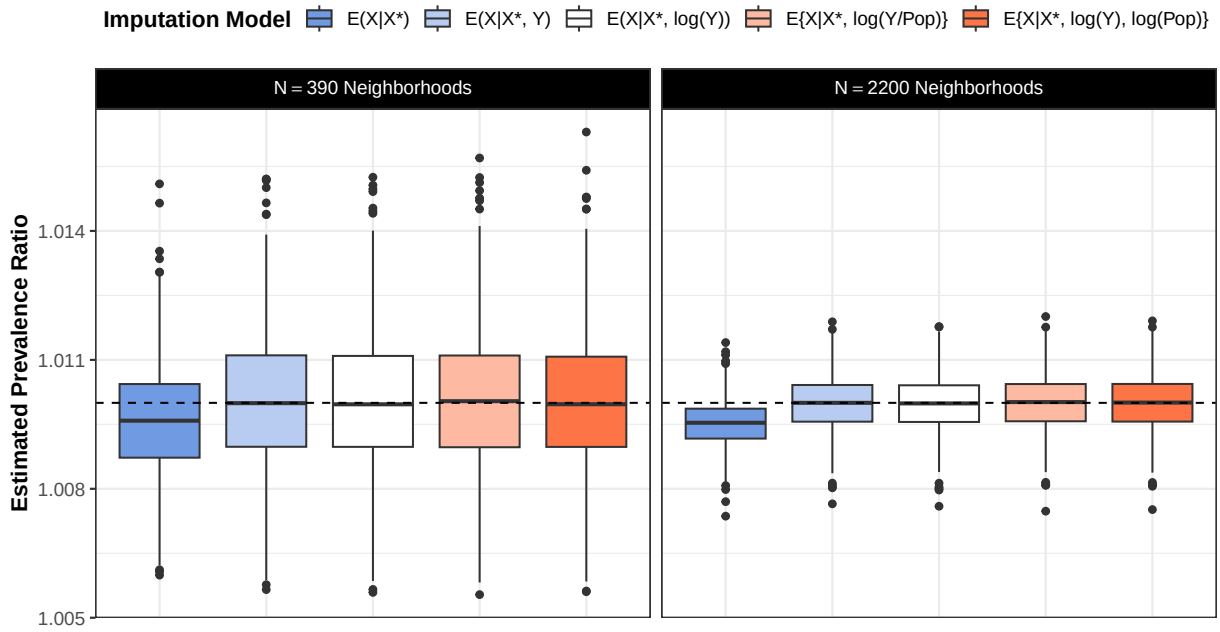

Figure S3: Estimated prevalence ratios for map-based food access X on health using multiple imputation. The five possible ways to include the analysis model outcome Y (with or without the model offset Pop) in the imputation model for X were considered. All results are based on 1000 replications.

Imputing X from a model that excluded the outcome entirely led to estimates of the prevalence ratio that were consistently lower than the truth in all sample sizes considered, albeit by a small amount given the small value of β_1. This bias is consistent with the recommendations from Moons et al. (2006) and D'Agostino McGowan et al. (2024). In-

cluding the untransformed Y or transformed outcome $\log(Y)$ in the imputation model led to unbiased estimates, on average. Incorporating the offset Pop into the imputation model, either in the log prevalence $\log(Y/Pop)$ or on its own as $\log(Pop)$, could lead to estimates with slightly more variability, but they were still unbiased. Based on these results, missing X values were replaced with random draws from the conditional distribution of X given X^* and $\log(Y)$, corresponding to imputation model 3 considered.

Table S1: Simulation results under varied additive errors in straight-line proximity to healthy foods, as controlled by the mean μ_U of the errors U.

		Gold Standard		**Naive**		**Complete Case**			**Imputation**				
N	$\boldsymbol{\mu_U}$	**Bias**	**ESE**	**Bias**	**ESE**	**Bias**	**ESE**	**RE**	**Bias**	**ESE**	**ASE**	**CP**	**RE**
390	−0.10	0.000	0.001	−0.038	0.001	−0.016	0.004	0.079	0.006	0.002	0.002	0.959	0.643
	−0.35	0.004	0.001	−0.041	0.001	−0.026	0.004	0.075	0.009	0.001	0.002	0.966	0.648
	−0.70	0.004	0.001	−0.053	0.001	0.008	0.004	0.093	0.008	0.002	0.002	0.980	0.647
	−1.00	−0.001	0.001	−0.066	0.001	−0.005	0.004	0.080	0.002	0.002	0.002	0.971	0.529
2200	−0.10	0.001	0.000	−0.036	0.000	−0.002	0.001	0.111	0.001	0.001	0.001	0.959	0.704
	−0.35	0.000	0.000	−0.043	0.000	−0.003	0.002	0.098	0.001	0.001	0.001	0.975	0.712
	−0.70	0.000	0.000	−0.054	0.000	0.003	0.002	0.097	0.000	0.001	0.001	0.977	0.601
	−1.00	−0.001	0.000	−0.064	0.000	0.004	0.002	0.086	−0.002	0.001	0.001	0.982	0.531

Note: **Bias** and **ESE** are, respectively, the empirical relative bias and standard error of the log prevalence ratio estimator $\hat{\beta}_1$; **ASE** is the average of the standard error estimator $\widehat{\mathrm{SE}}\left(\hat{\beta}_1\right)$; **CP** is the empirical coverage probability of the 95% confidence interval for the log prevalence ratio β_1; **RE** is the empirical relative efficiency to the Gold Standard. All entries are based on 1000 replicates.

Table S2: Descriptive statistics of the $N = 387$ census tracts in the Piedmont Triad, North Carolina

Variable	Level	Summary
County	Alamance	36 (9.3%)
	Caswell	6 (1.6%)
	Davidson	34 (8.8%)
	Davie	7 (1.8%)
	Forsyth	93 (24.0%)
	Guilford	118 (30.5%)
	Montgomery	6 (1.6%)
	Randolph	28 (7.2%)
	Rockingham	21 (5.4%)
	Stokes	9 (2.3%)
	Surry	22 (5.7%)
	Yadkin	7 (1.8%)
Population[a]		4095 (3901 – 5282)
Land Area (in Miles^2)[a]		4.70 (1.80 – 19.15)
Population Density (per Mile^2)[a]		842.90 (230.80 – 2037.45)
Proximity to Healthy Foods (in Miles)	Straight-Line	1.01 (0.57 – 2.12)
	Map-Based	1.51 (0.89 – 2.99)
Prevalence of Health	Coronary Heart Disease	0.07 (0.06 – 0.08)
Outcome[b]	Diagnosed Diabetes	0.11 (0.09 – 0.13)
	High Blood Pressure	0.35 (0.32 – 0.38)
	Obesity	0.34 (0.31 – 0.38)

Note: County was summarized by the count (proportion) of census tracts at each level. All other variables were summarized numerically by the median (interquartile range [IQR]). [a]Taken from the 2010 U.S. Census. [b]Taken from the 2022 PLACES dataset (Centers for Disease Control and Prevention, 2022).

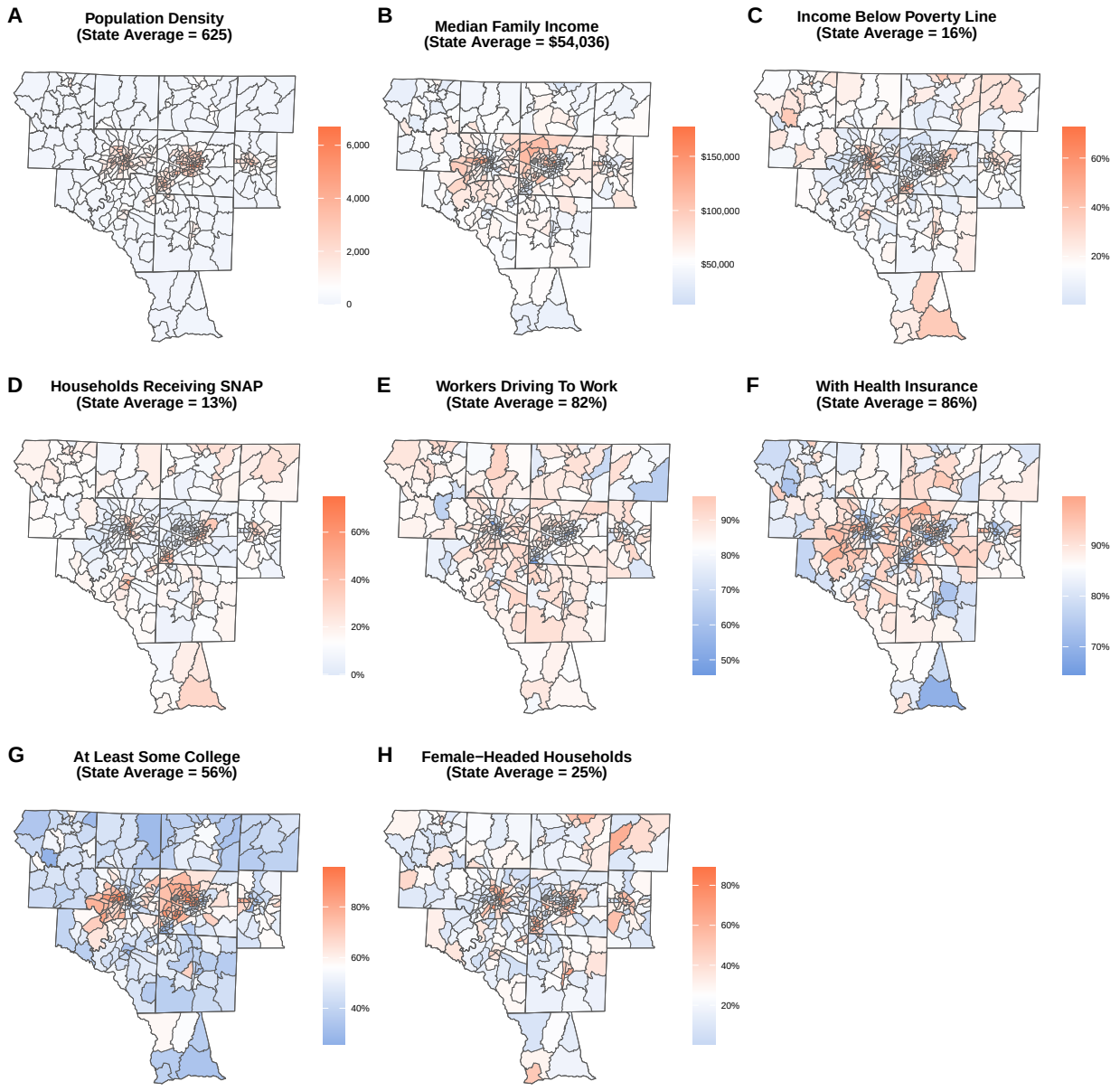

Figure S4: Choropleth maps of socioeconomic factors across the census tracts of the Piedmont Triad, North Carolina. Data were taken from the 2015 American Community Survey. The gradient for each map is centered at the state median. Two census tracts were missing median family income. All other maps are based on $N = 387$ census tracts. The one census tract with zero population was excluded.

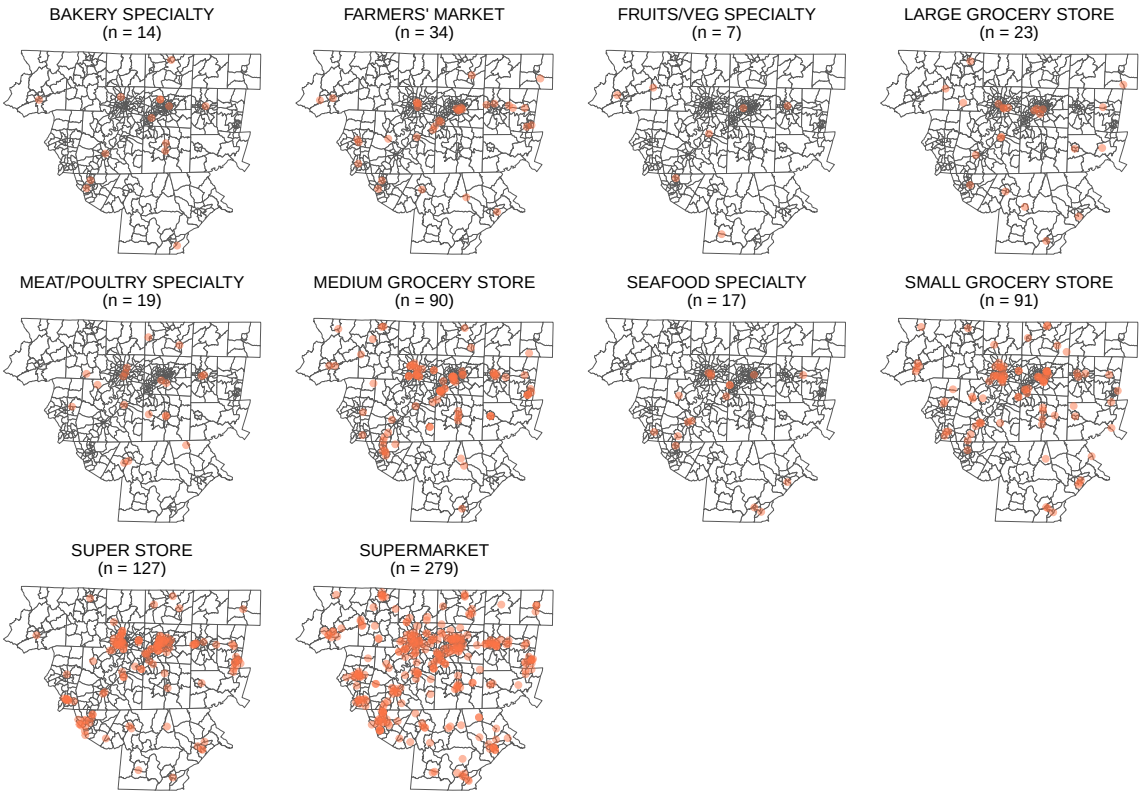

Figure S5: Map of $M = 701$ authorized SNAP retailers in the Piedmont Triad, North Carolina, broken down by store type. Data were taken from the 2022 Historical SNAP Retail Locator Data (United States Department of Agriculture, 2022).

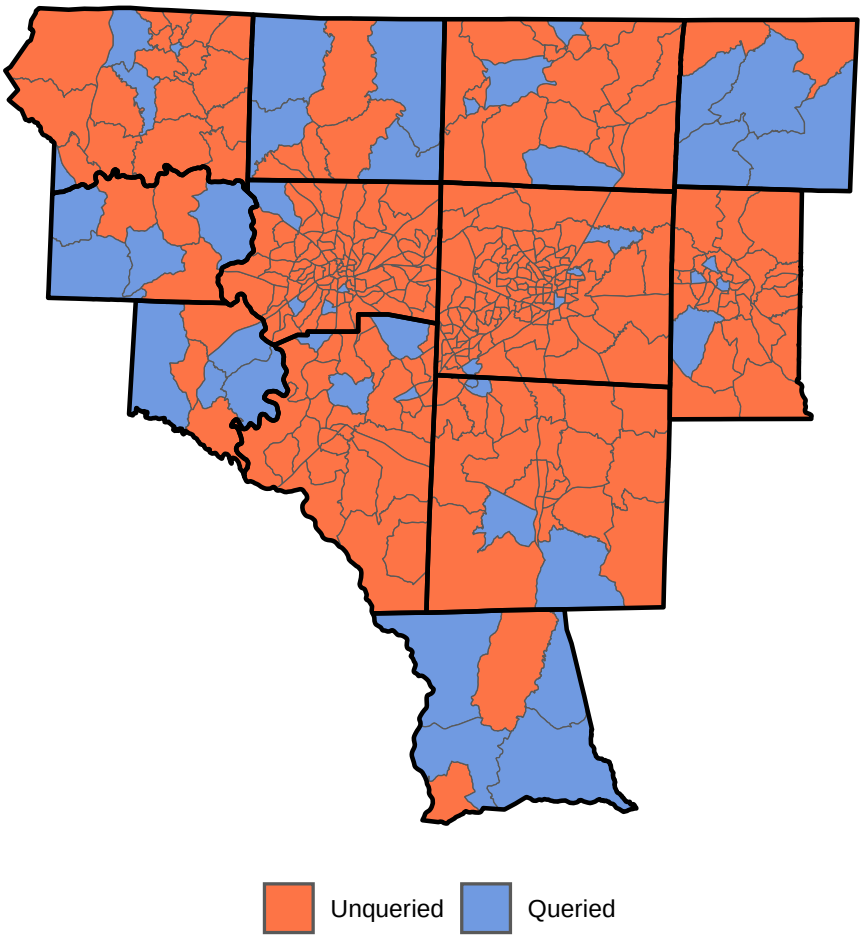

Figure S6: Map of census tracts in the Piedmont Triad, North Carolina, colored according to whether it was treated as queried in the partially queried analysis. For the partially queried analysis, $n = 48$ tracts were chosen to be queried via county-stratified random sampling. The thicker boundaries denote outline the 12 counties.

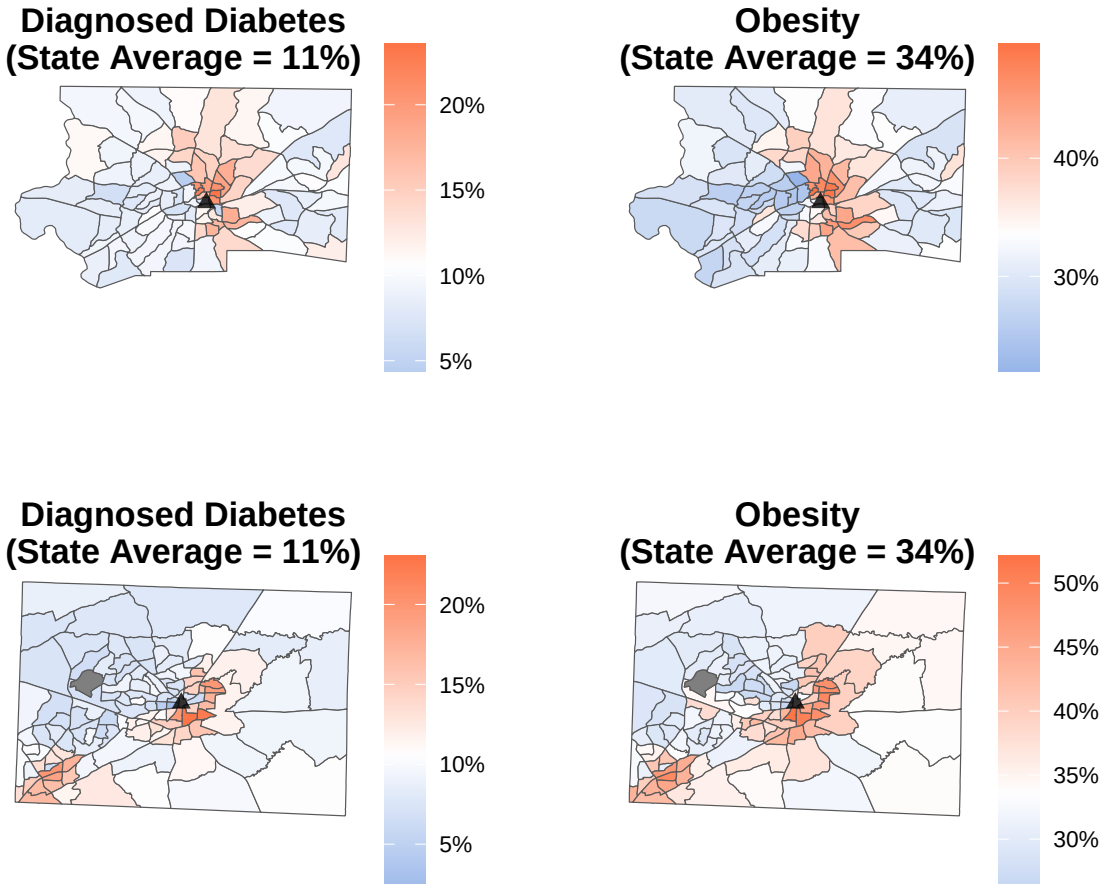

Figure S7: Choropleth maps of the crude prevalence of adverse health outcomes for census tracts in Forsyth County (top row) and Guilford County (bottom row), North Carolina. The triangles denote "downtown" Winston-Salem and Greensboro (represented by their City Halls) in the top and bottom row, respectively. The gradient for each map is centered at the state median. Data were taken from the 2022 PLACES dataset (Centers for Disease Control and Prevention, 2022).

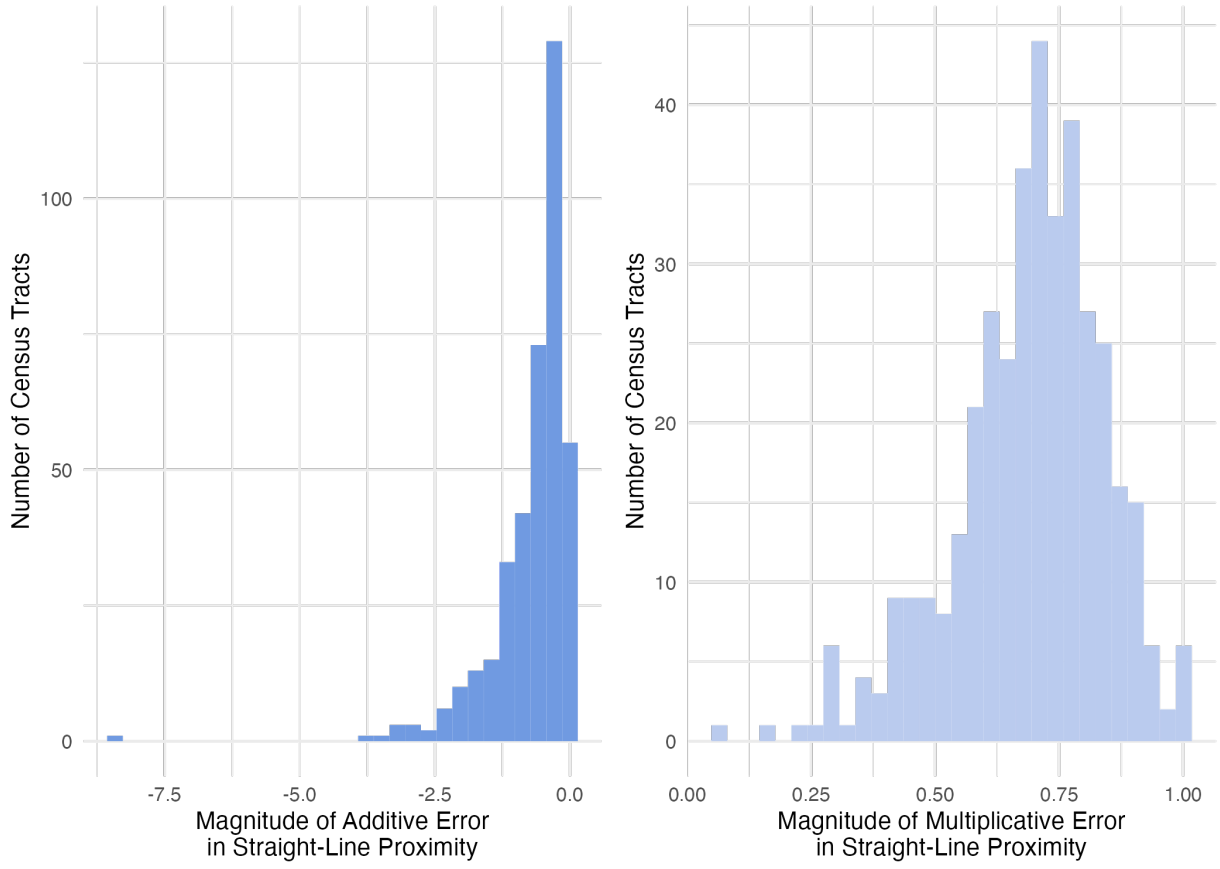

Figure S8: Histogram of additive errors (U) and multiplicative errors (W) in straight-line proximity to healthy foods (X^*) from the fully queried data ($N = 387$) for the Piedmont Triad, North Carolina.

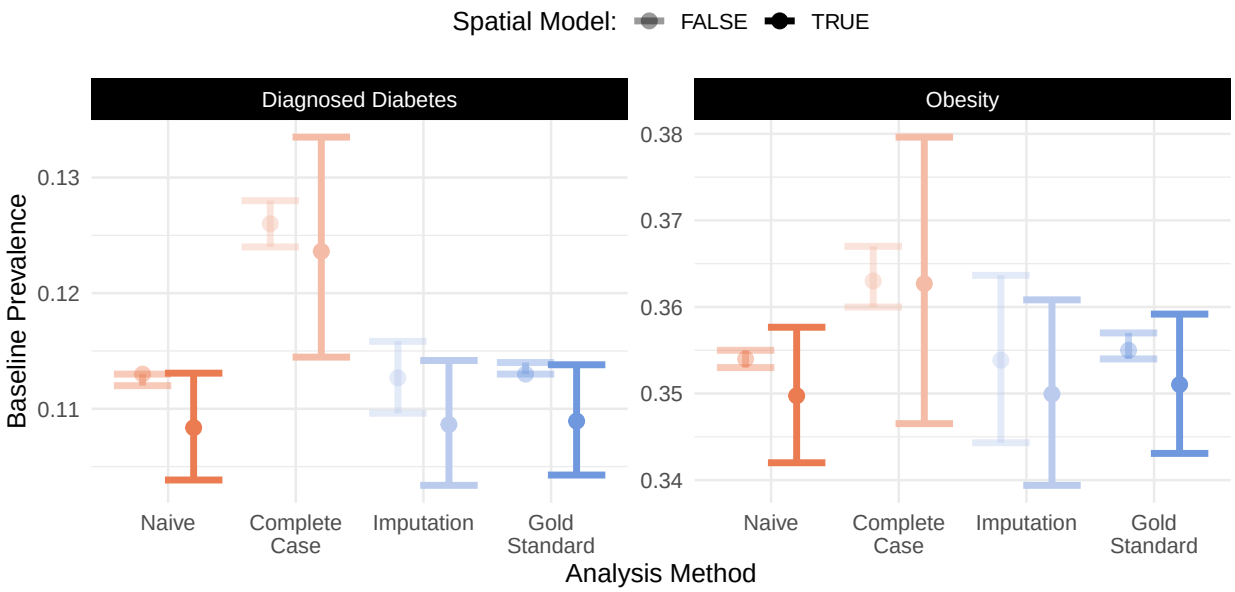

Figure S9: Estimated baseline prevalence (with 95% confidence intervals) for health outcomes in the Piedmont Triad, North Carolina using different analysis methods. Within each health outcome and method, estimates on the right came from the mixed effects model allowing for spatial autocorrelation between bordering neighborhoods (census tracts); estimates on the left came from the model assuming independence between neighborhoods.

References

Centers for Disease Control and Prevention (2022). PLACES. `https://www.cdc.gov/places`. [Online; accessed 20-April-2023].

D'Agostino McGowan, L., S. C. Lotspeich, and S. A. Hepler (2024). The 'Why' behind including 'Y' in your imputation model. *Statistical Methods in Medical Research*. in press.

Moons, K. G. M., R. A. R. T. Donders, T. Stijnen, and F. E. Harrell (2006). Using the outcome for imputation of missing predictor values was preferred. *Journal of Clinical Epidemiology 59*(10), 1092–1101.

United States Department of Agriculture (2022). Historical SNAP Retailer Locator Data. `https://www.fns.usda.gov/snap/retailer/historicaldata`. [Online; accessed 21-July-2023].