

Supplementary materials

In the primary analysis, models were fit on household-resolution features, but every performance and predictor-importance metric was computed at the census-tract level, so the effective number of independent outcome observations is the 34 census tracts rather than the 475 households. To assess whether the qualitative findings were sensitive to retaining household-resolution features, a fully tract-aggregated complementary analysis was performed.

Household-level predictors were aggregated to the census-tract level using tract-specific median values. Median values were selected because they are less sensitive to extreme observations and provide a robust summary of tract-level conditions. Categorical variables were summarized using the most frequently occurring category within each tract. This process resulted in 34 independent tract-level observations, corresponding to the 34 census tracts represented in the study area. The outcome variable remained census-tract obesity prevalence. The primary models were then re-estimated using the aggregated tract-level dataset and evaluated using five-fold cross-validation. For this reduced dataset, the 34 tract-level observations were partitioned into five approximately equal folds by random assignment; the household-coordinate KMeans spatial blocks described in Table S3 pertain to the primary household-resolution analysis and were not applied to the tract-aggregated sensitivity analysis.

Table S1 presents fold-specific prediction performance for the Null Mean Model, Elastic Net, Random Forest, Gradient Boosting, and GAM-like spline models. The Null Mean Model served as a simple baseline and predicted obesity prevalence using only the mean obesity prevalence from the training tracts, without using any environmental, demographic, housing, transportation, or soil-related predictors. The purpose of the Null Mean Model was to establish the level of prediction accuracy that could be achieved without any explanatory information. Models that outperform the Null Mean Model demonstrate that the predictor variables contain meaningful information related to obesity prevalence.

The table reports prediction performance separately for each of the five cross-validation folds. Reporting results by fold provides insight into model stability and consistency across different subsets of the tract-level data. Although some variation in performance was observed across folds, Random Forest, Gradient Boosting, and GAM-like spline models generally produced lower prediction errors than the Null Mean Model and Elastic Net. These findings indicate that the environmental, demographic, housing, and transportation variables retained predictive value even after aggregation to the tract level. Importantly, the pattern of results remained similar to that observed in the primary analysis. Models that performed well in the primary analysis continued to perform relatively well in the tract-level analysis, while weaker-performing models remained comparatively less effective. The tract-aggregated analysis provided qualitative consistency with the primary analysis, although the small number of tracts and variable fold-level performance preclude determining the extent to which the primary findings were influenced by repeated tract outcomes.

Supplementary Table S1. Fold-specific model performance under the tract-aggregated complementary sensitivity analysis.

Model	Fold	RMSE	MAE	R2
-------	------	------	-----	----

Null Mean Model	1	5.81	4.78	-0.33
Null Mean Model	2	8.48	6.91	-0.18
Null Mean Model	3	9.69	9.20	-0.77
Null Mean Model	4	7.33	6.45	-0.64
Null Mean Model	5	8.18	6.58	-0.11
Elastic Net	1	19.08	11.40	-13.28
Elastic Net	2	4.03	3.34	0.73
Elastic Net	3	6.12	4.69	0.29
Elastic Net	4	4.02	3.24	0.51
Elastic Net	5	14.95	10.91	-2.70
Random Forest	1	4.15	3.33	0.33
Random Forest	2	4.77	3.93	0.63
Random Forest	3	5.69	4.53	0.39
Random Forest	4	3.81	3.07	0.56
Random Forest	5	6.66	5.33	0.26
Gradient Boosting	1	5.75	4.36	-0.29
Gradient Boosting	2	3.92	3.50	0.75
Gradient Boosting	3	6.21	4.96	0.27
Gradient Boosting	4	5.21	4.34	0.17
Gradient Boosting	5	7.41	6.23	0.09
GAM-like Splines	1	5.38	4.37	-0.13
GAM-like Splines	2	6.55	5.81	0.29
GAM-like Splines	3	7.42	6.78	-0.03
GAM-like Splines	4	4.56	4.24	0.37
GAM-like Splines	5	7.53	6.19	0.06

Table S2 presents permutation-based predictor importance estimates derived from the tract-level Gradient Boosting model. Gradient Boosting was selected because it was one of the strongest-performing models in the primary analysis and remained among the top-performing models in the tract-aggregated complementary sensitivity analysis. Predictor importance was calculated using permutation importance, a method that measures how much model performance decreases when the values of a predictor are randomly shuffled. Larger importance values indicate variables that contribute more strongly to predictive performance.

The results show that distance to airport remained the most influential predictor in the tract-level analysis. Other highly ranked predictors included measures of racial composition (Proportion White and Proportion Black), transportation accessibility (Distance to Nearest Street and Distance to Major Roads), housing characteristics (Home Value on Zillow and Taxable Home Value), and environmental metal concentrations, particularly lead (Pb). Population characteristics and land area measures also contributed to model performance. The persistence of several leading predictor domains provided qualitative consistency with the primary analysis; however, the small number of tracts and variable fold-level performance preclude conclusions regarding robustness or stability. Several of the same variables identified as influential in the primary analysis remained among the highest-ranked predictors in the tract-level analysis. The persistence of several leading predictor domains

provided qualitative consistency with the primary analysis, although the small number of tracts and variable fold-level performance preclude determining the extent to which the primary rankings were influenced by repeated tract outcomes.

The general advantage of tree-based models over the null and regularized linear models was broadly preserved after aggregation to 34 independent tracts, although the ordering of the individual models differed and Random Forest rather than Gradient Boosting produced the lowest average RMSE, providing qualitative consistency with the primary findings, although the small number of tracts and variable fold-level performance preclude definitive conclusions regarding the influence of repeated outcome values. However, fold-level performance was highly variable, with several folds yielding negative R^2 values (Table S1), so the aggregate accuracy should be regarded as unstable and the tract-level results as exploratory. The persistence of several leading predictor domains provided qualitative consistency with the primary analysis. Given the small number of tracts and variable fold-level performance, these results should not be interpreted as demonstrating robustness or stability of the predictor rankings.

Supplementary Table S2. Predictor importance from the tract-aggregated gradient boosting complementary sensitivity model.

Feature	Mean	SD
Distance_to_Airport_in_Meters	4.31	0.48
Proportion_White	2.96	0.33
Distance_to_Nearest_Street_in_Meters	2.44	0.33
Proportion_Black	0.97	0.13
Area_Total	0.82	0.11
Home_Value_on_Zillow	0.42	0.09
Distance_to_Major_Roads_in_meters	0.39	0.08
Population2010	0.35	0.06
Total_population	0.26	0.07
Proportion_Hispanic_or_Latino	0.25	0.05
White_alone	0.27	0.07
Area_Land	0.24	0.05
Se (µg/kg)	0.18	0.06
Pb (mg/kg)	0.83	0.12
Zn (mg/kg)	0.05	0.02
Cu (mg/kg)	0.03	0.01
Cd (µg/kg)	0.03	0.02
Some_Other_Race_alone	0.16	0.03
Taxable_Home_value	0.14	0.04
Proportion_American_Indian	0.05	0.02
Percent_Soil_Clay	0.02	0.01
Year_House_built	0.02	0.01
Percent_Soil_Silt	0.00	0.00
Soil_Symbol	0.00	0.00

Population Density per sq mile	0.00	0.00
Existing Vegetation Type 2010	0.00	0.00
Native Hawaiian and Other Pacific Islander alone	0	0
Asian alone	0	0
Area Water	0	0
Hispanic or Latino	0	0
Percent Soil Sand	0	0
Soil pH	0	0
soil organic matter	0	0
American Indian and Alaska Native alone	0	0
Black or African American alone	0	0
Two or More Races	0	0

Because this study is a proof-of-concept with a small number of independent tract-level outcomes, model hyperparameters were prespecified and held constant rather than optimized through grid, random, or nested search. Supplementary Table S3 reports the complete fixed settings for every model, so that all results can be reproduced exactly. All analyses used a single global random seed (42) applied to NumPy, PyTorch, and scikit-learn. Models were evaluated using 5-fold GroupKFold cross-validation in which folds were defined by spatial blocks obtained from KMeans clustering of household coordinates, so that geographically proximate households were held out together.

Supplementary Table S3. Fixed model hyperparameters and reproducibility settings.

Component	Implementation	Fixed settings
Random seed	NumPy, PyTorch, scikit-learn	42 (global; applied to all stochastic steps)
Cross-validation	scikit-learn GroupKFold	5 folds; groups = spatial blocks
Spatial blocking	scikit-learn KMeans	15 blocks; n_init = 20; on household UTM coordinates
Numeric preprocessing	SimpleImputer, StandardScaler	Median imputation; standardized
Categorical preprocessing	SimpleImputer, OneHotEncoder	Most-frequent imputation; one-hot encoded
Elastic Net	scikit-learn	alpha = 0.05; l1_ratio = 0.5; max_iter = 50,000
Random Forest	scikit-learn	n_estimators = 700; min_samples_leaf = 2
Gradient Boosting	scikit-learn HistGradientBoosting	learning_rate = 0.05; max_depth = 6; max_iter = 800
GAM-like splines	scikit-learn SplineTransformer + Ridge	n_knots = 5; degree = 3; Ridge alpha = 20.0
MLP	scikit-learn MLPRegressor	hidden_layer_sizes = (128, 64); max_iter = 3,000

CNN-1D	PyTorch	Conv1d(1→8) and Conv1d(8→16), kernel = 5, padding = 2; ReLU; adaptive average pooling; dense head (16→32→1); 25 epochs
Transformer	PyTorch	d_model = 32; 4 attention heads; 1 encoder layer; 15 epochs
Autoencoder	PyTorch	15 epochs
PyTorch optimization	Adam	learning rate = 0.001; batch size = 128; MSE loss
Stacked ensemble	scikit-learn StackingRegressor	Base learners: Elastic Net, Gradient Boosting, Random Forest
Predictor importance	scikit-learn permutation_importance	50 repeats; scored by negative mean squared error

Supplementary Figure S1. Household-resolution predictions plotted against the corresponding census-tract obesity prevalence. Multiple points share the same observed value because households within a tract were assigned the same tract-level outcome. The plot is descriptive and should not be interpreted as representing 475 independent outcome observations.

