

Predicting census-tract obesity prevalence using integrated environmental, built-environment, and neighborhood data: a scale-aware geospatial machine-learning proof-of-concept

Doreen Jehu-Appiah,^{1,2,3} Emmanuel Obeng-Gyasi^{1,3}

¹Department of Built Environment, North Carolina A&T State University, Greensboro, North Carolina; ²Department of Computational Data Science and Engineering, North Carolina Agricultural and Technical State University, Greensboro, North Carolina; ³Environmental Health and Disease Laboratory, North Carolina A&T State University, Greensboro, NC, United States

Abstract

Obesity prevalence varies markedly across neighborhoods, reflecting environmental, built-environment, and social influences, yet few studies integrate primary environmental measurements with contextual spatial data under a scale-aware predictive framework. As a proof-of-concept, we evaluated whether household-resolution measurements, laboratory-measured soil metals (Pb, Cd, Zn, Se, Cu) together with housing, soil-context, built-environment, and sociodemographic variables, can be harmonized into a spatial-block machine-learning framework to generate small-area obesity predictions for Greensboro, North Carolina, USA. Predictors were assembled for 475 geocoded households spanning 34 census tracts and aligned across heterogeneous spatial supports. Models were fitted using household-resolution predictors, with predictions aggregated to the census-tract level for evaluation. Five-fold cross-validation used coordinate-based spatial blocks, which reduced geographic proximity between training and testing observations but did not guarantee complete separation of census tracts. In the primary exploratory analysis, Gradient Boosting produced the lowest mean prediction error (root mean squared error, $RMSE = 4.57 \pm 1.53$; $R^2 = 0.453 \pm 0.235$), although performance was modest and variable across folds; these values are primary exploratory estimates rather than validated generalization performance. A fully tract-aggregated analysis provided a complementary evaluation using 34 independent observations and indicated that tree-based models remained comparatively strong, although model rankings were not identical. Contextual demographic, housing, and spatial-gradient variables contributed more strongly than measured soil metals; because the PLACES outcome is itself modeled from demographic and socioeconomic structure, these leading predictors are partly circular with the outcome and are best read as markers of structural context. These findings demonstrate the feasibility of the data-integration framework but should not be interpreted as definitive estimates of predictive accuracy or causal environmental effects.

Key words: obesity prevalence; geospatial machine learning; census tracts; soil metals; spatial cross-validation.

Correspondence: Emmanuel Obeng-Gyasi, Department of Built Environment, North Carolina A&T State University, 1601 E. Market Street, Price Hall 118, Greensboro, North Carolina 27411, United States. E-mail eobenggyasi@ncat.edu

Introduction

Obesity is a major global public health concern and a leading risk factor for cardiovascular disease, type 2 diabetes, several cancers, and premature mortality (GBD 2017 Risk Factor Collaborators, 2018; Hruby, 2015). In the United States, obesity prevalence has increased steadily over the past several decades and now affects more than 40% of adults (Hales, 2020). Although obesity is often framed as an individual-level condition driven by diet and physical activity, a growing body of research demonstrates that its distribution is strongly shaped by social, environmental, and structural factors operating at multiple spatial scales (Diez Roux & Mair, 2010; Swinburn, 2011). Obesity prevalence does not occur uniformly across space. Substantial geographic variation has been documented across states, counties, and smaller neighborhood units, even within the same metropolitan region (Befort *et al.*, 2012; Robert & Reither, 2004). Such spatial heterogeneity reflects

differences in contextual exposures, access to resources, environmental burdens, and built-environment conditions. Understanding this variation requires moving beyond individual risk factors and examining the environments in which people live. Census tracts provide a meaningful and policy-relevant unit for this analysis. Designed to approximate neighborhoods and typically containing 2,500–8,000 residents, census tracts capture localized contextual conditions while remaining large enough for stable population-level estimates (U.S. Census Bureau, 1994). Many public health programs, environmental assessments, and equity-based resource allocation strategies operate at or near this scale. Accordingly, tract-level obesity prediction supports place-based public health planning and aligns with environmental justice priorities (Braveman *et al.*, 2011).

Obesity is a multifactorial condition. Individual-level determinants include dietary intake, physical activity, and genetic susceptibility. These operate alongside contextual determinants: the food environment and access to affordable, nutrient-dense foods

(Drewnowski, 2009), and its influence on physical activity (Sallis *et al.*, 2012), and socioeconomic conditions shaping exposure and opportunity (Braveman *et al.*, 2011; Diez Roux & Mair, 2010). Environmental contaminant exposure represents an additional, less well-characterized domain (Heindel *et al.*, 2017; Swinburn, 2011). The present study examines measured environmental exposures as one component among these determinants rather than as a primary driver.

Tract-level obesity prevalence and neighborhood heterogeneity

Marked variation in obesity prevalence across census tracts has been widely documented (Robert & Reither, 2004; Befort *et al.*, 2012). This heterogeneity reflects the cumulative influence of environmental, built-environment, and neighborhood social conditions. Neighborhood context influences dietary patterns, opportunities for physical activity, psychosocial stress, environmental exposures, and access to health-promoting resources (Diez Roux & Mair, 2010; Sallis *et al.*, 2012). These contextual domains interact over time, shaping population-level health outcomes in ways that are not fully captured by individual characteristics alone. Environmental exposures may contribute to obesity through multiple pathways. Certain environmental contaminants, including heavy metals, have been linked to metabolic dysregulation, inflammation, and altered endocrine function in experimental and epidemiologic studies (Heindel *et al.*, 2017; Thayer *et al.*, 2012). Soil represents a long-term environmental reservoir for metal accumulation, particularly in urban areas affected by historical industrial activity, traffic emissions, and legacy housing materials (Mielke *et al.*, 2019). Although soil metals are not traditionally central to obesity research, environmental health frameworks increasingly recognize the role of chronic low-level exposures in shaping cardiometabolic outcomes (Heindel *et al.*, 2017). The built environment further influences obesity risk by shaping physical activity, food access, transportation patterns, and neighborhood walkability (Sallis *et al.*, 2012). Features such as land use, street connectivity, proximity to major roads, and housing characteristics are unevenly distributed across neighborhoods and are often patterned by socioeconomic and racial stratification (Diez Roux & Mair, 2010). Social and demographic context, including measures of income, education, racial composition, and population density, further modifies exposure patterns and health opportunities (Braveman *et al.*, 2011). Importantly, obesity prevalence at the census-tract level reflects aggregated population outcomes rather than individual risk. Ecologic analyses at this scale are intended to identify contextual associations and spatial patterns, not to infer individual-level causation (Roux, 2002), yet tract-level modeling remains essential for geographic targeting, environmental remediation priorities, and equitable public health interventions. A central methodological challenge is integrating predictors measured at heterogeneous spatial supports, since failure to address scale alignment can lead to biased inference, ecological misinterpretation, or overly optimistic predictive performance (Openshaw, 1984; Roux, 2002; Bivand, 2008). In this study, we adopt a scale-aware framework that harmonizes laboratory-measured soil metals and contextual spatial variables to the tract level and evaluates predictions using household-coordinate spatial-block cross-validation.

Environmental exposures, context, and scale mismatch

Environmental health research frequently integrates data collected at very different spatial scales. Environmental contaminants

such as heavy metals are often measured at fine spatial resolution, sometimes at individual households, parcels, or sampling points (Filippelli & Laidlaw, 2010; Levin *et al.*, 2008), and urban soil contamination can vary substantially within short distances due to historical land use, traffic corridors, industrial emissions, and housing age. These point-level measurements capture localized exposure but do not align directly with population-level health data. Contextual predictors, by contrast, are typically derived from spatially aggregated datasets: built-environment variables such as land use, transportation networks, street density, vegetation cover, and zoning are represented as polygons or grid cells (Brownson *et al.*, 2009; Frank *et al.*, 2004), and soil classifications from national surveys represent mapped units that average multiple soil characteristics across defined areas. Health outcomes such as obesity prevalence are likewise available only at aggregated levels, including census tracts, counties, or metropolitan areas (Zhang *et al.*, 2014) these describe the proportion of residents affected within an area rather than individual risk, and associations at this level should be interpreted ecologically rather than causally at the person level (Wakefield, 2008). Combining predictors and outcomes measured at different spatial supports therefore requires explicit methodological care, because statistical associations can change depending on how spatial data are aggregated or partitioned (Fotheringham & Wong, 1991; Wakefield, 2008). Without careful scale alignment, analyses may introduce bias or produce misleading estimates, so harmonizing fine-resolution environmental exposures with tract-level health outcomes demands deliberate aggregation procedures and scale-aware modeling.

Limitations of prior tract-level prediction studies

Small-area prediction of obesity has grown in recent years, but several limitations remain evident.

First, many tract-level models rely heavily on sociodemographic variables derived from census data. Indicators such as poverty, education, racial composition, and housing characteristics are frequently used as (Drewnowski, 2009). While these variables are important markers of structural disadvantage, they often function as proxies for underlying environmental and infrastructural exposures. Reliance on sociodemographic predictors alone may obscure physical environmental influences that contribute independently to metabolic health.

Second, direct environmental exposure measurements are rarely incorporated into tract-level obesity models. Experimental and mechanistic studies suggest that certain environmental contaminants, including metals and endocrine-disrupting chemicals, may influence adipogenesis, insulin signaling, inflammation, and metabolic regulation (Lind *et al.*, 2012). Epidemiological evidence, however, remains inconsistent, and the independent contribution of metal exposure to obesity at the population level has not been established. Soil metals are therefore best regarded as potential environmental determinants whose contribution remains uncertain, rather than as confirmed risk factors. Such exposures are nonetheless seldom integrated into spatial obesity prediction frameworks, partly due to data complexity and scale mismatch, and their measurement is warranted to evaluate whether any contribution is detectable.

Third, predictive modeling studies often use random train–test splits without accounting for geographic clustering. When neighboring observations are randomly divided between training and testing sets, spatial dependence can artificially inflate predictive performance (Brenning, 2012; Valavi *et al.*, 2019). Spatial autocorrelation is common in environmental and neighborhood data (Anselin, 1988), and ignoring it introduces geographic information

leakage, whereby information from the test region is effectively present in the training data. Few prior small-area obesity studies report geographically structured validation or quantify residual spatial autocorrelation, so reported accuracies may reflect leakage rather than transferability to unsampled areas. A related and often unaddressed issue is spatial misalignment: predictors and outcomes are frequently drawn from different spatial supports and combined without explicit harmonization, which can bias associations and further inflate apparent model performance (Fotheringham & Wong, 1991).

Finally, prior studies frequently emphasize linear regression frameworks or limited nonlinear exploration. Yet obesity emerges from complex systems involving environmental, behavioral, and social feedback loops (Hammond, 2009). Nonlinear relationships and cross-domain interactions may be central to understanding tract-level variation. Restricting models to additive linear effects may fail to capture important structural dynamics.

Study innovation

This study advances tract-level obesity modeling in several important ways. We note at the outset that machine-learning integration of environmental, neighborhood, and built-environment variables for small-area obesity prediction is by now well established (Lotfata *et al.*, 2022; Allen, 2023), so novelty here does not rest on that integration itself. The distinctive contribution is instead the coupling of primary, laboratory-measured household soil-metal concentrations with an explicit treatment of spatial-support mismatch and household-coordinate spatial-block cross-validation.

First, we integrate laboratory-measured soil metals with contextual built-environment and neighborhood variables. Unlike many prior small-area studies that rely primarily on census indicators, we incorporate measured environmental contaminants that may reflect long-term cumulative exposure conditions (Filippelli & Laidlaw, 2010; Lind *et al.*, 2012). Their inclusion is exploratory and does not presuppose an association with obesity prevalence.

Second, we explicitly harmonize heterogeneous predictors to the tract-level outcome scale, aligning environmental, soil, and contextual variables through explicit aggregation rather than combining differing spatial supports without adjustment. This directly addresses concerns in spatial epidemiology regarding ecological inference and spatial misalignment (Fotheringham & Wong, 1991).

Third, we adopt leakage-aware evaluation using household-coordinate spatial-block cross-validation. By holding out spatial blocks of proximate households during model testing, we reduce the influence of spatial autocorrelation on performance metrics; because the blocks are defined on household coordinates rather than on tract boundaries, this does not guarantee that every household from a given tract is held out together. This analysis provides a more geographically conservative evaluation than random household splitting (Brenning, 2012; Valavi *et al.*, 2018).

Fourth, we compare interpretable statistical models with flexible machine learning approaches. Regularized regression and spline-based models offer transparency and partial interpretability, while ensemble and neural network models capture nonlinear patterns and higher-order interactions (James *et al.*, 2023).

Study objectives

The objectives of this study are as follows: i) to predict census-tract obesity prevalence using an integrated set of predictors including laboratory-measured soil metals, soil characteristics, built-environment metrics, and neighborhood demographic variables; ii) to identify influential predictors, nonlinear effects, and cross-domain interactions that contribute to tract-level predictive

variation in obesity prevalence; iii) to evaluate predictive performance at the census-tract level (34 independent outcomes) under leakage-aware cross-validation.

Materials and Methods

This is a cross-sectional, ecological (small-area) study that uses a predictive spatial machine-learning framework. It is ecological because the outcome, obesity prevalence, is defined at the census-tract level rather than for individuals; it is predictive rather than causal because the goal is to estimate tract-level prevalence from environmental, built-environment, and neighborhood variables rather than to identify causal effects; and it is a spatial machine-learning framework because predictors measured at differing spatial supports are aligned to the tract level and models are evaluated with household-coordinate spatial-block cross-validation. All findings should therefore be interpreted at the tract level and as predictive, not causal.

Data sources and construction

Data used in this study were obtained from several environmental, administrative, and public spatial datasets and combined into a single analytic dataset. Soil samples were collected from 475 household locations within the study area and analyzed in the laboratory for five metals: lead (Pb), cadmium (Cd), zinc (Zn), selenium (Se), and copper (Cu). Households were enrolled using a convenience-based sampling strategy rather than probability (random) sampling. Recruitment was directed toward achieving broad spatial coverage across Greensboro, so that sampled households spanned 34 census tracts distributed across the city rather than clustering in a single area. Because sampling was not random, the households are not a statistically representative sample of the city's population, and the results are not population-representative estimates. This sampling design is revisited as a limitation in the discussion. Lead, copper, and zinc are reported in mg/kg (ppm) and cadmium and selenium in µg/kg (ppb); because the analytical method carries higher detection limits and greater relative uncertainty for cadmium and selenium, these two metals are treated throughout as exploratory indicators rather than definitive exposure measures, and their tract-level values are consistent with reported North Carolina background soil concentrations. Laboratory analyses followed established environmental measurement protocols to ensure reliable and comparable results (Sparks *et al.*, 2022). Soil metals are commonly used as indicators of long-term environmental contamination because they accumulate over time and often reflect historical industrial activity and land use (Tchounwou *et al.*, 2012). Each soil measurement represents a point-level environmental observation at a specific household location.

In addition to soil measurements, contextual variables describing the built environment and neighborhood characteristics were incorporated. Housing characteristics, including housing age and property value indicators, were obtained from parcel-level administrative records. Soil context variables, such as soil type and related attributes, were derived from the Soil Survey Geographic Database (SSURGO) maintained by the United States Department of Agriculture Natural Resources Conservation Service (USDA NRCS, 2021). Built-environment variables were derived from publicly available spatial datasets describing land use, transportation networks, and infrastructure, including measures such as distance to major roads and proximity to transportation corridors.

These distance-based measures, including distance to the airport, should be interpreted with care. Within a single city evaluated across a small number of tracts, such variables largely index a household's position along a broad northwest-southeast spatial gradient that co-varies with neighborhood demographic and socioeconomic structure. They are therefore best understood as spatial-gradient proxies rather than as causal transportation or proximity exposures, and this interpretation is developed further in the Discussion. Neighborhood sociodemographic characteristics, including racial composition and population density, were obtained from census block and census tract summaries, which are widely used in small-area health research to represent population context (Krieger *et al.*, 2003). All variables were spatially linked to the geographic coordinates of household locations using Geographic Information System (GIS) procedures (Longley *et al.*, 2015). The resulting dataset contained environmental, housing, soil, built-environment, and neighborhood predictors for each household location. Formally, each observation can be represented as a predictor vector:

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})$$

where $i=1, \dots, 475$ indexes households and p denotes the number of predictors across environmental, built, soil, and neighborhood domains.

Outcome variable

The outcome variable used in this study was census-tract obesity prevalence, denoted as OBESITY_CrudePrev. This variable represents the model-based prevalence of adults with obesity (body mass index ≥ 30 kg/m², computed from self-reported height and weight) within each census tract, obtained from the U.S. Centers for Disease Control and Prevention PLACES project. PLACES provides small-area estimates generated from Behavioral Risk Factor Surveillance System (BRFSS) survey data using multilevel regression and poststratification, such that between-tract variation is driven largely by the demographic composition of each tract together with area-level socioeconomic structure. A direct consequence is that some predictors in the present models overlap with the types of variables used to generate the PLACES estimates. Tract racial composition and area-level socioeconomic indicators such as home value resemble inputs to the underlying small-area estimation, so their association with modeled obesity prevalence is partly circular: it may reflect the structure of the outcome model rather than an independent relationship. We therefore treat these variables as markers of structural context rather than as evidence of independent effects, and we return to this circularity as a central interpretive limitation in the Discussion. The variable was treated as a continuous outcome. Because obesity prevalence is defined at the census-tract level, it determines the analytic scale of the study and the level at which model predictions are interpreted. Let $t=1, \dots, T$ index census tracts and let Y_t represent the observed obesity prevalence for tract t . All model predictions and evaluation metrics were ultimately summarized at the tract level to maintain consistency between the predictors and the outcome scale.

Predictor domains, spatial support, and analytic scale

Predictors used in this study come from multiple spatial sources, including point-level environmental measurements, parcel-level administrative data, soil map units, built-environment

datasets, and census-based summaries. Because variables measured at different spatial scales can produce misleading relationships if not properly aligned, all predictors were harmonized to a common spatial reference based on household locations (Gotway & Young, 2002). Each household location was assigned to a census tract using geographic coordinates and tract boundary data, ensuring that all observations were linked to the correct tract. Because the outcome is defined at the census-tract level, all predictions, performance metrics, and substantive interpretations were ultimately summarized at the tract level. Accordingly, results should be interpreted as ecological associations that reflect patterns across geographic areas rather than individual-level relationships.

Analytic unit and a note on inference

Although environmental and housing predictors were measured at 475 individual household locations, the obesity outcome is defined only at the census-tract level, and the 475 households fall within 34 tracts. Because household-level soil measurements capture within-tract environmental heterogeneity that would be lost under simple tract aggregation, the household-level feature matrix was retained for this analysis, while all performance summaries were computed at the tract level. In the analyses reported here, models were trained at the household level: each household was assigned the obesity prevalence of its tract, so each tract's single obesity value is repeated across all of its sampled households, and model predictions were subsequently aggregated to the tract level for evaluation. This design preserves the fine spatial resolution of the soil measurements, but it means that the number of independent outcome observations is 34 rather than 475. This repetition of the outcome (a form of pseudo-replication) tends to understate uncertainty and can inflate apparent predictive performance and the stability of predictor-importance rankings. For this reason, all results are presented as preliminary, exploratory, and hypothesis-generating, and the predictors are additionally aggregated to the 34 tracts as a complementary tract-aggregated sensitivity analysis (*Supplementary Materials*), and multilevel or measurement-error models with a tract-level random effect are identified as the next methodological step. The household-level framework was adopted for a specific, limited purpose: it is the resolution at which the environmental predictors are measured, and it preserves within-tract exposure heterogeneity that aggregation removes. It is not presented as statistically preferable for tract-level inference. Aggregation would remove the pseudo-replication but discard within-tract variation; hierarchical and mixed-effects models would represent household nesting through a tract-level random effect; Bayesian spatial models would borrow strength across tracts and propagate uncertainty; and measurement-error models would account for the gap between measured soil concentration and true exposure. Each is better suited to formal inference and is identified as a next step. The present analysis is a proof-of-concept that measured environmental data carry predictive signal at their native resolution, not a substitute for these approaches.

Preprocessing

Missing values were imputed within each cross-validation fold using median imputation for continuous predictors and most-frequent-category imputation for categorical predictors, so that no information from held-out folds entered preprocessing. Continuous predictors were standardized:

$$X^* = \frac{X - \mu_X}{\sigma_X}$$

Categorical predictors were encoded using one-hot encoding within cross-validation folds to prevent information leakage (Kuhn & Johnson, 2013).

Validation and performance metrics

Model performance was evaluated using five-fold grouped cross-validation. The dataset was partitioned into five folds based on spatial blocks, ensuring that observations from the same spatial block were assigned to the same fold. In each iteration, approximately 80% of the spatial blocks were used for model training and the remaining 20% were used for testing. This approach reduces spatial leakage between training and testing data and provides a more geographically conservative evaluation than random household splitting (Brenning, 2012; Roberts *et al.*, 2017).

Predictions generated during cross-validation were aggregated to the census-tract level, and model accuracy was evaluated using three standard metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2). Lower values of MAE and RMSE indicate better predictive accuracy, while higher R^2 values indicate greater explanatory power. All models were fitted by minimizing squared error, and comparative model performance was ranked using RMSE averaged across the five spatial folds, with MAE and R^2 reported alongside. All reported metrics are computed on held-out folds under spatial-block cross-validation and therefore reflect held-out rather than in-sample performance. Because the spatial blocks were defined on household coordinates rather than tract boundaries, this blocking did not guarantee that all households within a census tract were assigned to the same fold, so the resulting estimates should be interpreted as provisional rather than as fully leakage-free measures of geographic generalizability. An additional optimism correction (for example, the bootstrap 0.632 estimator) was not applied, as the grouped out-of-fold design already excludes training data from each estimate and such corrections are unstable with a small number of independent units.

Mean Absolute Error (MAE)

Mean Absolute Error measures the average magnitude of prediction errors without considering their direction. It provides a direct interpretation of the typical absolute difference between observed and predicted values.

$$MAE = \frac{1}{T} \sum_{t=1}^T |Y_t - \hat{Y}_t|$$

Root Mean Squared Error (RMSE)

Root Mean Squared Error gives greater weight to larger errors because the prediction differences are squared before averaging. RMSE is commonly used in predictive modeling because it penalizes large deviations more strongly than MAE.

$$RMSE = \sqrt{\frac{1}{T} \sum_{t=1}^T (Y_t - \hat{Y}_t)^2}$$

Coefficient of Determination (R^2)

The coefficient of determination measures the proportion of variance in the observed outcome that is explained by the model.

$$R^2 = 1 - \frac{\sum_t (Y_t - \hat{Y}_t)^2}{\sum_t (Y_t - \bar{Y})^2}$$

where T denotes the number of census tracts used in model evaluation, Y_t represents the observed obesity prevalence for tract t , $\hat{Y}_t$ represents the predicted obesity prevalence for tract t , $\bar{Y}$ denotes the mean observed obesity prevalence across all tracts. These metrics quantify prediction accuracy and explained variance (James *et al.*, 2023).

Predictive modeling approaches

Several predictive modeling approaches were evaluated to capture both linear and nonlinear relationships between environmental predictors and obesity prevalence. Elastic Net regression was used as an interpretable baseline model because it combines L1 and L2 regularization to stabilize estimates when predictors are correlated and to perform variable selection (Zou & Hastie, 2005). Generalized Additive Models (GAMs) were also implemented to allow flexible nonlinear relationships between predictors and the outcome while maintaining interpretability (Wood, 2017). To capture more complex interactions among predictors, tree-based ensemble models were applied, including Random Forests and Gradient Boosting, which are well suited for modeling nonlinear relationships and interactions without requiring explicit specification (Friedman, 2001). Neural network approaches were also evaluated to explore more complex feature representations. These architectures (MLP, CNN-1D, Transformer, and Autoencoder) were included as prespecified benchmarks rather than as expected top performers. Deep-learning models typically require large training samples, so with only 34 independent tract-level outcomes they were anticipated to be at a disadvantage. We evaluated them deliberately to test whether more flexible, nonlinear representations could improve on the tree-based models under identical household-coordinate spatial-block validation; their substantially higher held-out prediction errors provided an informative benchmark under the present exploratory evaluation. A Multilayer Perceptron (MLP) was used as a fully connected neural network model, and a one-dimensional Convolutional Neural Network (CNN-1D) was applied to learn structured patterns across ordered predictor vectors. Finally, an ensemble approach was explored by combining predictions from multiple models to leverage complementary strengths and improve overall predictive performance (Wolpert, 1992). To evaluate predictor importance, permutation-based feature importance was used to quantify the decrease in model performance after each predictor was randomly permuted. Permutation importance from the household-resolution primary model was interpreted descriptively. As a complementary sensitivity analysis, permutation importance was recomputed using the fully tract-aggregated Gradient Boosting model, with means and standard deviations across 50 permutations reported in *Supplementary Table S2*. Because the two analyses use different analytic units, the supplementary estimates were used to assess qualitative consistency in the leading predictor domains rather than to quantify uncertainty in the primary ranking. Predictors associated with larger decreases in performance were considered more influential. This approach provides an interpretable measure of the relative importance of environmental, built-environment, soil, and neighborhood variables in predicting obesity prevalence.

All data preprocessing, modeling, validation, and visualization were conducted in Python (version 3.9.18). Data manipulation and

analysis were performed using pandas, NumPy, and SciPy, while machine learning models were implemented using scikit-learn. The convolutional (CNN-1D), Transformer, and autoencoder networks were implemented in PyTorch, and the multilayer perceptron (MLP) used scikit-learn's MLPRegressor. All analyses used a fixed random seed (42) applied to NumPy, PyTorch, and scikit-learn, so that all stochastic procedures are reproducible. Model evaluation used 5-fold GroupKFold cross-validation with folds defined by KMeans spatial blocks of household coordinates, so that geographically proximate households were held out together. Because this study is a proof-of-concept with a small number of independent tract-level outcomes, hyperparameters were prespecified and held constant rather than optimized by grid or nested search; tuning across so few independent outcomes would risk overfitting the validation folds. The complete settings for each model are reported in *Supplementary Table S3*. The analysis code is available in the following GitHub repository: <https://github.com/Doreen-nolda/Predicting-Census-Tract-Obesity>. Spatial data processing, geocoding, census tract assignment, and map production were conducted using ArcGIS Pro (Esri, Redlands, CA, USA). All models were supplied the same set of approximately 23 candidate predictors; the two categorical predictors were one-hot encoded, and no per-model feature selection was performed. Hyperparameters were prespecified and held constant across folds and models; no hyperparameter search was conducted, and nested cross-validation was therefore not applicable. Complete predictor and hyperparameter specifications are given in *Supplementary Table S3*.

Results

Descriptive statistics

Table 1 summarizes the distribution of obesity prevalence, soil metal concentrations, environmental characteristics, and sociodemographic variables across the study area. Mean obesity prevalence was 39.4% (SD = 6.8%), with values ranging from 23.1% to 50.4%. Housing values exhibited substantial variability, with a median value of \$85,100 and a maximum value exceeding \$45 million, indicating a highly right-skewed distribution. Such extreme maxima are common in administrative parcel data, where non-residential, multi-unit, or misrecorded entries can produce values far above typical residential prices; housing value was therefore interpreted as a skewed neighborhood indicator rather than a precise property measure, and its modeled influence should be read with that caution in mind. The median year of house construction was 1953, suggesting that much of the housing stock in the study area was built before the mid-20th century.

Among the measured metals, copper and selenium exhibited the greatest variability, with mean concentrations of 239.9 mg/kg (copper) and 221.6 µg/kg (selenium), respectively, and large standard deviations relative to their means. Lead, cadmium, and zinc also showed considerable variation across census tracts, indicating heterogeneous environmental conditions. Soil characteristics were relatively stable, with mean clay, silt, and sand contents of 29.5%, 26.4%, and 44.0%, respectively. The study area was predominantly composed of Black residents (mean proportion = 0.650) compared with White residents (mean proportion = 0.254), while population density varied substantially across census tracts.

Table 1. Summary statistics (mean, standard deviation, median, minimum, and maximum) for obesity prevalence and all predictor variables used in the analysis.

Variable	Mean	SD	Median	Min	Max
Obesity Prevalence (%)	39.4	6.8	40.6	23.1	50.4
Home Value (\$)	207,310	2074164.5	85,100	0	45,247,489
Year House built	1,954	24.1	1,953	1,887	2,018
Pb (mg/kg)	122.9	84.5	97.8	10.8	416.2
Cd (µg/kg)	106.7	51.9	138	28	190
Zn (mg/kg)	107.2	72.1	85	37	494
Se (µg/kg)	221.6	293	155	2.828	1,650
Cu (mg/kg)	239.9	329.6	128	21	1,581
Percent_Soil_Clay	29.5	1.9	29.4	23.8	39.3
Percent Soil Silt	26.4	3.6	26.1	16.9	39.3
Percent Soil Sand	44	3.8	44.5	26.7	52.1
Soil pH	6.3	0.5	6.6	4.7	6.6
Soil organic matter	0.5	0.2	0.5	0.375	1.6
Distance to Airport (meters)	14,247	2870.2	14797.1	3221.2	23725.9
Distance to Nearest Street (meters)	27.5	13	26.9	0.4	148.4
Distance to Major Roads (meters)	441.3	565.4	336.4	1.3	5795.3
Population Density (persons/sq mile)	4233.8	2600.551	4209.1	0	31366.2
Proportion White	0.254	0.316	0.1	0	1
Proportion Black	0.650	0.344	0.8	0	1
Proportion American Indian	0.007	0.023	0	0	0.2
Proportion Hispanic or Latino	0.069	0.109	0.0	0	0.6

Distribution of obesity prevalence across census tracts

Obesity prevalence varied substantially across census tracts within the study area, ranging from approximately 23% to 50%, indicating marked geographic heterogeneity. This level of tract-level variability is consistent with prior small-area studies that report spatial clustering of obesity within metropolitan regions

(Zhang *et al.*, 2014). The observed variation is consistent with ecological health frameworks emphasizing contextual influences on population health (Pickett & Pearl, 2001).

Figure 1 illustrates the spatial distributions of cadmium (Cd), lead (Pb), copper (Cu), selenium (Se), zinc (Zn), and obesity prevalence across census tracts within Greensboro, North

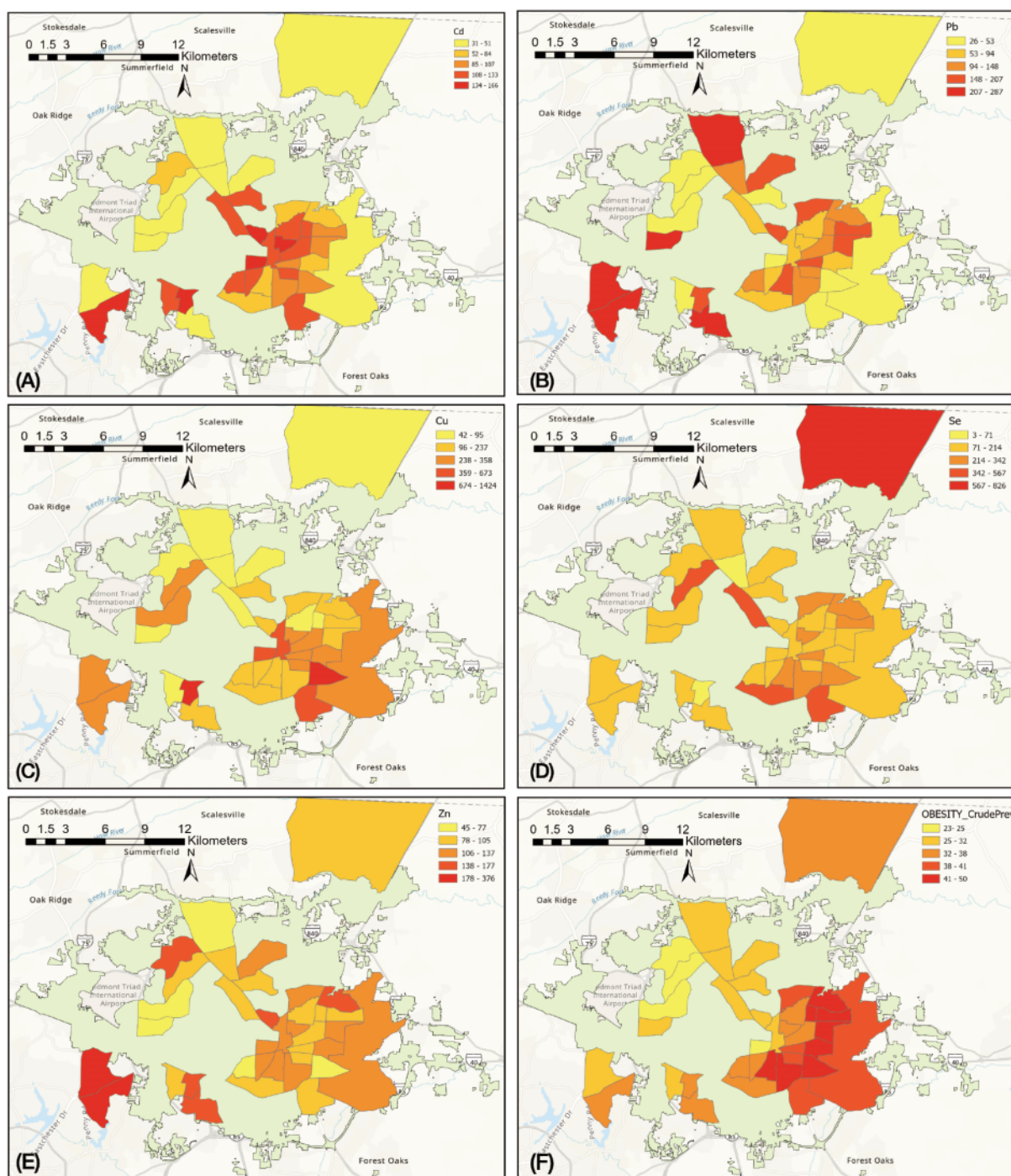


Figure 1. Spatial distribution of soil metal concentrations and obesity prevalence across census tracts within the City of Greensboro, North Carolina. Panels show the spatial distribution of (A) cadmium (Cd), (B) lead (Pb), (C) copper (Cu), (D) selenium (Se), (E) zinc (Zn), and (F) obesity prevalence.

Carolina. Distinct geographic patterns were observed for each metal, indicating that environmental contamination is not uniformly distributed throughout the city. Instead, concentrations tended to cluster within specific census tracts. Obesity prevalence also exhibited marked spatial variability, allowing for visual comparison between environmental exposures and health outcomes across the study area.

Each soil metal showed a distinct, spatially clustered distribution rather than uniform citywide contamination: cadmium ranged from 31–166 µg/kg, lead from 26–287 mg/kg, copper from 42–1,424 mg/kg, selenium from 3–826 µg/kg, and zinc from 45–376 mg/kg. Elevated cadmium, copper, and zinc concentrations were concentrated in central and southeastern tracts, whereas lead hotspots were more widely dispersed and selenium was characterized by a single pronounced northern hotspot. Obesity prevalence ranged from 23% to 50% and followed a clear gradient, with the highest values in central and southeastern tracts and the lowest in northern and northwestern areas, partly overlapping areas of elevated cadmium, lead, copper, and zinc. These maps are descriptive and do not establish causal relationships; the spatial co-occurrence identifies areas that may warrant further environmental health investigation. To further examine how well environmental and contextual predictors predict this variation, multiple predictive models were evaluated and compared against a null baseline that predicts each tract using only the training-set mean obesity prevalence (root mean squared error equal to the outcome standard deviation, approximately 6.8 prevalence points). Because performance varied across spatial folds, these metrics are best interpreted as fold-averaged estimates carrying fold-to-fold uncertainty rather than precise point values. Model performance varied across the evaluated approaches. Tree-based models achieved the strongest predictive performance, with Gradient Boosting producing the lowest RMSE, followed by the GAM-like spline model and Random Forest.

Table 2 summarizes predictive performance across the five spatial cross-validation folds. A null mean model was included as a baseline for comparison. The null model does not use any environmental, housing, soil, built-environment, or sociodemographic predictors. Instead, within each cross-validation fold, it predicts the obesity prevalence of every held-out census tract using only the average obesity prevalence observed among the training tracts. In other words, it assumes that all test tracts have the same obesity prevalence equal to the training-set mean. The purpose of the null

model is to establish the minimum level of predictive performance that can be achieved without using any predictor information. Models that cannot outperform this baseline provide little evidence that the predictors contain meaningful information about spatial variation in obesity prevalence. The null model achieved an RMSE of 8.13 ± 0.99 and an MAE of 6.92 ± 0.74 (with negative per-fold R^2 , as expected for a training-mean predictor evaluated on held-out folds; see *Supplementary Table S1*). Gradient Boosting substantially outperformed this baseline (RMSE = 4.57 ± 1.53 ; MAE = 3.84 ± 1.44 ; $R^2 = 0.453\pm0.235$), reducing prediction error by approximately 44% relative to the null model. Across the five spatial folds these estimates carried wide, descriptive fold-based 95% intervals, reflecting the small number of independent tracts (RMSE 2.67–6.47; MAE 2.05–5.63; R^2 0.16–0.75), and the point estimates should be interpreted accordingly. The best-performing model also reproduced the observed range of tract obesity prevalence, with observed values spanning approximately 23% to 50% and predicted values approximately 24% to 51% (*Supplementary Figure S1*). Random Forest and GAM-like spline models also outperformed the baseline, although to a lesser extent. In contrast, deep-learning approaches, including CNN-1D, Transformer, Autoencoder, and MLP architectures, exhibited considerably larger prediction errors and strongly negative R^2 values, consistent with poorer held-out predictive performance. These results suggest that for this relatively small, spatially structured dataset, tree-based methods captured meaningful predictive signals beyond the average tract obesity prevalence. Under the coordinate-based spatial-block evaluation, the neural-network models produced substantially higher held-out prediction errors than the tree-based models. As a complementary sensitivity analysis, predictors were aggregated to the 34 census tracts and the models were re-estimated using one observation per tract. In the fully tract-aggregated analysis, Random Forest had the lowest average RMSE, followed by Gradient Boosting and the GAM-like spline model; tree-based methods therefore remained among the stronger approaches, although the identity of the top-performing model differed from the household-resolution analysis. Fold-specific results were variable, indicating that model rankings and accuracy estimates are unstable with only 34 tracts (*Supplementary Tables S1–S2*).

Figure 2 shows the top predictors identified using permutation feature importance from the Gradient Boosting model. Distance to airport appears as the most influential predictor. The proportion of Black residents is the second most important predictor, and home

Table 2. Model performance under five-fold spatial cross-validation.

Model	RMSE	MAE	R ²
Null Mean Model	8.13±0.99	6.92±0.74	—
Gradient Boosting	4.57±1.53	3.84±1.44	0.453±0.24
Random Forest	5.16±2.58	4.42±2.22	0.299±0.50
GAM-like Splines	5.23±1.38	4.23±1.34	0.294±0.17
Stacked Ensemble	5.58±3.09	3.96±1.60	0.164±0.69
Elastic Net	11.83±15.96	4.81±2.30	-5.55±13.18
CNN-1D	34.78	33.37	-33.62
Transformer	34.74	34.16	-34.59
Autoencoder	38.94	37.21	-42.45
MLP	80.44±150.51	15.39±16.48	-458.60±1013.19

Values are mean±SD across folds where fold-specific standard deviations were available. R^2 was not summarized for the null model because fold-specific values were negative and dependent on the composition of the held-out folds.

value also ranks among the most influential variables. Built-environment and accessibility variables, including distance to major roads and distance to the nearest street, show measurable but smaller importance. Primary soil metals, by contrast, contributed little: only lead (Pb) showed modest importance, while copper, cadmium, zinc, and the soil-composition indicators (percent clay, soil pH) ranked at or near the bottom of the distribution. Because permutation importance is shared among correlated predictors and the effective sample comprises only 34 independent tracts, the rankings are interpreted qualitatively rather than as precise orderings. *Supplementary Table S2* reports variability for the separate tract-aggregated sensitivity model and is used to assess qualitative consistency in the leading predictor domains, not uncertainty in the primary ranking. The collective pattern indicates that spatial variation in modeled obesity prevalence across these census tracts is driven predominantly by demographic and socioeconomic composition, with environmental soil measures adding limited additional signal. Figure 3 shows the RMSE values for each model across the five spatial cross-validation folds. Each fold represents a different geographic partition of the data, allowing the models to be evaluated on different spatial subsets. The results indicate that tree-based models, such as Gradient Boosting and Random Forest, maintain relatively stable performance across folds, with only small changes in prediction error. In contrast, neural network models show larger fluctuations in RMSE between folds.

Figure 4 presents the correlation structure among the predictor variables and the obesity prevalence outcome. Most predictors show weak to moderate pairwise correlations, suggesting that strong multicollinearity is limited in the dataset. Several demographic variables display notable relationships. For example, the proportion of Black residents shows a positive correlation with

obesity prevalence, while the proportion of White residents shows a negative correlation, reflecting demographic variation across census tracts. Environmental and neighborhood factors also exhibit modest associations with the outcome. Population density, soil characteristics, and proximity to transportation features show small to moderate correlations with obesity prevalence. In addition, some predictors are correlated with each other, particularly among soil composition variables and demographic proportions, which is expected given the geographic clustering of environmental and neighborhood characteristics. The correlation patterns indicate that multiple environmental, demographic, and built-environment factors may contribute jointly to spatial variation in obesity prevalence.

This analysis was included for interpretation rather than variable selection. The tree-based models used here are relatively insensitive to multicollinearity, so no predictors were removed. The matrix instead informs the permutation-importance results: when two predictors move together, one may rank lower because the other absorbs part of its contribution, not because it lacks predictive value. Such predictors are therefore read as a group representing shared structural context rather than as independent effects.

Discussion

This study examined whether census tract obesity prevalence can be predicted using an integrated set of environmental, soil, built-environment, and neighborhood variables within a scale-aware modeling framework. Several important findings emerge. First, under household-coordinate spatial-block cross-validation,

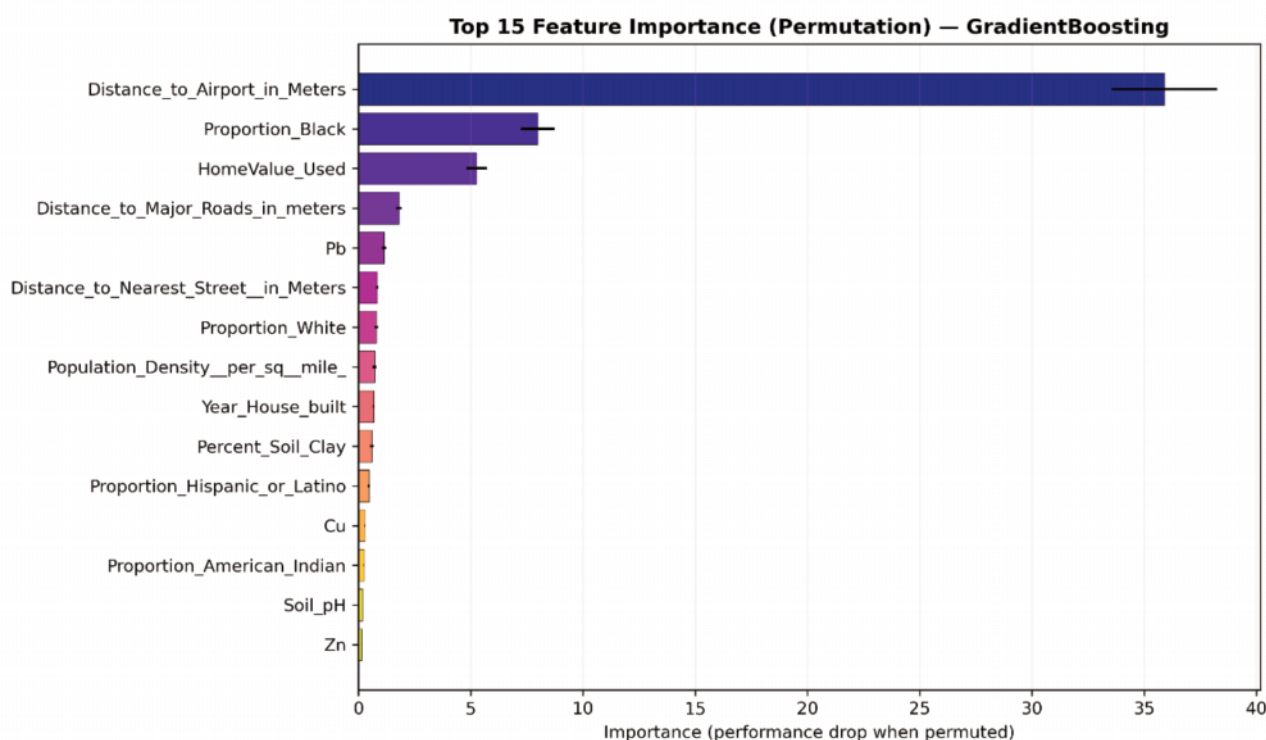


Figure 2. Permutation feature importance for the Gradient Boosting model. Bars show the relative contribution of predictors to model performance, measured as the decrease in prediction accuracy when each feature is permuted.

gradient boosting and spline-based models achieved the strongest performance among the models evaluated, although overall predictive accuracy was modest (R^2 0.453±0.235), whereas neural network architectures performed poorly when evaluated using spatially disciplined validation. This pattern aligns with growing evidence that tree-based ensemble methods often outperform deep learning approaches in structured tabular data settings, particularly when sample sizes are modest and predictor relationships are heterogeneous but not extremely high-dimensional (Grinsztajn *et al.*, 2022; Shwartz-Ziv & Armon, 2022). These results reinforce a central principle in applied data science: model complexity must be matched to data structure. Increasing architectural sophistication does not guarantee improved generalization, especially when spatial clustering is explicitly controlled.

This level of performance is best understood in relation to the validation strategy used. Prior geospatial obesity studies report a broad range of accuracy: county-level explainable models have reported high explanatory power (Allen, 2023), and deep-learning models on satellite imagery have reported R^2 above 0.9 (Dahu *et al.*, 2024). Many of these evaluate performance using random or non-spatial cross-validation, which allows geographically proximate observations to appear in both training and test sets. Because obesity prevalence is spatially autocorrelated, this can inflate apparent accuracy relative to prediction in genuinely unsampled areas. Prior studies have used a range of validation strategies, including random and non-spatial cross-validation. Compared with random household splitting, the coordinate-based spatial blocking used here reduced geographic proximity between training and testing observations. However, because the blocks were not defined by census-tract boundaries, the reported estimates remain exploratory and should not be interpreted as fully leakage-free or directly comparable with estimates from studies using different validation designs. Second, predictor importance analyses were dominated by neighborhood demographic composition, housing value, and a

transportation-distance variable that largely indexes the city's spatial gradient; among primary soil metals, only lead showed modest importance, and the remaining metals and soil-composition indicators contributed negligibly. The prominence of transportation related variables is consistent with prior research linking traffic exposure, environmental stressors, air pollution, and built-environment design to cardiometabolic outcomes (Brook *et al.*, 2010; Cain *et al.*, 2021). This consistency is noted as a possible interpretation rather than a demonstrated pathway, as the transportation-distance variable primarily indexes the city's spatial gradient. Similarly, neighborhood socioeconomic gradients have long been associated with obesity disparities, reflecting differences in food environments, physical activity opportunities, and structural inequality (Diez Roux & Mair, 2010; Drewnowski, 2009). The emergence of these structural variables within a predictive framework indicates that the model is capturing real demographic and socioeconomic patterning rather than spurious statistical noise, though, as detailed below, their partial circularity with the modeled outcome limits any causal reading.

A specific caution applies because the outcome is a modeled estimate rather than a directly observed measure. CDC PLACES tract-level obesity prevalence is generated from self-reported height and weight using multilevel regression and poststratification, so between-tract variation is driven largely by the demographic composition of each tract and area-level socioeconomic structure rather than by independently observed measurements. Several of the most influential predictors identified here, particularly racial composition and housing value, are strong correlates of that same demographic and socioeconomic structure. Because these predictors are closely related to variables already used to generate the PLACES estimates, their reported importance may partly reflect prediction of the PLACES model itself rather than prediction of independently observed obesity prevalence; they are therefore best interpreted as proxies for the broader structural con-

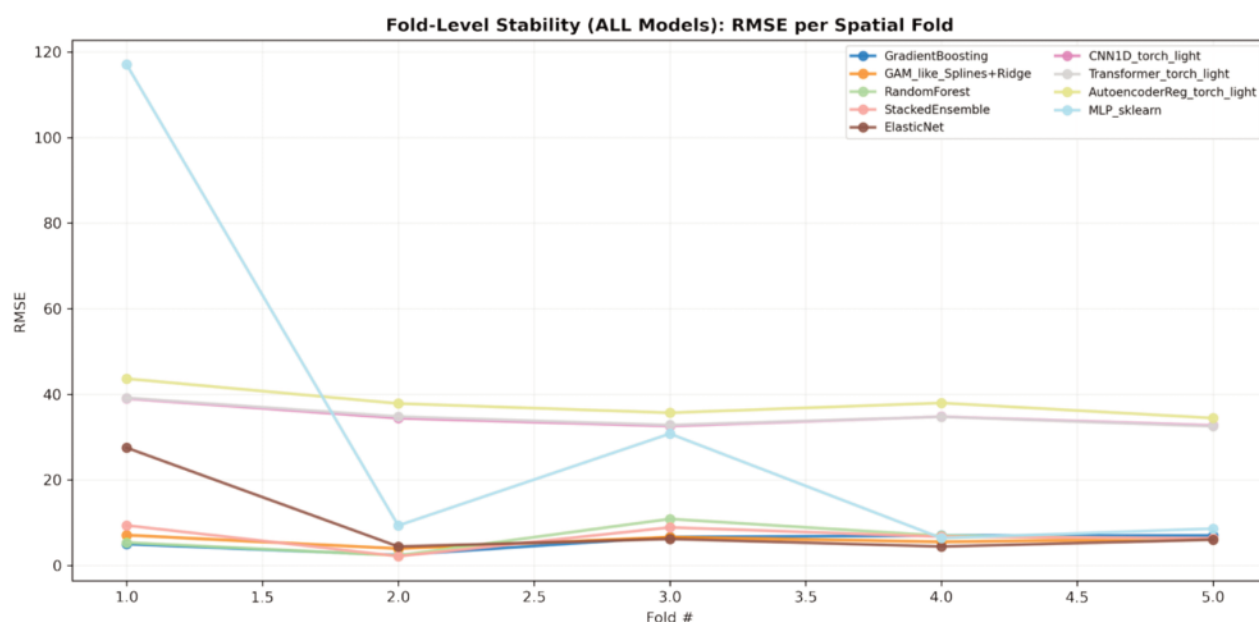


Figure 3. Fold-level model stability across spatial cross-validation folds. Lines show RMSE values for each model across the five spatial folds, illustrating variation in predictive performance across geographic partitions.

ditions that jointly shape modeled obesity prevalence and neighborhood disadvantage, not as independent environmental or built-environment effects. This potential circularity does not negate the spatial patterning described here, but it cautions against reading these contributions as evidence of direct causal pathways. The tract-level associations are intended to prioritize areas and generate testable hypotheses for confirmation with individual-level, directly measured outcomes, and not to establish causation.

Third, the tract-aggregated sensitivity analysis showed qualitative consistency in the leading predictor domains and the relative performance of model classes. However, the differing model rankings, variable fold-level performance, and small number of tracts preclude conclusions regarding robustness or the extent to which the primary findings were influenced by repeated tract outcomes.

In this sensitivity analysis, the predictors were fully aggregated to the 34 census tracts and the models were re-estimated on these tract-level observations. In the fully tract-aggregated analysis the general advantage of tree-based models over the null and regularized linear models was broadly preserved, although the ordering of the individual models differed and Random Forest rather than Gradient Boosting produced the lowest average RMSE (Supplementary Tables S1–S2); the tract-aggregated analysis showed qualitative consistency in the relative performance of model classes, although it cannot establish that the primary findings were unaffected by repeated tract outcomes.

A central contribution of this study lies in its explicit handling of spatial scale. Environmental exposures were measured at the household level, whereas obesity prevalence was defined at the

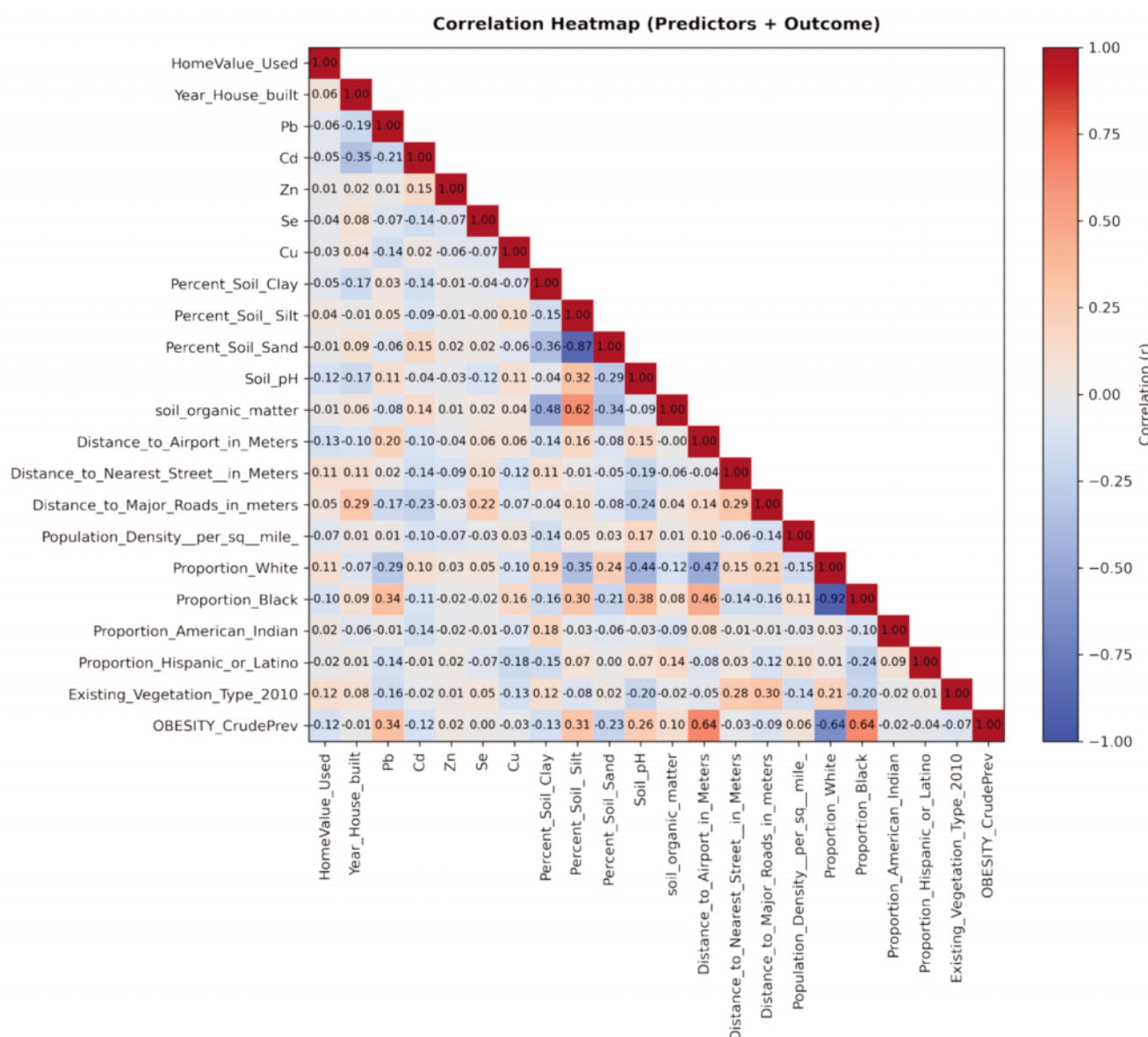


Figure 4. Correlation heatmap of predictors and obesity prevalence. Colors represent Pearson correlation coefficients, with red indicating positive correlations and blue indicating negative correlations.

census-tract level. Without careful harmonization, such mismatches in spatial support can introduce bias, distort associations, or inflate predictive performance (Gotway & Young, 2002). Spatial misalignment is a well-recognized challenge in spatial epidemiology and small-area health modeling (Banerjee *et al.*, 2003). By assigning households to census tracts and conducting household-coordinate spatial-block cross-validation, we reduced geographic proximity between training and testing observations relative to random household splitting but did not guarantee complete separation of census tracts (Roberts *et al.*, 2017). This approach aligns with emerging best practices emphasizing geographically structured validation instead of random splits, which can yield overly optimistic estimates when spatial autocorrelation is present (Meyer *et al.*, 2019). More broadly, the study demonstrates that fine-scale environmental measurements can be integrated with aggregated public health outcomes in a statistically coherent manner when scale is handled explicitly.

Soil metals, including lead, emerged as contextual predictors of tract-level obesity prevalence. Experimental studies have proposed that certain metals could influence metabolic pathways through oxidative stress, inflammation, or endocrine disruption (Ruzzin, 2012), but this is offered here only as a hypothesis. The present ecological design, together with the limited contribution of soil metals in the models, does not permit any inference regarding individual-level exposure-response mechanisms, and no such mechanism is supported by these data. Instead, soil metals should be interpreted as indicators of environmental context. Urban soil contamination frequently co-occurs with older housing stock, historical industrial activity, infrastructure density, and socioeconomic disadvantage (Filippelli & Laidlaw, 2010). In this sense, soil metals may function as markers of broader environmental conditions rather than isolated toxic drivers. The identification of cross-domain interactions, particularly between environmental indicators and neighborhood socioeconomic variables, further supports frameworks of environmental inequality, which posit that environmental burdens and social disadvantage often intersect spatially (Brulle & Pellow, 2006). Such interactions likely reflect complex structural processes rather than simple additive biological effects.

The limited predictive contribution of the soil metals warrants direct consideration. Several non-exclusive explanations are plausible: the modeled outcome is generated primarily from demographic and socioeconomic structure, so aligned predictors dominate and leave little variance for environmental measurements; aggregating household measurements to the tract attenuates fine-scale exposure signal; the small number of independent tracts limits detection of weak or spatially heterogeneous associations; soil concentration is an imperfect proxy for internal or dietary exposure; and any true association may simply be weak at the population scale. Distinguishing among these will require individual-level exposure and biomarker data and larger samples with more independent spatial units.

The spatial distributions shown in Figure 1 provide additional context for the predictive modeling results. Several metals, including cadmium, lead, copper, and zinc, exhibited localized hotspots concentrated within central and southeastern portions of Greensboro, areas that also tended to experience higher obesity prevalence. These spatial patterns do not establish causality; they indicate only that elevated metal concentrations and higher obesity prevalence co-occur in some neighborhoods, a co-occurrence that generates hypotheses for future individual-level study rather than evidence of a shared mechanism. Conversely, some northwestern census tracts near Oak Ridge and Summerfield exhibited relatively low metal concentrations despite moderate to high obesity prevalence,

indicating that environmental contamination alone is unlikely to account for the geographic variation in obesity. These observations support the multifactorial nature of obesity and underscore the importance of considering environmental, socioeconomic, and built-environment factors simultaneously. Importantly, all interpretations are ecological. Census-tract obesity prevalence represents aggregated patterns, and associations identified at the tract level cannot be translated directly to individuals. This distinction is fundamental in ecological epidemiology, where group-level relationships do not necessarily mirror individual-level mechanisms (Morgenstern, 1995). The purpose of this study was predictive and descriptive at the small-area level, not causal inference.

Methodologically, this study demonstrates the feasibility of integrating laboratory-generated environmental exposure data with public spatial and census datasets in a unified predictive framework. Differences in spatial resolution and measurement support are often cited as barriers to cross-domain integration in environmental health research (Jerrett *et al.*, 2005). Our approach provides a reproducible template for addressing such challenges. As a secondary comparison, neural-network models including a one-dimensional convolutional network were also evaluated. These did not outperform the simpler tree-based methods under spatial validation; rather than a deep-learning contribution, the relevant methodological result is that parsimonious tree-based models were more reliable than neural networks at this sample size (LeCun *et al.*, 2015). The comparative evaluation across interpretable statistical models and flexible machine learning approaches clarifies when increased complexity enhances performance and when simpler, well-regularized models remain preferable.

Several strengths and limitations should be acknowledged. A major strength is the use of primary laboratory-measured soil metals, which reduce exposure misclassification compared to modeled or proxy data (Filippelli & Laidlaw, 2010). The explicit spatial validation provides a more geographically conservative evaluation than random splitting. However, limitations include the ecological nature of the outcome, compression of within-tract variability, potential temporal misalignment across datasets, and modest sample size relative to predictor dimensionality. External validation in additional geographic settings will be essential to assess transportability.

This practical application illustrates the framework's potential translational value. A local health department could use the tract-level predictions as a screening tool to prioritize areas for follow-up: tracts where elevated measured soil-metal concentrations co-occur with higher predicted obesity prevalence could be flagged for targeted environmental sampling or community health assessment, directing limited field resources to where measured environmental burden and health burden overlap. The predictions would serve to narrow where more definitive, individual-level assessment is conducted, rather than to guide intervention directly. This screening role is consistent with the study's preliminary status: the framework identifies areas warranting closer examination, and confirmatory measurement would precede any specific action.

The primary contribution of this study is methodological. It presents a scale-aware framework that harmonizes predictors of differing spatial support to a common outcome scale, reduces geographic proximity between training and testing folds through household-coordinate spatial-block cross-validation while not guaranteeing complete separation of census tracts, and compares modeling paradigms under a disciplined evaluation standard. This framework is the transferable result of the work and applies to small-area environmental health prediction well beyond obesity.

The public-health implications are more limited and remain

preliminary. The findings do not identify causal determinants of obesity and are not sufficient to guide intervention. Their practical value lies in screening: the framework can indicate tracts where measured environmental burden and predicted health burden co-occur, helping direct where individual-level assessment should follow. Realizing this in practice would require tract-level confirmation, individual-level exposure data, and larger samples.

Strengths and limitations

Strengths

A major strength of this study is the use of laboratory generated soil metal measurements rather than modeled or proxy environmental exposures. Direct environmental measurements reduce exposure misclassification and provide more accurate contextual indicators (Filippelli & Laidlaw, 2010). Another strength is the explicit spatial validation strategy, which reduces leakage bias and provides a more geographically conservative evaluation (Roberts *et al.*, 2017).

Limitations

Several limitations should be acknowledged. First, the outcome is ecological. Census-tract obesity prevalence reflects aggregate patterns and does not permit individual level inference (Morgenstern, 1998). Within tract heterogeneity is compressed into a single summary measure. In addition, the outcome is a modeled small-area estimate derived from self-reported height and weight, which tends to underestimate true obesity, and its tract-level variation is generated largely from demographic and socioeconomic structure; because several influential predictors (notably racial composition and housing value) correlate with that structure, their associations with modeled obesity prevalence may partly reflect the outcome model itself rather than independent effects. Second, temporal misalignment across datasets is possible. Environmental measurements, housing records, and census estimates may not correspond to identical time periods. Temporal mismatch can attenuate or distort observed associations (Carroll *et al.*, 2006). Third, the sample size is modest relative to the dimensionality of the predictor space. While regularization and cross-validation mitigate overfitting risk, external validation is needed. In addition, the number of sampled households differed substantially across tracts, so tract-level environmental summaries derived from sparsely sampled tracts carry greater uncertainty and should be interpreted accordingly. Relatedly, the stability of the permutation-importance rankings is limited by the small number of independent tracts and by correlation among predictors; variability reported in *Supplementary Table S2* pertains to the separate tract-aggregated sensitivity model and should not be interpreted as uncertainty in the primary household-resolution ranking. Spatial dependence was addressed through spatial-block cross-validation rather than by fitting an explicit spatial model; evaluation of residual autocorrelation (e.g., Moran's I of the residuals) and explicit spatial models such as spatial lag or geographically weighted regression are identified as next steps, best pursued with more independent spatial units than the 34 available here.

Fourth, and most importantly, the unit of analysis warrants explicit caution. Because the obesity outcome is defined only at the census-tract level while the models were fit to 475 household-level observations, each tract's single outcome value is repeated across its sampled households, and the effective number of independent outcomes is 34 rather than 475. This pseudo-replication can understate uncertainty and inflate both apparent model performance and the apparent stability of predictor importance, so the reported met-

rics should be read as optimistic and provisional. Household-resolution features were used to preserve the spatial resolution of the soil measurements, but all performance and importance metrics were computed at the census-tract level (34 independent outcomes), and a fully tract-aggregated complementary sensitivity analysis is reported (Supplementary Materials). Multilevel or measurement-error models with a tract-level random effect would provide still more rigorous inference and are identified as the clear next step.

Finally, because households were recruited by convenience rather than by random sampling, the sample may not represent the underlying population, and coverage was uneven across tracts; this limits generalizability and is a target for future probability-based sampling.

Future directions

Several extensions are warranted. First, external validation in other geographic regions would test the transportability of the modeling framework. Spatial heterogeneity in environmental exposures suggests that model performance may vary across cities (Anselin, 1988). Second, longitudinal designs incorporating time-varying environmental measurements would improve temporal alignment and allow dynamic modeling of tract-level change (Lawson, 2018). Third, causal modeling approaches such as directed acyclic graphs or structural equation models could be explored when appropriate data and assumptions are available (Pearl, 2009). However, such approaches require stronger assumptions than predictive modeling and should be applied with careful consideration.

Conclusions

This proof-of-concept demonstrates the feasibility of integrating laboratory-measured household soil metals with built-environment, housing, and neighborhood data measured across heterogeneous spatial supports. Contextual demographic, housing, and spatial-gradient variables contributed more strongly to prediction than the measured soil metals, whose contribution was generally limited. Predictive performance was modest and varied across validation folds, and the identity of the strongest model differed between the household-resolution and fully tract-aggregated analyses. Because the outcome was available for only 34 census tracts and coordinate-based spatial blocking did not guarantee complete tract separation, the reported performance and importance estimates should be interpreted as provisional. The principal contribution is therefore the scale-aware integration framework and the identification of hypotheses and methodological priorities for larger, fully tract-grouped studies.

References

- Allen B, 2023. An interpretable machine learning model of cross-sectional U.S. county-level obesity prevalence using explainable artificial intelligence. *PLoS ONE* 18:e0292341.
- Anselin L, 1988. *Spatial econometrics: methods and models*. Springer Science & Business Media. (Vol. 4). Kluwer Academic Publishers.
- Banerjee S, Carlin BP, Gelfand AE, 2003. *Hierarchical modeling and analysis for spatial data*. Chapman and Hall/CRC.
- Befort CA, Nazir N, Perri MG, 2012. Prevalence of obesity among adults from rural and urban areas of the United States: findings

- from NHANES (2005–2008). *J Rural Health* 28:392–7.
- Bivand RS, Pebesma EJ, Gómez-Rubio V, 2008. *Applied spatial data analysis with R*. New York, NY: Springer New York.
- Braveman PA, Kumanyika S, Fielding J, Laveist T, Borrell LN, Manderscheid R, Troutman A, 2011. Health disparities and health equity: the issue is justice. *Am J Public Health* 101:S149–55.
- Brenning A, 2012. Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: The R package *sperrorest*. *IEEE international geoscience and remote sensing symposium*. <https://doi.org/10.1109/IGARSS.2012.6352393>
- Brook RD, Rajagopalan S, Pope CA 3rd, Brook JR, Bhatnagar A, Diez-Roux AV, Holguin F, Hong Y, Luepker RV, Mittleman MA, Peters A, Siscovick D, Smith SC Jr, Whitsel L, Kaufman JD; American Heart Association Council on Epidemiology and Prevention, Council on the Kidney in Cardiovascular Disease, and Council on Nutrition, Physical Activity and Metabolism, 2010. Particulate matter air pollution and cardiovascular disease: An update to the scientific statement from the American Heart Association. *Circulation* 121:2331–78.
- Brownson RC, Hoehner CM, Day K, Forsyth A, Sallis JF, 2009. Measuring the built environment for physical activity: State of the science. *Am J Preventive Med* 36:S99–S123.e12.
- Brulle RJ, Pellow DN, 2006. Environmental justice: human health and environmental inequalities. *Ann Rev Public Health* 27:103–24.
- Cain KL, Salmon J, Conway TL, Cerin E, Hinckson E, Mitás J, Schipperijn J, Frank LD, Anjana RM, Barnett A, Dygryn J, Islam MZ, Molina-Garcia J, Moran M, Wan Muda WAM, Oyeyemi AL, Reis R, Santos MP, Schmidt T, Schofield GM, Timperio A, Van Dyck D, Sallis JF, 2021. International Physical Activity and Built Environment Study of adolescents: IPEN Adolescent design, protocol and measures. *BMJ Open* 11:e046636.
- Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM, 2006. *Measurement error in nonlinear models: a modern perspective*. Chapman and Hall/CRC.
- Dahu BM, Khan S, Toubal IE, Alshehri M, Martinez-Villar CI, Ogundele OB, Sheets LR, Scott GJ, 2024. Geospatial modeling of deep neural visual features for predicting obesity prevalence in Missouri: quantitative study. *JMIR AI* 3:e64362.
- Diez Roux AV, 2002. A glossary for multilevel analysis. *J Epidemiol Community Health* 56:588–94.
- Diez Roux AV, Mair C, 2010. Neighborhoods and health. *Ann N Y Acad Sci* 1186:125–45.
- Drewnowski A, 2009. Obesity, diets, and social inequalities. *Nutrition Rev* 67:36–S39.
- Filippelli GM, Laidlaw MAS, 2010. The elephant in the playground: confronting lead-contaminated soils as an important source of lead burdens to urban populations. *Perspect Biol Med* 53:31–45.
- Fotheringham AS, Wong DW, 1991. The modifiable areal unit problem in multivariate statistical analysis. *Environ Planning A* 23:1025–44.
- Frank LD, Andresen MA, Schmid TL, 2004. Obesity relationships with community design, physical activity, and time spent in cars. *Am J Prevent Med* 27:87–96.
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189–1232. <https://doi.org/10.1214/aos/1013203451>
- GBD 2017 Risk Factor Collaborators, 2018. Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet* 392:1923–94.
- Gotway CA, Young LJ, 2002. Combining incompatible spatial data. *J Am Statist Assoc* 97:632–48.
- Grinsztajn L, Oyallon E, Varoquaux G, 2022. Why do tree-based models still outperform deep learning on typical tabular data? *Adv Neural Informat Proces Systems* 35:507–520.
- Hales CM, Carroll MD, Fryar CD, Ogden CL, 2020. Prevalence of obesity and severe obesity among adults: United States, 2017–2018 (NCHS Data Brief No. 360). National Center for Health Statistics. Available from: <https://www.cdc.gov/nchs/products/databriefs/db360.htm>
- Hammond RA, 2009. Complex systems modeling for obesity research. *Preventing Chronic Dis* 6:A97.
- Heindel JJ, Blumberg B, Cave M, Machtinger R, Mantovani A, Mendez MA, Nadal A, Palanza P, Panzica G, Sargis R, Vandenberg LN, Vom Saal F, 2017. Metabolism disrupting chemicals and metabolic disorders. *Reprod Toxicol* 68:3–33.
- Hruby A, Hu FB, 2015. The epidemiology of obesity: A big picture. *Pharmacoeconomics* 33:673–89.
- James G, Witten D, Hastie T, Tibshirani R, Taylor J, 2023. *An introduction to statistical learning: with applications in Python*. Springer.
- Jerrett M, Burnett RT, Ma R, Pope III CA, Krewski D, Newbold KB, Thun MJ, 2005. Spatial analysis of air pollution and mortality in Los Angeles. *Epidemiology* 16:727–36.
- Krieger N, Chen JT, Waterman PD, Rehkopf DH, Subramanian SV, 2003. Choosing area-based socioeconomic measures to monitor social inequalities in low birth weight and childhood lead poisoning: The Public Health Disparities Geocoding Project (US). *J Epidemiol Comm Health* 57:186–99.
- Kuhn M, Johnson, K, 2013. *Applied predictive modeling*, Vol. 26. New York: Springer.
- Lawson AB, 2018. *Bayesian disease mapping: hierarchical modeling in spatial epidemiology* (3rd ed.). CRC Press.
- LeCun Y, Bengio Y, Hinton G, 2015. Deep learning. *Nature* 521:436–44.
- Levin R, Brown MJ, Kashtock ME, Jacobs DE, Whelan EA, Rodman J, Schock MR, Padilla A, Sinks T, 2008. Lead exposures in U.S. Children, 2008: implications for prevention. *Environ Health Perspect* 116:1285–93.
- Lind PM, Olsen L, Lind L, 2012. Elevated circulating levels of copper and nickel are found in elderly subjects with left ventricular hypertrophy. *Ecotoxicol Environ Safety* 86:66–72.
- Longley PA, Goodchild MF, Maguire DJ, Rhind DW, 2015. *Geographic information science and systems*. John Wiley & Sons.
- Lotfata A, Georganos S, Kalogirou S, Helbich M, 2022. Ecological associations between obesity prevalence and neighborhood determinants using spatial machine learning in Chicago, Illinois, USA. *ISPRS Internat J Geo-Information* 11:550.
- Meyer H, Reudenbach C, Wöllauer S, Nauss T, 2019. Importance of spatial predictor variable selection in machine learning applications—Moving from data reproduction to spatial prediction. *Ecological Modelling* 411:108815.
- Mielke HW, Gonzales CR, Powell ET, 2019. Curtailing lead aerosols: effects of primary prevention on declining soil lead and children's blood lead in metropolitan New Orleans. *Int J Environ Res Public Health* 16:2068.
- Morgenstern H, 1995. *Ecologic studies in epidemiology: Concepts, principles, and methods*. *Ann Rev Public Health*

16:61–81.

- Morgenstern H, 1998. Ecologic studies. In K. J. Rothman & S. Greenland (Eds.), *Modern Epidemiology* (2nd ed., pp. 459–480). Lippincott Williams & Wilkins.
- Openshaw S, 1984. The modifiable areal unit problem. Concepts and techniques in modern geography. Geobooks.
- Pearl J, 2009. *Causality*. Cambridge University Press.
- Pickett KE, Pearl M, 2001. Multilevel analyses of neighbourhood socioeconomic context and health outcomes: A critical review. *J Epidemiol Comm Health* 55:111–22.
- Robert SA, Reither EN, 2004. A multilevel analysis of race, community disadvantage, and body mass index among adults in the United States. *Soc Sci Med* 59:2421–34.
- Roberts DR, Bahn V, Ciuti S, Boyce MS, Elith J, Guillera-Aroita G, Hauenstein S, Lahoz-Monfort JJ, Schröder B, Thuiller W, Warton DI, Wintle BA, Hartig F, Dormann CF, 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* 40:913–29.
- Ruzzin J, 2012. Public health concern behind the exposure to persistent organic pollutants and the risk of metabolic diseases. *BMC Public Health* 12:298.
- Sallis JF, Floyd MF, Rodriguez DA, Saelens BE, 2012. Role of built environments in physical activity, obesity, and cardiovascular disease. *Circulation* 125:729–37.
- Shwartz-Ziv R, Armon A, 2022. Tabular data: Deep learning is not all you need. *Information Fusion* 81:84–90.
- Sparks DL, Singh B, Siebecker MG, 2022. *Environmental soil chemistry*. Elsevier.
- Swinburn BA, Sacks G, Hall KD, McPherson K, Finegood DT, Moodie ML, Gortmaker SL, 2011. The global obesity pandemic: shaped by global drivers and local environments. *Lancet*, 378: 378:804–14.
- Tchounwou PB, Yedjou CG, Patlolla AK, Sutton DJ, 2012. Heavy Metal Toxicity and the Environment. *Exp Suppl* 101:133–64.
- Thayer KA, Heindel JJ, Bucher JR, Gallo MA, 2012. Role of environmental chemicals in diabetes and obesity: a National Toxicology Program workshop review. *Environ Health Perspect* 120:779–89.
- U.S. Census Bureau, 1994. *Geographic Areas Reference Manual*. U.S. Department of Commerce, Economics and Statistics Administration.
- United States Department of Agriculture, Natural Resources Conservation Service, 2021. *Soil Survey Geographic Database (SSURGO)*. Available from: <https://www.nrcs.usda.gov/resources/data-and-reports/soil-survey-geographic-database-ssurgo>
- Valavi R, Elith J, Lahoz-Monfort JJ, Guillera-Aroita G, 2019. blockCV: An R package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. *Methods Ecol Evolut* 10:225–32.
- Wakefield J, 2008. Ecologic studies revisited. *Ann Rev Public Health* 29:75–90.
- Wolpert DH, 1992. Stacked generalization. *Neural Networks* 5:241–59.
- Wood SN, 2017. *Generalized additive models: an introduction with R*. Chapman and Hall/CRC.
- Zhang X, Holt JB, Lu H, Wheaton AG, Ford ES, Greenlund KJ, Croft JB, 2014. Multilevel regression and poststratification for small-area estimation of population health outcomes: a case study of chronic obstructive pulmonary disease prevalence using the behavioral risk factor surveillance system. *Am J Epidemiol* 179:1025–33.

Online supplementary materials

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

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

Supplementary Table S3. Fixed model hyperparameters and reproducibility settings.

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.

Received: 22 June 2026; Accepted: 25 August 2026.

Contributions: Doreen Jehu-Appiah, data curation, formal analysis, software, visualization, and manuscript drafting; Emmanuel Obeng-Gyasi, conceptualization, methodology, supervision, funding acquisition, and manuscript review and editing. Both authors read and approved the final version of the manuscript and agreed to be accountable for all aspects of the work.

Conflict of interest: the authors declare that they have no competing interests.

Ethics approval and consent to participate: the study protocol was reviewed by the North Carolina A&T State University Institutional Review Board (IRB) and determined to be exempt, as all participant data were deidentified and risks to participants were judged to be minimal. The analyses reported here additionally used publicly available, aggregate secondary data (CDC PLACES; U.S. Census Bureau; USDA NRCS SSURGO).

Consent for publication: not applicable. The manuscript contains no data from any individual person.

Availability of data and materials: the obesity outcome is publicly available from the CDC PLACES project (<https://www.cdc.gov/places>); sociodemographic data are available from the U.S. Census Bureau (<https://data.census.gov>); and soil-context data are available from the USDA NRCS SSURGO database (<https://websoilsurvey.nrcs.usda.gov>). The laboratory-measured soil-metal concentrations and the assembled analytic dataset generated during this study are available from the corresponding author upon reasonable request. The analysis code is available in the following GitHub repository: GitHub repository.

Funding: the research reported in this publication was supported by the National Science Foundation under Award Number RISE-2401878. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Science Foundation.

Publisher's note: all claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article or claim that may be made by its manufacturer is not guaranteed or endorsed by the publisher.