International Journal of
Environmental Research and Public Health

MDPI

Article

Housing, Neighborhood Context, and Soil Metals as Predictors of Model-Based Small-Area Diabetes Prevalence in a Medium-Sized City: An Explainable Spatial Machine-Learning Analysis

Rifa Tasnia [1,2] and Emmanuel Obeng-Gyasi [1,2,*]

[1] Department of Built Environment, North Carolina A&T State University, Greensboro, NC 27411, USA
[2] Environmental Health and Disease Laboratory, North Carolina A&T State University, Greensboro, NC 27411, USA
* Correspondence: eobenggyasi@ncat.edu; Tel.: +1-336-285-3132

Highlights

Public health relevance—How does this work relate to a public health issue?

- Diabetes prevalence varies across neighborhoods, highlighting the importance of examining environmental exposures and neighborhood conditions associated with geographic health disparities.
- This study examines relationships among housing characteristics, built environment factors, residential soil metals, and model-based diabetes prevalence across 37 census tracts in Greensboro, North Carolina.

Public health significance—Why is this work of significance to public health?

- Housing, geographic accessibility, environmental metals, and soil characteristics all contributed to model predictions. Environmental metals received the largest summed attribution, which partly reflects the ten metal predictors relative to two housing and two geographic accessibility predictors, whereas attribution per predictor was highest for housing and geographic accessibility.
- The primary Random Forest model achieved a spatially out-of-fold R^2 of 0.275, so approximately 72.5% of between-tract variation in model-based prevalence was left unexplained, and the outcome is a modeled small-area estimate rather than a clinically ascertained diagnosis.

Public health implications—What are the key implications or messages for practitioners, policy makers and/or researchers in public health?

- Combining soil metal measurements with housing and neighborhood data can help identify communities that may benefit from additional environmental assessment and diabetes prevention efforts.
- Findings support incorporating environmental and neighborhood conditions into local public health planning while recognizing that observed predictive relationships do not establish causation.

Check for updates

Academic Editor: Yasushi Suwazono

Received: 29 September 2026
Revised: 19 October 2026
Accepted: 20 October 2026
Published: 23 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Abstract

Diabetes prevalence varies across urban neighborhoods, yet the relative predictive contributions of environmental contamination, housing conditions, and geographic accessibility remain incompletely characterized. We applied an explainable spatial machine-learning framework to census-tract variation in model-based small-area diabetes prevalence across

Greensboro, North Carolina. Environmental and neighborhood variables from 475 sampled residential locations were aggregated across 37 census tracts; tract prevalence estimates and 95% confidence limits came from the CDC PLACES 2024 release, so the findings describe relationships with modeled estimates rather than with clinically ascertained diabetes. Eighteen predictors, selected on conceptual grounds before model fitting, were evaluated using Random Forest under five-block spatial cross-validation. The primary model achieved an out-of-fold R^2 of 0.275 (RMSE 4.213; MAE 3.314 percentage points), leaving approximately 72.5% of between-tract variation unexplained; the median R^2 across 50 alternative spatial partitions was 0.224. Held-out SHAP analysis attributed 51.12% of summed absolute contribution to environmental metals, 21.78% to housing, 17.69% to geographic accessibility, and 9.40% to soil characteristics. Because the metal domain contained ten predictors against two housing and two accessibility predictors, attribution per predictor was highest for housing (40.04%) and geographic accessibility (32.52%). Individual predictor rankings were unstable across partitions and are not interpreted, and median soil cadmium and lead did not differ between High–High prevalence clusters and the remaining tracts. A 999-permutation test indicated performance beyond the permutation null ($p = 0.001$). These ecological associations are not evidence of individual-level exposure effects or causality.

Keywords: diabetes prevalence; spatial machine learning; census tract; SHAP; model-based small-area estimation; housing characteristics; built environment; soil metals; spatial cross-validation

1. Introduction

Diabetes is influenced by a combination of demographic, socioeconomic, behavioral, clinical, and environmental factors, reflecting the complex nature of metabolic dysregulation and disease development [1]. Increasing evidence suggests that diabetes risk is shaped not only by biological and behavioral factors but also by social determinants of health, including neighborhood socioeconomic conditions, housing environments, food access, and healthcare accessibility [2]. Diabetes prevalence is influenced by broader contextual factors such as housing conditions, education, income, healthcare access, and neighborhood environments, and significant associations between these factors and census-tract diabetes prevalence have been reported across the United States [3].

The growing burden of diabetes represents a major global and national public health challenge. According to the World Health Organization, approximately 830 million people worldwide were living with diabetes in 2022, compared with approximately 200 million in 1990 [4]. In the United States, the Centers for Disease Control and Prevention estimated that 38.4 million people, representing 11.6% of the population, had diabetes according to the 2024 National Diabetes Statistics Report [5]. North Carolina also experiences a substantial diabetes burden, with the 2025 North Carolina State Health Assessment reporting that approximately 10–12% of adults have diagnosed diabetes [6]. At the local level, the 2023–2024 Guilford County Community Health Assessment identifies diabetes and other chronic diseases as important public health concerns [7].

The CDC PLACES 2024 release also provides model-based estimates of diagnosed diabetes prevalence for Guilford County and its census tracts [8]. Across the 37 Greensboro-area tracts analyzed here, those estimates ranged from 3.30% to 23.90%, a spread that brackets both the statewide estimate of approximately 10% to 12% [6] and the national estimate of 11.6% [5]. These sources highlight the importance of investigating geographic variation in diabetes burden within Greensboro rather than assuming that countywide or statewide estimates adequately represent individual neighborhoods.

Neighborhood-level characteristics may affect diabetes risk through pathways involving chronic stress, healthcare access, physical activity opportunities, food environments, and environmental quality [9]. Housing environments may influence diabetes risk and self-management through affordability, neighborhood resources, environmental quality, and access to healthcare services [10]. Environmental inequities can also result in disproportionate exposure to lead (Pb) and other environmental hazards among socially disadvantaged communities, contributing to geographic health disparities [11,12].

Heavy metals, especially lead (Pb), can persist as legacy contaminants in urban soils. In areas where contaminated soil has not been effectively remediated, removed, replaced, or otherwise isolated, surface-soil contamination may persist for decades after the original anthropogenic source has been reduced or eliminated [13–18].

Residential urban soils in the United States frequently contain elevated Pb concentrations resulting from historical anthropogenic activities. Pb concentrations in urban residential soils are almost seven times greater than those found in uncontaminated background soils, according to Brown et al. [19].

Laidlaw et al. [20] and Obeng-Gyasi et al. [18] identified multiple historical sources of lead contamination in urban soils, including emissions from leaded gasoline, deterioration of lead-based paint, metal-smelting activities, battery-recycling operations, and other industrial processes. Consequently, urban soils can serve as long-term reservoirs of legacy lead contamination when contaminated soil remains in place. Although regulatory measures have substantially reduced contemporary atmospheric lead emissions, previously deposited lead may remain in urban soils for extended periods unless it is removed, replaced, isolated, or otherwise effectively remediated.

In Greensboro, residential soil lead concentrations exhibited substantial spatial variability in the previously published household soil investigation [18]. That study reported that 43.1% of 2310 residential soil samples exceeded the 100 mg/kg comparison value used for residential gardening, 3.12% exceeded the 400 mg/kg comparison value for play areas, and 0.04% exceeded the 1200 mg/kg comparison value for general residential areas. It also reported a North Carolina background soil lead median of 19.4 mg/kg. These values describe the historical screening comparisons reported in the parent study and should not be interpreted as current federal cleanup thresholds.

Additionally, the bioavailability of soil Pb, which can be impacted by soil characteristics including pH, organic matter content, soil type, and metal speciation, determines the possible health effects of soil Pb in addition to its overall concentrations [21]. Understanding the geographic distribution of environmental pollutants is crucial for identifying vulnerable populations and assessing potential environmental contributions to chronic-disease disparities because environmental exposures vary spatially [20]. Despite decreases in Pb emissions from historical sources, contaminated urban soils can remain an important potential source of human Pb exposure, as Mielke et al. [16] further showed that soil Pb concentrations are highly correlated with blood Pb levels.

1.1. Exposure Pathways and Toxicological Considerations for Lead and Cadmium

Residential soil contamination may contribute to human metal exposure through incidental ingestion of contaminated soil or household dust, inhalation of resuspended particles, and consumption of food grown in contaminated soil [22,23]. The relative importance of these pathways depends on land use, human activity, exposure duration, soil characteristics, and the bioavailability of individual metals [24].

For lead, adverse health effects include neurological, renal, cardiovascular, and developmental toxicity [23]. Experimental and epidemiological studies have also investigated possible relationships between lead exposure, oxidative stress, inflammatory responses,

and impaired glucose regulation [23,25]. Cadmium exposure occurs through contaminated food, tobacco smoke, industrial emissions, and, in certain settings, contaminated soil or dust [22,26]. Cadmium accumulates particularly in the kidneys and can induce oxidative stress, inflammatory responses, and cellular injury [26]. These mechanisms provide biological plausibility for investigating metal exposures in relation to metabolic health, but they do not establish that the measured residential soil concentrations caused diabetes in the study population.

Toxicological reference values and environmental screening levels must be distinguished from demonstrated exposure doses or clinical disease thresholds [24]. Soil screening concentrations are generally used to identify conditions requiring additional assessment; they do not indicate that all residents experience equivalent exposure or that a particular disease outcome will occur [24]. Moreover, the present study measured environmental concentrations rather than individual metal uptake, biomarkers, exposure duration, or clinical diabetes status. Accordingly, the metal predictors were interpreted as exploratory indicators of environmental conditions rather than direct measurements of biologically effective exposure.

1.2. Federal and North Carolina Regulatory Context

In the United States, the Environmental Protection Agency (EPA) evaluates contaminated residential soils through risk-based screening and site-specific remediation frameworks, including those implemented under the Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA) and the Resource Conservation and Recovery Act (RCRA). The EPA's October 2025 residential soil lead directive established a regional screening level of 200 mg/kg and a regional removal management level of 600 mg/kg for relevant federal cleanup programs, superseding its January 2024 guidance [24]. Separately, the soil-lead hazard standards issued under Section 403 of the Toxic Substances Control Act identify 400 mg/kg for bare soil in children's play areas and an average of 1200 mg/kg for bare soil in the remainder of the yard [27], and the Department of Housing and Urban Development Lead Safe Housing Rule imposes evaluation and hazard-reduction requirements on federally assisted housing, a category relevant to Greensboro's older residential stock [28]. For cadmium, the EPA publishes regional screening levels for residential soil in its generic screening tables [29]; the tract-level median cadmium concentrations measured here (31 to 168 μg/kg, equivalent to 0.031 to 0.168 mg/kg) fall orders of magnitude below those screening concentrations. These values serve different regulatory purposes and should not be interpreted as universal health-effect thresholds or as interchangeable with state-specific remediation requirements.

In North Carolina, lead hazards are addressed through state public health requirements and contaminated-site remediation programs. North Carolina General Statute § 130A–131.9C specifies soil lead remediation levels for designated residential and child-occupied settings, while the North Carolina Department of Environmental Quality provides preliminary soil remediation goals and risk-based procedures for assessing contaminated sites [30,31].

These regulatory frameworks depend on the applicable program, exposure scenario, land use, and contaminant. Consequently, the present study does not classify its soil measurements as regulatory violations or infer health risks solely from exceedance of a single screening concentration. Instead, regulatory guidance provides context for interpreting environmental contamination and identifying situations that may warrant further site-specific investigation.

Environmental pollutants have increasingly been recognized as important contributors to metabolic diseases, including diabetes, beyond traditional lifestyle and genetic

risk factors [32]. Exposure to heavy metals, particularly Pb and Cd, may increase diabetes risk by impairing glucose regulation and disrupting metabolic function, and elevated Pb exposure has been linked to type 2 diabetes in epidemiological research [1,33]. Associations between heavy metal exposures and diabetes prevalence have been reported in several settings [34,35], with experimental and epidemiologic evidence suggesting potential effects on pancreatic β-cell function and glucose homeostasis [36,37]. Heavy metals are common environmental contaminants with harmful effects on human health [38,39]. Mixture-based evidence from U.S. adults further indicates that metals and persistent environmental chemicals may exhibit nonlinear and exposure-specific relationships with cardiometabolic outcomes, which limits the value of considering environmental contaminants only in isolation [1,35].

Cadmium (Cd) is a persistent environmental contaminant linked to diabetes and several other chronic health conditions [26,33]. Oxidative stress and chronic inflammation are increasingly recognized as important biological mechanisms linking environmental metal exposures with diabetes and other cardiometabolic disorders [25,26].

Cadmium contamination in urban soils may arise from both historical and contemporary anthropogenic activities, including metal smelting and refining, battery production and disposal, industrial emissions, waste handling, and atmospheric deposition [26]. Cadmium may also enter soils through phosphate fertilizers and other soil amendments. Once introduced into the environment, cadmium can persist and may become available for plant uptake or human contact depending on soil chemistry and environmental conditions [26].

Although cadmium has been investigated in relation to metabolic and renal health outcomes, relatively limited attention has been given to whether spatial variability in residential soil cadmium, particularly within-neighborhood concentration heterogeneity, contributes predictive information about modeled small-area diabetes prevalence when housing, other metals, soil characteristics, and geographic context are considered simultaneously. This gap motivates the inclusion of both median cadmium concentration and cadmium interquartile range in the present exploratory analysis.

Previous research suggests that environmental trace-element exposure may be associated with metabolic health outcomes, although these relationships vary according to the element, exposure level, and biological context. Selenium (Se) shows a dose-dependent relationship with type 2 diabetes risk in nonexperimental studies [40], while zinc (Zn) and chromium (Cr) have been examined primarily in relation to insulin signaling and glucose regulation [34].

The exposome framework conceptualizes the cumulative influence of environmental, social, and built-environment factors across the life course and has increasingly been used to understand complex chronic disease outcomes [41–46].

Because disease risk may result from the combined influence of environmental pollutants, socioeconomic conditions, lifestyle factors, and neighborhood characteristics rather than from a single exposure source, environmental exposures should be viewed within a broader exposome perspective [32]. Previous research has shown that the prevalence of diabetes is not evenly distributed throughout geographic regions and often shows strong spatial clustering, indicating that place-based environmental and socioeconomic factors may be involved in regional differences in disease burden [10,47]. This spatial perspective is especially important because chronic disease burden often clusters geographically rather than being uniformly distributed across urban space. Prior spatial health studies have shown that areas with elevated diabetes burden are frequently shaped by overlapping social, environmental, and built-environment disadvantages [10,48–50].

Traditional statistical approaches may have limitations in capturing complex nonlinear relationships and high-dimensional interactions among environmental exposures, whereas

interpretable machine-learning methods provide greater flexibility for modeling these complex exposure–outcome associations and identifying important predictors [1,51–54].

Although previous studies have examined environmental contaminants, housing characteristics, neighborhood conditions, and diabetes separately, comparatively few studies have simultaneously incorporated environmental metals, soil characteristics, housing variables, and accessibility metrics within an explainable spatial machine-learning framework [3,9,55]. In particular, there remains limited evidence on whether household-derived environmental metal measurements, soil characteristics, housing conditions, and accessibility measures can jointly explain small-area variation in diabetes prevalence when geographic separation is explicitly considered during model validation.

Furthermore, the relative predictive contributions of these domains remain poorly characterized in an explainable machine-learning framework. To address these gaps, the present study integrates household-derived environmental measurements with housing, soil, and accessibility information at the census-tract level, evaluates predictive performance using spatial cross-validation, and applies SHAP-based interpretation to characterize the predictors and domains contributing to geographic variation in diabetes prevalence.

Therefore, the present study asked whether household-derived residential soil metal measurements add predictive information about census-tract variation in model-based diabetes prevalence beyond housing, soil, and built-environment context, and which predictor domains account for the predictive signal that is present. To address this question, we developed an explainable spatial machine-learning framework, evaluated it under geographically held-out validation, and interpreted it using SHAP attribution (see Figure 1).

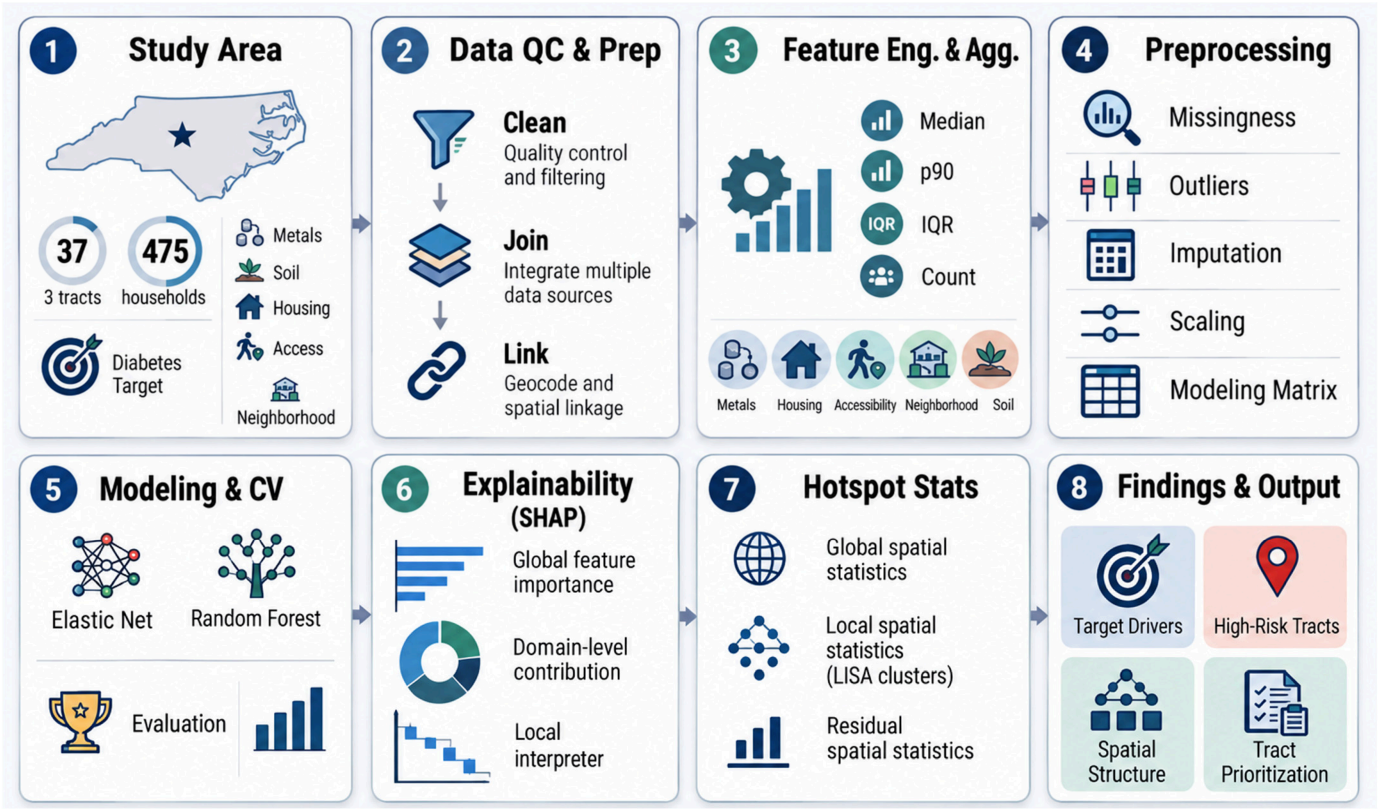


Figure 1. Overview of the study workflow integrating residential soil metals, soil characteristics, housing conditions, and geographic accessibility data with CDC PLACES modeled diabetes prevalence at the census-tract level. The framework includes data preparation, feature engineering, spatial cross-validation, Random Forest modeling, SHAP-based interpretation, spatial statistical analysis, and reporting of findings across 37 census tracts in Greensboro, North Carolina.

2. Materials and Methods

2.1. Study Area and Data Sources

This study was conducted in Greensboro, North Carolina, United States, using census tracts as the primary spatial units for descriptive, predictive, and geographic analyses. The analytical dataset was developed by linking location-level environmental and housing information to census tract geography and aggregating the relevant predictors to the tract level. The parent study [18] reported soil measurements for 462 participating residential locations, comprising 2310 individual soil samples, from the 2019 recruitment campaign. Soil from 13 further participating residences was analyzed after that publication, so the present analysis covers 475 sampled residential locations distributed across 37 census tracts.

The predictor domains included environmental metals, soil characteristics, housing conditions, and geographic accessibility measures. The number of participating households within each census tract was retained as a descriptive indicator of sampling density and was excluded from all predictive feature sets. Figure 2 illustrates the study area, the locations of sampled households, and the spatial distribution of CDC PLACES modeled diabetes prevalence across the study census tracts.

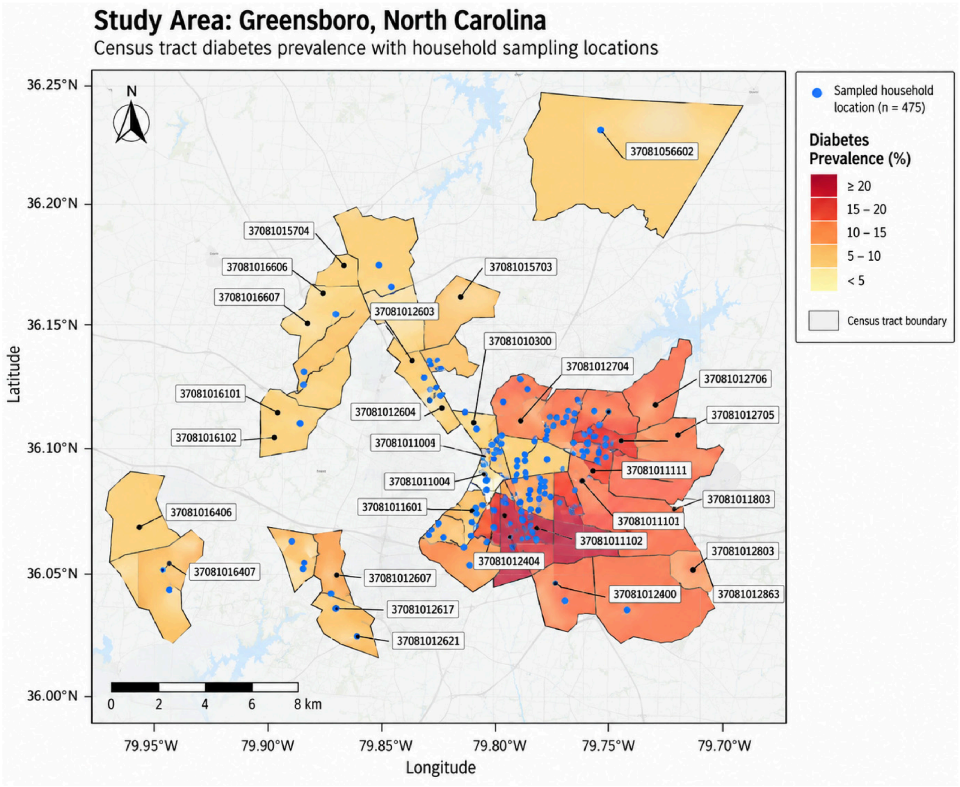


Figure 2. Study area and tract-level diabetes prevalence in Greensboro, North Carolina. Census tracts are labeled with unique tract identifiers (GEOID), and blue circles indicate the 475 sampled residential locations.

2.1.1. Household Recruitment and Soil Sampling

Household recruitment and soil sampling followed the protocol described in the parent study [18]. The original recruitment campaign was conducted between April and August 2019 in selected census blocks across Greensboro's 16 residential ZIP codes. Recruitment combined door-to-door approaches with referrals from participating residents. Census blocks containing at least five residential properties were eligible for recruitment, while apartment buildings and properties without sampleable soil were excluded. In the parent study, five bare-soil samples were collected at each household from the upper 5 cm of soil:

one at the dripline, three from the yard, and one at the street side, yielding the 2310 samples reported there [18].

Participation was voluntary, and household enrollment followed a non-probability convenience- and referral-based sampling approach intended to achieve broad geographic coverage rather than statistical representativeness. Adult residents participated in the recruitment and sampling process. The sampled households are therefore not a probability sample of Greensboro residences, and the tract-level summaries derived from them are descriptive characterizations of the sampled locations rather than population exposure estimates.

The analytic file contains one record per sampled residential location, giving 475 records: the 462 locations reported in the parent study together with 13 locations whose soil was analyzed after that publication. The 13 additional locations were sampled and analyzed using the same field and laboratory procedures, and their housing, soil-property, accessibility, and demographic characteristics were obtained from the same secondary data sources used for every other location, so they are not a distinct class of observation. Location-level metal concentrations were summarized from the soil measurements collected at each location and then aggregated to the census tract.

2.1.2. Environmental, Soil, Housing, and Geographic Data

Environmental metal measurements included lead (Pb), cadmium (Cd), zinc (Zn), selenium (Se), and copper (Cu). The documented parent-study analytical procedure used a Thermo Scientific Niton XL2 800 X-ray fluorescence (XRF) analyzer (Thermo Fisher Scientific, Waltham, MA, USA), following U.S. Environmental Protection Agency Method 6200 [56]. According to the source dataset's reporting conventions, Pb, Cu, and Zn concentrations were expressed in mg/kg, whereas Cd and Se concentrations were expressed in μg/kg.

The parent study described quality-control procedures involving a National Institute of Standards and Technology (NIST) soil reference material, with measurements evaluated against a ±10% acceptance criterion, and approximately 10% of soil samples were additionally analyzed using inductively coupled plasma optical emission spectroscopy (ICP-OES). The same field and laboratory procedures were applied to the 13 locations whose soil was analyzed after that publication.

Cadmium and selenium were treated as exploratory environmental indicators because of analytical uncertainty associated with their measurement and detection limits.

Soil characteristics included soil pH, organic matter, and soil texture, obtained from the Soil Survey Geographic Database (SSURGO) maintained by the U.S. Department of Agriculture Natural Resources Conservation Service [57]. Housing characteristics included year of house construction and taxable property value. Geographic accessibility measures included distances to the airport, major roads, and nearby streets. Census tract boundaries and centroid coordinates were obtained from the 2020 U.S. Census TIGER/Line files [58], and tract demographic and housing characteristics used for the representativeness assessment were obtained from the American Community Survey five-year estimates [59].

2.1.3. Diabetes Prevalence Outcome: CDC PLACES

Tract-level diabetes prevalence estimates were obtained from the Centers for Disease Control and Prevention (CDC) PLACES 2024 release [8,60], using the 2020 census tract geographic framework. The outcome was the official crude prevalence (%) of diagnosed diabetes among adults aged 18 years and older. Estimates were linked directly to the study census tracts using unique 11-digit census tract GEOIDs, without averaging or otherwise aggregating household-level diabetes values.

The final analytical dataset contained 37 unique census tracts, each with one official modeled diabetes prevalence estimate and its corresponding lower and upper 95% confidence limits. All 37 tracts had complete outcome and confidence-limit values, and no duplicate GEOIDs or internally inconsistent confidence intervals were identified. Tract-level diabetes prevalence ranged from 3.30% to 23.90%, with a mean of 13.76%.

CDC PLACES estimates are generated using a model-based small-area estimation framework incorporating Behavioral Risk Factor Surveillance System (BRFSS) data and census-related demographic information [60]. The outcome therefore represents modeled population prevalence rather than diabetes diagnoses directly observed among the participating households. Directly observed clinical or local health-department diabetes prevalence data were not available at the same census tract resolution for this analysis.

During revision, an error in the construction of the outcome variable used in the originally submitted analysis was identified and corrected. In the original submission, the tract-level outcome had been derived from household-linked values rather than by joining the official PLACES tract estimate one-to-one to each tract GEOID, as described above. All models, performance metrics, SHAP attributions, spatial statistics, tables, and figures reported in this manuscript were regenerated using the one-to-one tract-level outcome. The quantitative results reported here therefore differ substantively from those in the original submission, and no numerical result from the original submission is carried forward.

Because the PLACES estimation framework incorporates demographic and socioeconomic information, certain housing and geographic predictors may act as indirect proxies for characteristics associated with the estimation process. These relationships were considered when interpreting model performance and feature attribution, and additional housing-excluded and predictor-overlap sensitivity analyses were conducted as described in Sections 2.2–2.5 and Supplementary Table S3.

2.2. Data Quality Control and Spatial Preparation

2.2.1. Data Quality Control

The analysis used a tract-level dataset derived from the 475 sampled residential locations described in Section 2.1, covering 37 census tracts in Greensboro, North Carolina. The location-level file included location identifiers, geographic coordinates, soil metal measurements (Pb, Cd, Zn, Se, and Cu), soil characteristics, housing attributes, and geographic accessibility measures.

Non-informative variables, administrative identifiers, and shapefile attributes without substantive predictive meaning were excluded from the modeling matrices. Potentially redundant environmental, soil, housing, and geographic measures were examined before defining the predictor specifications.

The number of participating households within each census tract was retained separately as a sampling-density measure and excluded from all predictor matrices.

2.2.2. Feature Engineering

Household-level environmental, soil, housing, and geographic variables were summarized at the census-tract level to develop the predictive dataset. Candidate summaries included medians, 90th percentiles, and interquartile ranges (IQRs), representing central tendency, upper-tail values, and within-tract variability.

The broader predictor specification contained 43 variables. A reduced specification containing 18 predictors was defined for this revision and is reported as the primary model. Its composition was selected on conceptual grounds, before the revised models were fitted and without reference to the outcome values, in order to limit redundancy while retaining interpretable variables across the four predictor domains. This specification

was defined during the present revision and did not precede the original study. Nine accessibility variables in the original submission were logarithmic transformations of nine other accessibility variables; because Random Forest splits are invariant to monotone transformations, these duplicated variables carried no additional information and were removed, as were soil texture fields with near-zero variance.

The 18 predictors consisted of

- Environmental metals (10): Median and IQR concentrations of Pb, Cd, Zn, Se, and Cu.
- Soil characteristics (4): Median soil pH, organic matter, clay content, and sand content.
- Housing conditions (2): Median taxable home value and median year of house construction.
- Geographic accessibility (2): Median distance to the airport and median distance to major roads.

An additional housing-excluded specification containing 16 predictors was defined as a co-primary model. This specification retained the 10 metal, four soil, and two geographic accessibility predictors while excluding the two housing variables.

The 43-predictor specification was retained as a sensitivity analysis to evaluate the influence of predictor-set complexity. Sampled-household count was not included in any of these predictor matrices.

2.2.3. Spatial Distribution of Representative Predictors

Representative environmental, soil, and housing predictors were mapped to illustrate their geographic distributions across the study census tracts.

The descriptive maps (Figures 3 and 4) provide geographic context for the spatial variation in selected predictors. These maps were intended to characterize the sampled study area rather than establish causal associations or identify independently validated exposure hotspots.

2.2.4. Assessment of Sampling Representativeness

The geographic and socioeconomic characteristics of sampled census tracts were compared with those of unsampled census tracts intersecting the Greensboro municipal boundary. Census tract boundaries were obtained from the 2020 U.S. Census TIGER/Line geographic files, and demographic and housing characteristics were obtained from the 2024 American Community Survey (ACS) five-year estimates.

The municipal-intersection dataset contained 92 census tracts, comprising 36 sampled and 56 unsampled tracts. One unsampled tract (GEOID 37081980100) lacked valid observations for all six ACS characteristics selected for the final comparison and was excluded from the descriptive analysis. Accordingly, the final comparison included 36 sampled and 55 unsampled tracts (n = 91). One of the 37 tracts included in the predictive analysis did not intersect the municipal boundary used for this assessment and was therefore excluded from the sampled-versus-unsampled municipal comparison.

The six characteristics were median household income, median owner-occupied home value, renter-occupied housing percentage, White-alone population percentage, Black-alone population percentage, and Hispanic population percentage. For each characteristic, distributions were summarized using medians and interquartile ranges (IQRs). Standardized mean differences (SMDs) were calculated to describe differences between sampled and unsampled census tracts.

These comparisons were descriptive and were intended to identify potential socioeconomic and demographic differences in geographic sampling coverage. Because household recruitment was non-probability-based, the analysis was not interpreted as establishing statistical representativeness or equivalence between sampled and unsampled areas. Detailed results are reported in Supplementary Table S10.

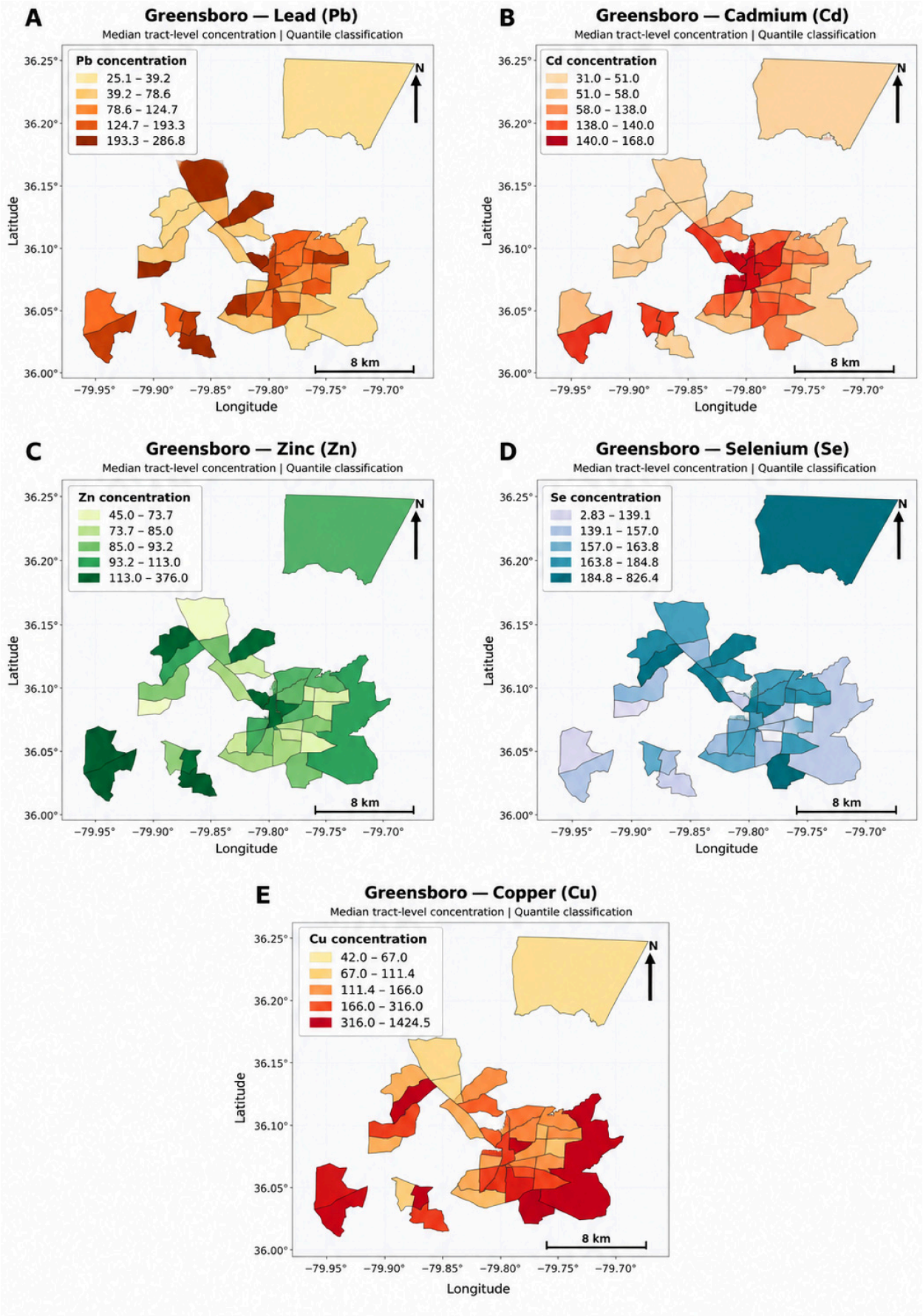


Figure 3. Spatial distribution of tract-level median metal concentrations across Greensboro census tracts. Panels show (**A**) Pb, (**B**) Cd, (**C**) Zn, (**D**) Se, and (**E**) Cu. All maps are displayed using quantile classification to facilitate comparison of tract-level spatial variation.

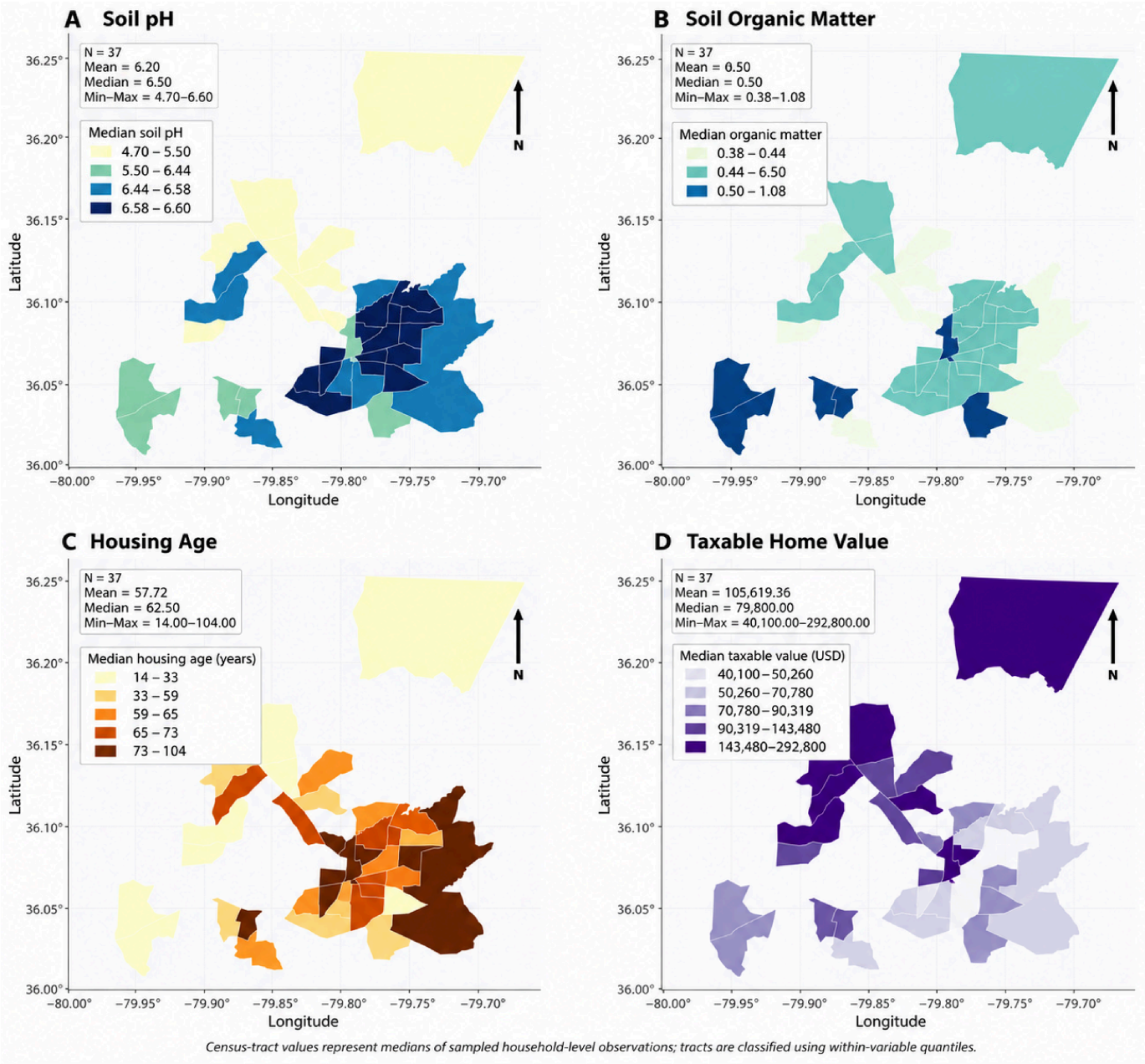


Figure 4. Spatial distribution of selected soil and housing characteristics across Greensboro census tracts. Panels show (**A**) median soil pH, (**B**) median soil organic matter, (**C**) median housing age, and (**D**) median taxable home value. Predictor values are classified using tract-level quantiles for descriptive visualization.

2.3. Predictive Modeling Framework

Given the relatively small number of census tracts compared with the number of candidate predictors, the models were treated as exploratory predictive tools rather than causal or etiologic models. With 37 spatial units, the analysis supports exploratory, spatially validated prediction and domain-level screening, but it does not support predictor-specific causal effect estimation or conventional multivariable inference.

2.3.1. Data Preprocessing

Before model fitting, tract-level predictors were evaluated for completeness, variability, and conceptual redundancy. Geographic identifiers, administrative attributes, and variables without informative variation were excluded from the predictor matrix. The number of

sampled households per tract was retained only as a descriptive sampling-density variable and was not included as a predictor.

Where required, missing predictor values were handled through median imputation fitted exclusively on training observations within each spatial cross-validation fold. Standardization was applied to linear and penalized regression models, whereas Random Forest and other tree-based models were fitted without predictor standardization.

Predictor reduction was guided by conceptual redundancy and interpretability rather than outcome-guided univariate significance screening. Zillow-derived housing-value measures and redundant geographic distance transformations were excluded from the primary reduced specification, while median taxable home value was retained as the representative property-value measure.

Three principal predictor specifications were considered: the primary 18-predictor model, the housing-excluded 16-predictor co-primary model, and the 43-predictor sensitivity model. Additional reduced specifications were examined in sensitivity analyses. Correlated predictors were retained when they represented substantively different environmental or geographic characteristics.

2.3.2. Prediction Model

Random Forest [61] was the primary learner for this analysis because it accommodates nonlinear relationships and interactions among correlated predictors without requiring a prespecified functional form, and because it remains comparatively stable with the small number of spatial units available. The originally submitted manuscript also reported a comparison across six candidate algorithms, comprising ordinary least squares, ridge regression, elastic net, Random Forest, gradient boosting, and histogram gradient boosting. That comparison was based on the superseded outcome definition described in Section 2.1.3 and was not re-run for all six candidates under the revised outcome definition. It has therefore been removed, and no claim is made in this manuscript that Random Forest outperformed the other candidate algorithms.

2.3.3. Spatial Cross-Validation

Predictive performance was evaluated using five-block spatial cross-validation, which holds out geographically contiguous blocks of observations in order to reduce the optimistic bias that random cross-validation produces for spatially autocorrelated data [62]. The 37 study census tracts were assigned to five geographically structured validation blocks, containing eight, eight, seven, seven, and seven tracts, respectively.

In each fold, one spatial block was held out for validation while the remaining four blocks were used for training. The procedure was repeated until every census tract had one out-of-fold prediction.

The same primary spatial partition was used for the principal predictor-set comparisons and related sensitivity analyses. Preprocessing operations requiring fitted parameters were performed exclusively using training-fold observations and then applied to the held-out validation block. Out-of-fold predictions were pooled across the five folds to calculate predictive performance.

To evaluate sensitivity to spatial partition selection, 50 alternative five-block spatial partitions were generated by varying partition orientation and retaining unique tract-to-block assignments. The primary 18-predictor and 43-predictor Random Forest specifications were evaluated on identical partitions, allowing paired performance comparisons.

For each partition, pooled out-of-fold R^2, root-mean-square error (RMSE), and mean absolute error (MAE) were calculated. Results were summarized using medians and 5–95th percentile ranges. Paired RMSE differences between predictor specifications were also

evaluated. These empirical percentile ranges describe variation across spatial partition configurations and were not interpreted as confidence intervals for predictive generalization.

2.3.4. Performance Evaluation and Model Configuration

Predictive performance was assessed using the coefficient of determination (R^2), RMSE, and MAE. All primary performance metrics were calculated from pooled out-of-fold predictions rather than training-set predictions. RMSE and MAE were expressed in diabetes-prevalence percentage points.

The primary Random Forest model used 1500 trees (n_estimators = 1500), square-root feature sampling at each split (max_features = "sqrt"), a minimum of two observations per terminal leaf (min_samples_leaf = 2), unrestricted maximum tree depth, and a fixed random seed of 42. Parallel computation was enabled using n_jobs = −1. The same Random Forest configuration was retained across the primary, co-primary, and predictor-set sensitivity analyses.

Data processing, modeling, validation, and visualization were conducted in Python (version 3.12.3) using pandas (version 2.2.2) [63], NumPy (version 1.26.4) [64], SciPy (version 1.13.1) [65], and scikit-learn (version 1.4.2) [66]. Spatial processing and mapping were performed in ArcGIS Pro (version 3.6.0; Esri, Redlands, CA, USA) [67]. Hyperparameters were prespecified and held constant rather than selected through optimization on the held-out spatial validation folds.

2.4. Explainable Machine Learning

Model interpretation was performed using SHapley Additive exPlanations (SHAP), a game-theoretic method for attributing model predictions to individual predictors [68]. The primary SHAP analysis focused on the reduced 18-predictor Random Forest specification. To maintain consistency with spatial cross-validation, SHAP values for each census tract were calculated using the Random Forest model trained without the tract's corresponding validation block.

These out-of-fold SHAP values were used to summarize predictor importance and domain-level attribution across all 37 census tracts.

Global predictor importance was calculated as the mean absolute SHAP value for each predictor. Each predictor's relative importance was expressed as a percentage of the sum of mean absolute SHAP values across all 18 predictors. Predictors were grouped into four predefined domains: environmental metals (10 predictors), soil characteristics (4 predictors), housing conditions (2 predictors), and geographic accessibility (2 predictors).

Primary domain-level importance was calculated by summing the mean absolute SHAP values within each domain and expressing the resulting sum as a percentage of the total across all predictors. This summed measure is the primary domain-level summary reported in this manuscript. Because the summed measure is influenced by the number of predictors within a domain, and the four domains contained ten, four, two, and two predictors respectively, a complementary feature-count-normalized analysis was also performed. For that analysis, mean absolute SHAP values were averaged within each domain, and the four domain averages were normalized to sum to 100%.

The summed measure describes how much of the total attribution each domain received, whereas the feature-count-normalized measure describes average attribution per predictor within each domain. The two measures answer different descriptive questions and produce different domain orderings, so both are reported together throughout. Neither is interpreted as an independent or causal effect.

SHAP attribution was also examined for the housing-excluded 16-predictor co-primary model to assess changes in predictor rankings and domain-level contributions after removal of the housing variables.

Because correlated predictors may share or redistribute SHAP attribution, individual metal rankings and domain-level contributions were not interpreted as causal effects, independent associations, or direct measures of exposure-related disease burden. The SHAP values were checked against their corresponding held-out Random Forest predictions using the additive reconstruction property.

2.5. Spatial Statistical Analysis and Sensitivity Analyses

2.5.1. Spatial Autocorrelation of Prediction Residuals

Spatial autocorrelation in the primary Random Forest model's out-of-fold residuals was evaluated using global Moran's I [69]. Residuals were defined as observed minus predicted CDC PLACES modeled diabetes prevalence.

Queen-contiguity spatial weights were used as the primary neighborhood definition. Alternative k-nearest-neighbor spatial weights with k = 2, 3, 4, and 5 were examined to assess sensitivity to spatial-neighborhood construction.

Permutation-based significance was evaluated using 999 permutations. The Queen-contiguity result was designated as the primary spatial diagnostic, while k-nearest-neighbor results were retained as sensitivity analyses. Findings were interpreted with recognition that residual spatial autocorrelation may depend on the neighborhood definition.

2.5.2. Permutation Test of Predictive Performance

An outcome-label permutation test was conducted to evaluate whether the observed spatial cross-validated predictive performance exceeded that expected under random reassignment of diabetes prevalence values across census tracts.

The tract-level diabetes prevalence outcome was permuted 999 times, and the Random Forest model was evaluated using the primary five-block spatial cross-validation configuration for each permutation.

The empirical upper-tail *p*-value was calculated as

p = (1 + number of permuted R^2 values greater than or equal to the observed R^2)/(999 + 1). The permutation test was interpreted as a diagnostic of predictive signal under the specified permutation scheme rather than evidence of a causal relationship.

2.5.3. Uncertainty Propagation Using CDC PLACES Confidence Limits

Uncertainty associated with the CDC PLACES model-based diabetes prevalence estimates was examined using 500 outcome draws informed by the official tract-specific lower and upper 95% confidence limits.

For each draw, predictive performance was evaluated using the established modeling and spatial cross-validation framework. The resulting R^2, RMSE, and MAE distributions were summarized using medians and 2.5–97.5th percentile ranges. These ranges describe sensitivity to uncertainty in the modeled outcome and were not interpreted as confidence intervals for predictive generalization.

2.5.4. Assessment of Potential Overlap with CDC PLACES Model Inputs

Potential conceptual overlap between the predictors and the CDC PLACES small-area estimation framework was evaluated through predictor classification and housing-excluded sensitivity analyses.

The primary 18-predictor specification included two housing variables: median taxable home value and median year of house construction. The housing-excluded 16-predictor co-

primary specification retained the 10 environmental-metal predictors, four soil predictors, and two geographic accessibility predictors.

An additional metals-and-soil-only specification containing 14 predictors was evaluated as a sensitivity analysis. The 18 primary predictors were classified according to their potential relationships with documented PLACES estimation inputs. No exact documented direct input overlap was identified, while four housing and geographic predictors were classified as potential indirect socioeconomic, built-environment, or geographic proxies (Supplementary Table S3).

These analyses were intended to assess predictive robustness and potential interpretive overlap, not to establish independence from the modeled outcome or infer causal effects.

2.5.5. Additional Predictor-Set Sensitivity Analyses

Additional sensitivity analyses compared the primary 18-predictor Random Forest model with the housing-excluded 16-predictor model, the 14-predictor metals-and-soil model, a 17-predictor model excluding airport distance, and the 43-predictor specification.

All comparisons used the CDC PLACES outcome, the same primary five-block spatial cross-validation configuration, and consistent Random Forest hyperparameters.

These comparisons assessed the sensitivity of predictive performance to predictor-set composition rather than independent or causal effects of individual predictors.

2.5.6. Sampling-Density Sensitivity Analyses

Sampling density was characterized by the number of participating households within each census tract. The 37 study tracts were divided descriptively into those containing fewer than five sampled households and those containing at least five sampled households.

A household-count-weighted Random Forest specification was compared with the unweighted primary model to assess sensitivity to assigning greater training weight to tracts containing more participating households.

Associations between sampled-household counts and absolute out-of-fold prediction errors were assessed using Spearman rank correlation. Differences in absolute prediction errors between sampling-density groups were examined using the Mann–Whitney U test.

These analyses were exploratory diagnostics and were not interpreted as evidence that the household sample was geographically representative.

2.5.7. Airport-Distance and Geographic Trend-Surface Sensitivity Analysis

To assess whether airport distance represented broad geographic position rather than airport-specific exposure, four Random Forest specifications were evaluated. These included the primary 18-predictor model, a 17-predictor model excluding airport distance, and both specifications augmented with a quadratic geographic trend surface.

The trend surface comprised five terms derived from projected census-tract centroid coordinates: easting, northing, squared easting, squared northing, and the easting-by-northing interaction. Centroid coordinates were projected to a metric coordinate system and standardized using training-fold observations only.

Pearson and Spearman correlations were calculated between airport distance and tract-centroid easting and northing. All four specifications used the same 37 census tracts, five spatial cross-validation blocks, and prespecified Random Forest hyperparameters. Predictive performance was compared using pooled out-of-fold R^2, RMSE, and MAE.

Detailed airport-distance ablation and geographic trend-surface results are presented in Supplementary Table S9.

3. Results

3.1. Characteristics of the Tract-Level Analytic Dataset

The final analytical dataset comprised 37 census tracts linked to 475 sampled residential locations. Diabetes prevalence was obtained directly from the CDC PLACES 2024 census tract dataset using 2020 census tract GEOIDs. Each tract was assigned one official model-based diabetes prevalence estimate. Across the 37 tracts, modeled diabetes prevalence ranged from 3.30% to 23.90%, with a mean of 13.76%, a median of 13.60%, and an interquartile range (IQR) of 8.30 percentage points (Table 1).

Table 1. Descriptive statistics for the tract-level analytic dataset across the 37 Greensboro-area census tracts.

Variable	Domain	Mean	SD	Median	IQR	Min	Max
Diabetes prevalence (%)	Outcome	13.76	5.02	13.60	8.30	3.30	23.90
Number of sampled households	Sampling density	12.84	13.78	8.00	14.00	1.00	53.00
Taxable home value, median ($)	Housing	105,619.36	66,856.97	79,800.00	73,250.00	40,100.00	292,800.00
Year house built, median	Housing	1962.28	23.64	1957.50	37.00	1916.00	2006.00
Lead (Pb), median (mg/kg)	Metals	116.85	77.20	94.00	118.70	25.10	286.80
Cadmium (Cd), median (µg/kg)	Metals	96.46	47.46	111.00	89.00	31.00	168.00
Cadmium (Cd), IQR (µg/kg)	Metals	43.72	39.89	29.50	84.00	0.00	106.50
Copper (Cu), median (mg/kg)	Metals	227.47	281.59	132.50	192.50	42.00	1424.50
Selenium (Se), median (µg/kg)	Metals	182.33	125.65	159.00	42.00	2.83	826.41
Distance to airport, median (m)	Accessibility	12,429.99	4511.09	13,213.27	5993.10	3221.16	23,628.38
Distance to major roads, median (m)	Accessibility	644.34	946.10	402.33	393.71	102.98	5682.23
Soil pH, median	Soil	6.20	0.58	6.50	0.48	4.70	6.60
Soil organic matter, median	Soil	0.50	0.11	0.50	0.05	0.38	1.08

Abbreviations: SD = standard deviation; IQR = interquartile range. Descriptive statistics were calculated across 37 census tracts using the CDC PLACES 2024 diabetes prevalence outcome. The number of sampled households is reported only as a sampling-density descriptor and was not included as a model predictor. Distance variables are reported in meters rather than log-transformed units.

Household sampling density varied across the study area. Thirteen census tracts contained fewer than five participating households, accounting for 26 household records, whereas 24 tracts contained at least five households, accounting for 449 records. Within-tract interquartile ranges may therefore be unstable in the sparsely sampled tracts, which limits interpretation of the within-tract variability predictors in those tracts. The number of sampled households was retained exclusively as a descriptive sampling-density measure and was excluded from all predictive feature sets.

Environmental, soil, housing, and geographic accessibility characteristics also varied across the study tracts. For example, tract-level median Pb concentrations ranged from 25.10 to 286.80 mg/kg, while median taxable home values ranged from $40,100 to $292,800 (Table 1).

The primary predictive specification contained 18 predictors: 10 environmental-metal summaries, four soil characteristics, two housing variables, and two geographic accessibility measures. A housing-excluded 16-predictor specification was evaluated as a co-primary model, while the 43-predictor specification was retained for sensitivity analysis. Given the limited number of spatial observations, findings were interpreted as exploratory predictive relationships rather than independent or causal effects.

3.2. Spatial Distribution of Environmental Metals

Figure 3A–E presents the geographic distributions of tract-level median Pb, Cd, Zn, Se, and Cu concentrations, respectively. The mapped concentrations varied across the 37 study census tracts, indicating descriptive spatial heterogeneity in the sampled soil measurements.

These maps summarize the geographic distribution of tract-level metal concentrations and do not establish statistically significant exposure clusters or independent associations with modeled diabetes prevalence.

3.3. Spatial Distribution of Soil and Housing Characteristics

Figure 4 presents the spatial distributions of selected soil and housing characteristics. Median soil pH and soil organic matter are displayed in Panels A and B, respectively, while median housing age and median taxable home value are presented in Panels C and D.

Tract-level median soil pH ranged from 4.70 to 6.60, and median soil organic matter ranged from 0.38 to 1.08. Median taxable home value ranged from $40,100 to $292,800 across the 37 study tracts (Table 1). These geographic distributions provide descriptive context for the subsequent predictive analyses but do not establish independent associations with diabetes prevalence.

3.4. Predictive Performance of the Primary and Alternative Random Forest Specifications

The primary reduced 18-predictor Random Forest model was evaluated using five-block spatial cross-validation across 37 census tracts. Based on pooled out-of-fold predictions, the model achieved an R^2 of 0.275, an RMSE of 4.213 percentage points, and an MAE of 3.314 percentage points (Figure 5; Supplementary Table S2).

The full 43-predictor sensitivity specification achieved an R^2 of 0.295, an RMSE of 4.155 percentage points, and a MAE of 3.324 percentage points, and therefore yielded a slightly higher R^2 and lower RMSE than the primary model on this fixed partition. The reduced 18-predictor model was nonetheless retained as the primary specification because its composition was fixed on conceptual grounds before the revised models were fitted, because it is more parsimonious and interpretable, and because it achieved a lower RMSE than the full specification in 34 of 50 alternative spatial partitions (Section 3.8).

The housing-excluded 16-predictor co-primary model achieved an R^2 of 0.202, an RMSE of 4.420 percentage points, and an MAE of 3.515 percentage points. The metals-and-soil-only 14-predictor sensitivity specification achieved an R^2 of 0.207, an RMSE of 4.406 percentage points, and an MAE of 3.503 percentage points.

These results demonstrate sensitivity to predictor-set composition. However, differences in predictive performance do not establish independent or causal effects of environmental or socioeconomic variables.

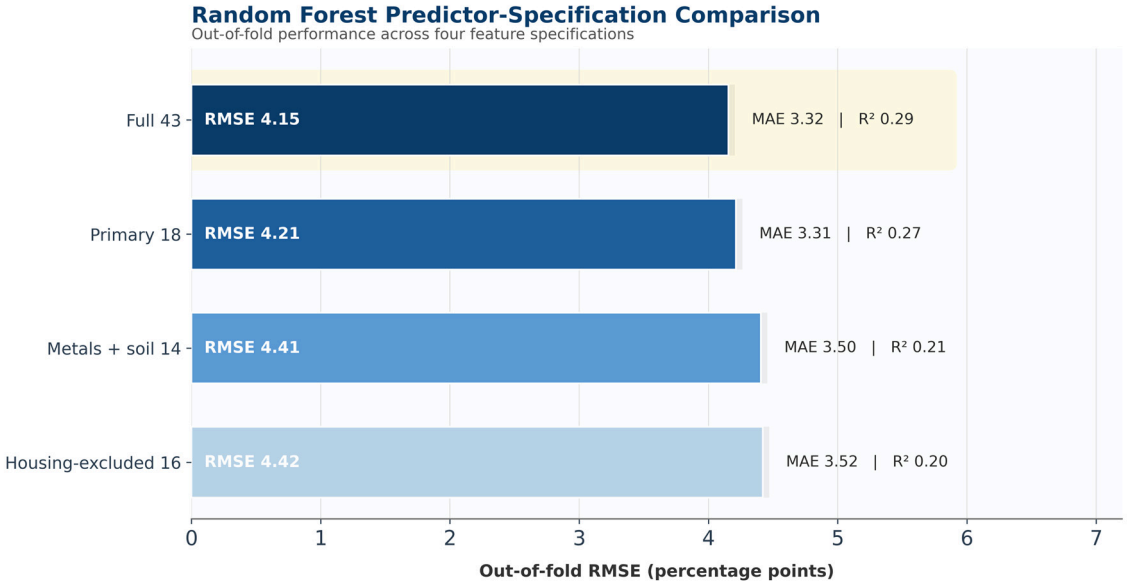


Figure 5. Comparison of spatially cross-validated Random Forest performance across four predictor specifications. Horizontal bars represent pooled out-of-fold root mean squared error (RMSE) for the full 43-predictor sensitivity model, primary reduced 18-predictor model, metals-and-soil-only 14-predictor sensitivity model, and housing-excluded 16-predictor co-primary model. Corresponding mean absolute error (MAE) and coefficient of determination (R^2) are displayed alongside each bar. All models were evaluated using the CDC PLACES 2024 modeled diabetes prevalence estimates across 37 census tracts under identical five-block spatial cross-validation. RMSE and MAE are expressed in diabetes-prevalence percentage points. Lower RMSE indicates better predictive performance. The 18-predictor model was retained as the primary specification while the other models were evaluated as alternative predictor-set specifications. The light-yellow background highlights the full 43-predictor model for visual comparison, while the different shades of blue distinguish the four predictor specifications.

3.5. Diagnostic Evaluation of the Final Random Forest Model

Beyond the aggregate predictive performance metrics, the out-of-fold (OOF) predictions from the primary 18-predictor Random Forest model were examined using the five-block spatial cross-validation framework (Figure 6). Across the 37 census tracts, the model achieved an overall R^2 of 0.275, RMSE of 4.213 percentage points, and MAE of 3.314 percentage points.

Figure 6A illustrates the relationship between CDC PLACES modeled diabetes prevalence and out-of-fold predictions. The fitted trend was flatter than the 1:1 agreement line, indicating compression of the predicted prevalence range relative to the observed modeled prevalence range. The model overpredicted prevalence in some lower-prevalence tracts and underpredicted prevalence in some higher-prevalence tracts.

Residuals were defined as observed minus predicted modeled prevalence (Figure 6B). The mean residual was −0.09 percentage points, indicating a small average signed prediction error. The Spearman correlation between residuals and predicted prevalence was weak and not statistically significant ($\rho = 0.18$, $p = 0.292$). Nevertheless, this result does not exclude other forms of systematic prediction error.

The distribution of absolute prediction errors is shown in Figure 6C. The mean absolute error was 3.314 percentage points, while the median absolute error was 2.37 percentage points, indicating that several relatively large prediction errors increased the average.

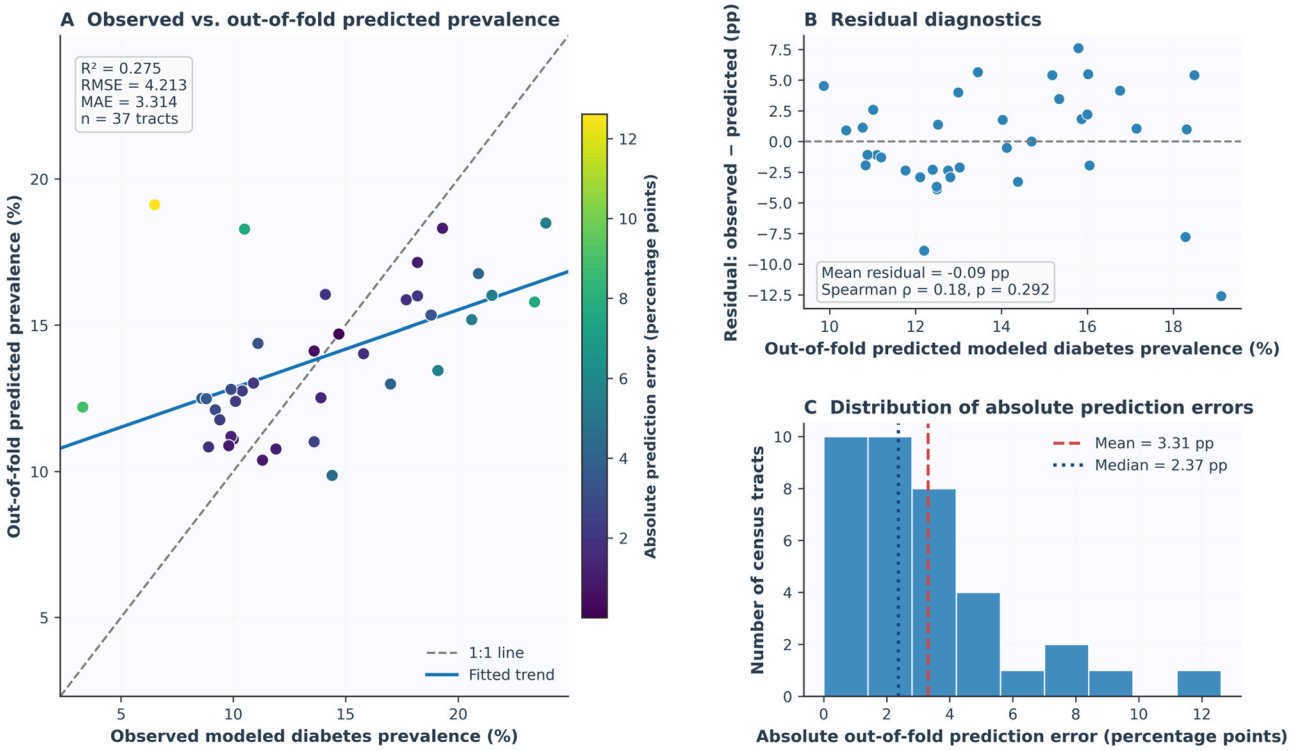


Figure 6. Diagnostic evaluation of the primary 18-predictor Random Forest model under five-block spatial cross-validation using CDC PLACES 2024 diabetes prevalence estimates. (**A**) Agreement between observed modeled diabetes prevalence and out-of-fold (OOF) predictions across 37 census tracts. Point colors represent absolute prediction errors in percentage points (pp). The dashed diagonal indicates perfect agreement (1:1 line), while the solid blue line represents the fitted linear trend. The pooled OOF performance was R^2 = 0.275, RMSE = 4.213 pp, and MAE = 3.314 pp. (**B**) Residuals, calculated as observed minus OOF-predicted prevalence, plotted against predicted prevalence. The mean residual was −0.09 pp, and the Spearman correlation between residuals and predicted values was ρ = 0.18 (p = 0.292). Positive residuals indicate underprediction, whereas negative residuals indicate overprediction. (**C**) Distribution of absolute OOF prediction errors across the 37 census tracts. Vertical reference lines indicate the mean absolute error (3.31 pp) and median absolute error (2.37 pp). All diagnostic statistics and plots were calculated exclusively from held-out predictions generated using the primary 18-predictor Random Forest specification. In Panel (**A**), each dot represents a census tract, with color indicating the absolute prediction error; the dashed diagonal line represents perfect agreement (1:1), and the solid blue line represents the fitted linear trend. In Panel (**B**), each dot represents a tract-level residual, and the horizontal dashed line indicates zero residual. In Panel (**C**), the red dashed vertical line indicates the mean absolute prediction error, while the blue dotted vertical line indicates the median absolute prediction error.

3.6. Spatial Distribution of Modeled and Out-of-Fold Predicted Diabetes Prevalence

Figure 7 presents the spatial distributions of CDC PLACES modeled diabetes prevalence, out-of-fold predictions, prediction residuals, and absolute prediction errors across the 37 study census tracts.

Modeled diabetes prevalence ranged from 3.30% to 23.90%, with a mean of 13.76%, median of 13.60%, and IQR of 8.30 percentage points (Figure 7A). Higher modeled prevalence occurred in several southeastern tracts, whereas lower prevalence occurred in selected central and western tracts.

Out-of-fold predicted prevalence ranged from 9.86% to 19.11%, with a mean of 13.85%, median of 13.02%, and IQR of 3.76 percentage points (Figure 7B). The narrower predicted range indicates that the model did not fully reproduce the most extreme tract-level prevalence estimates.

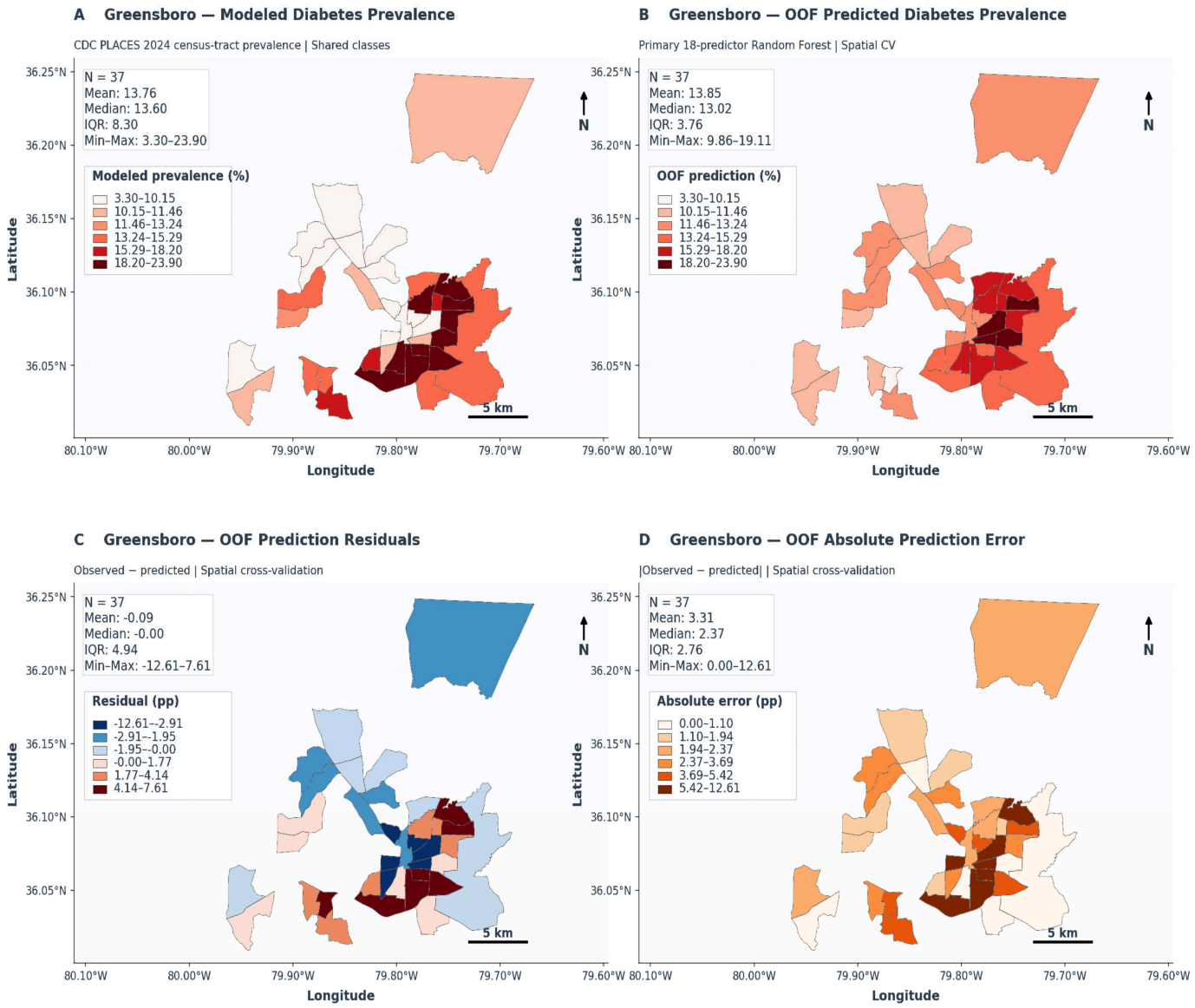


Figure 7. Spatial distribution of CDC PLACES modeled diabetes prevalence and out-of-fold predictions from the primary 18-predictor Random Forest model across 37 Greensboro-area census tracts. (**A**) CDC PLACES 2024 census-tract modeled diabetes prevalence (%). (**B**) Spatially out-of-fold (OOF) predicted diabetes prevalence (%) from the primary 18-predictor Random Forest model evaluated using balanced five-block spatial cross-validation. Panels (**A**,**B**) use identical prevalence classification boundaries to facilitate direct comparison. (**C**) OOF prediction residuals, calculated as observed minus predicted prevalence (percentage points); negative values indicate overprediction, and positive values indicate underprediction. (**D**) Absolute OOF prediction errors, calculated as the absolute difference between observed and predicted prevalence (percentage points). Six color classes are displayed in each panel; prevalence classes are shared between Panels (**A**,**B**), while residuals and absolute errors are classified separately. Summary statistics are reported within each panel. The primary model achieved a pooled OOF R^2 of 0.275, RMSE of 4.213 percentage points, and MAE of 3.314 percentage points. Geographic patterns are descriptive and should not be interpreted as evidence of statistically significant spatial clusters or causal effects.

Prediction residuals ranged from −12.61 to 7.61 percentage points, with a mean of −0.09 percentage points and an IQR of 4.94 percentage points (Figure 7C). Negative residuals indicate overprediction, whereas positive residuals indicate underprediction.

Absolute prediction errors ranged from approximately 0.00 to 12.61 percentage points, with a mean of 3.31 percentage points, median of 2.37 percentage points, and IQR of

2.76 percentage points (Figure 7D). Larger errors occurred in selected tracts rather than uniformly across the study area.

The primary model achieved a pooled out-of-fold R^2 of 0.275, indicating that approximately 27.5% of the variation in the modeled tract-level diabetes prevalence outcome was accounted for by the held-out predictions under this evaluation. Substantial predictive variation remained unexplained.

3.7. Explainable Machine-Learning Analysis

3.7.1. Global SHAP Feature Importance

Global out-of-fold (OOF) SHAP analysis was conducted to identify the relative contributions of individual predictors to the spatially held-out predictions of the primary 18-predictor Random Forest model.

Median taxable home value had the highest global importance, with a mean absolute SHAP value of 0.786 percentage points, representing 18.18% of total absolute attribution. This was followed by cadmium IQR (0.683 percentage points; 15.78%), median distance to the airport (0.591 percentage points; 13.65%), and lead IQR (0.378 percentage points; 8.74%) (Figure 8).

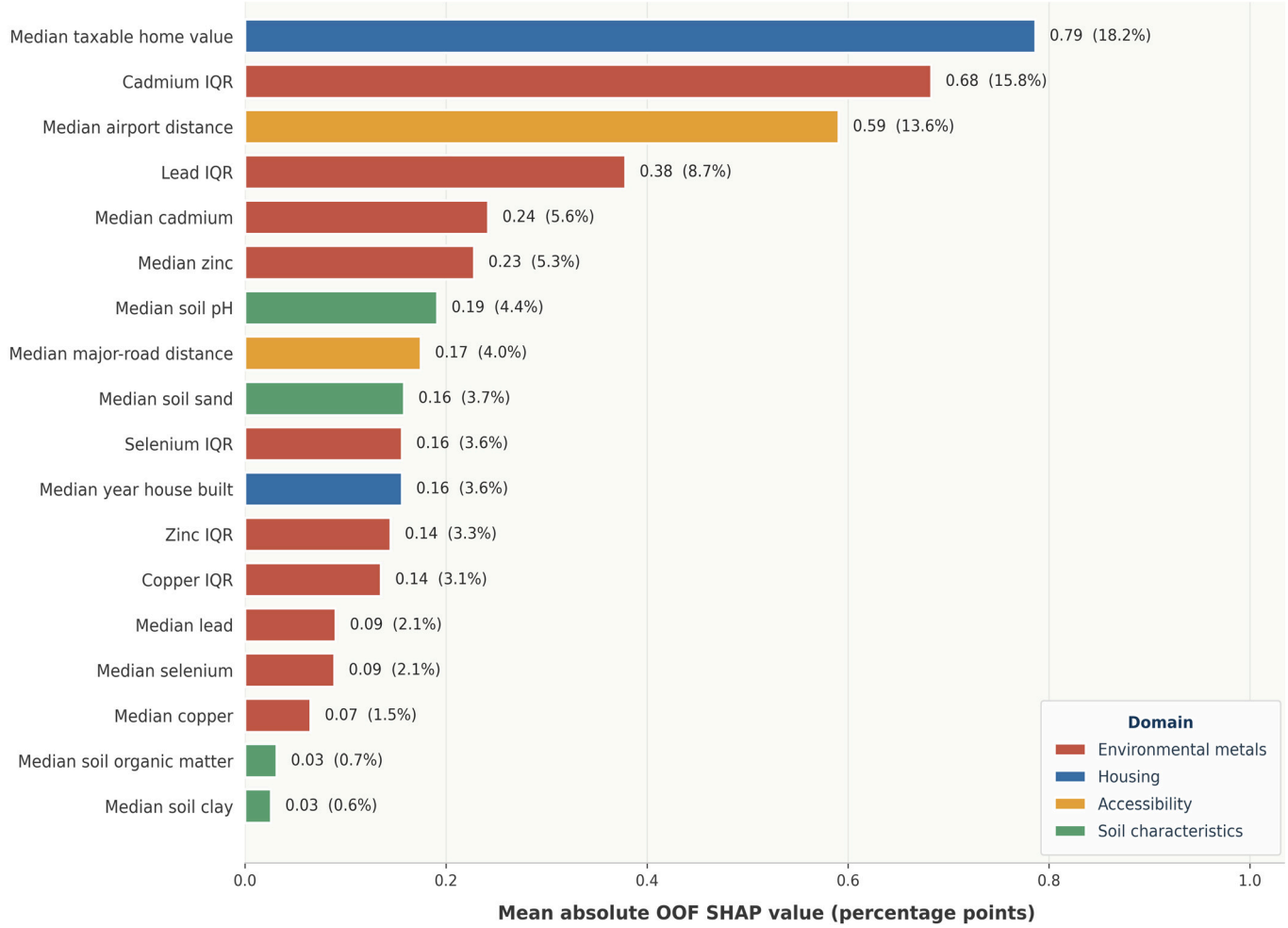


Figure 8. Global out-of-fold SHAP feature importance for the primary 18-predictor Random Forest model. Horizontal bars represent mean absolute SHAP values calculated from spatially held-out predictions across 37 census tracts. Features are ranked in descending order of global importance. Numerical labels indicate mean absolute SHAP values in percentage points and the corresponding percentage contribution to total absolute SHAP attribution. Bar colors distinguish environmental metals,

housing characteristics, geographic accessibility, and soil characteristics. Larger absolute SHAP values indicate greater contributions to model predictions, not independent or causal effects on diabetes prevalence.

Other notable predictors included median cadmium concentration (5.60%), median zinc concentration (5.27%), median soil pH (4.42%), and median distance to major roads (4.04%).

These results indicate that housing, environmental-metal, and geographic accessibility variables contributed to the model's predictions. SHAP importance quantifies predictive attribution rather than independent associations or causal effects, and attribution may be redistributed among correlated predictors. Individual predictor ranks were not stable across spatial partitions or across predictor specifications, so they are reported descriptively and are not used to support claims about specific metals (Supplementary Section S4).

3.7.2. SHAP Summary, Domain Contribution and Local Explanation

Out-of-fold SHAP values were further examined using a global summary plot, domain-level aggregation, and a tract-specific local explanation (Figure 9).

The global SHAP summary (Figure 9A) showed heterogeneity in predictor contributions across the 37 census tracts. Lower median taxable home values generally corresponded to positive SHAP contributions, whereas higher home values tended to contribute negatively to predicted modeled diabetes prevalence. Higher cadmium variability and greater airport distance generally contributed positively to predictions, although magnitudes varied among tracts. These patterns describe fitted model behavior and should not be interpreted as independent exposure–response relationships.

Summed absolute SHAP contributions indicated that environmental-metal predictors accounted for 51.12% of total attribution, followed by housing characteristics (21.78%), geographic accessibility (17.69%), and soil characteristics (9.40%) (Figure 9B; Supplementary Table S4).

Because the predictor domains contained different numbers of features, a complementary feature-count-normalized comparison was also conducted. Under this approach, housing accounted for 40.04% of relative attribution, followed by geographic accessibility (32.52%), environmental metals (18.79%), and soil characteristics (8.64%) (Supplementary Table S4). Household sampling density was excluded from both the predictor matrix and the SHAP analysis.

A local out-of-fold SHAP waterfall explanation was constructed for GEOID 37081011000, the tract with the highest held-out predicted prevalence (Figure 9C). Its observed CDC PLACES modeled prevalence was 6.50%, whereas the Random Forest prediction was 19.11%, representing an overprediction of approximately 12.61 percentage points.

Relative to the fold-specific baseline prediction of 13.82%, cadmium IQR (+1.74 percentage points) and median taxable home value (+1.36 percentage points) made the largest positive contributions. Additional positive contributions included lead IQR (+0.70), median airport distance (+0.39), copper IQR (+0.32), and median year of house construction (+0.31). Selenium IQR (−0.18) and median cadmium concentration (−0.16) contributed negatively.

These local SHAP values explain how the fitted model generated its held-out prediction, not why the tract's modeled diabetes prevalence differed from that prediction.

Taken together, the summed measure assigns the largest share of total attribution to environmental-metal predictors, whereas the feature-count-normalized measure assigns the largest average per-predictor attribution to housing and geographic accessibility. The difference between the two follows directly from the number of predictors in each domain. Neither measure establishes independent or causal effects.

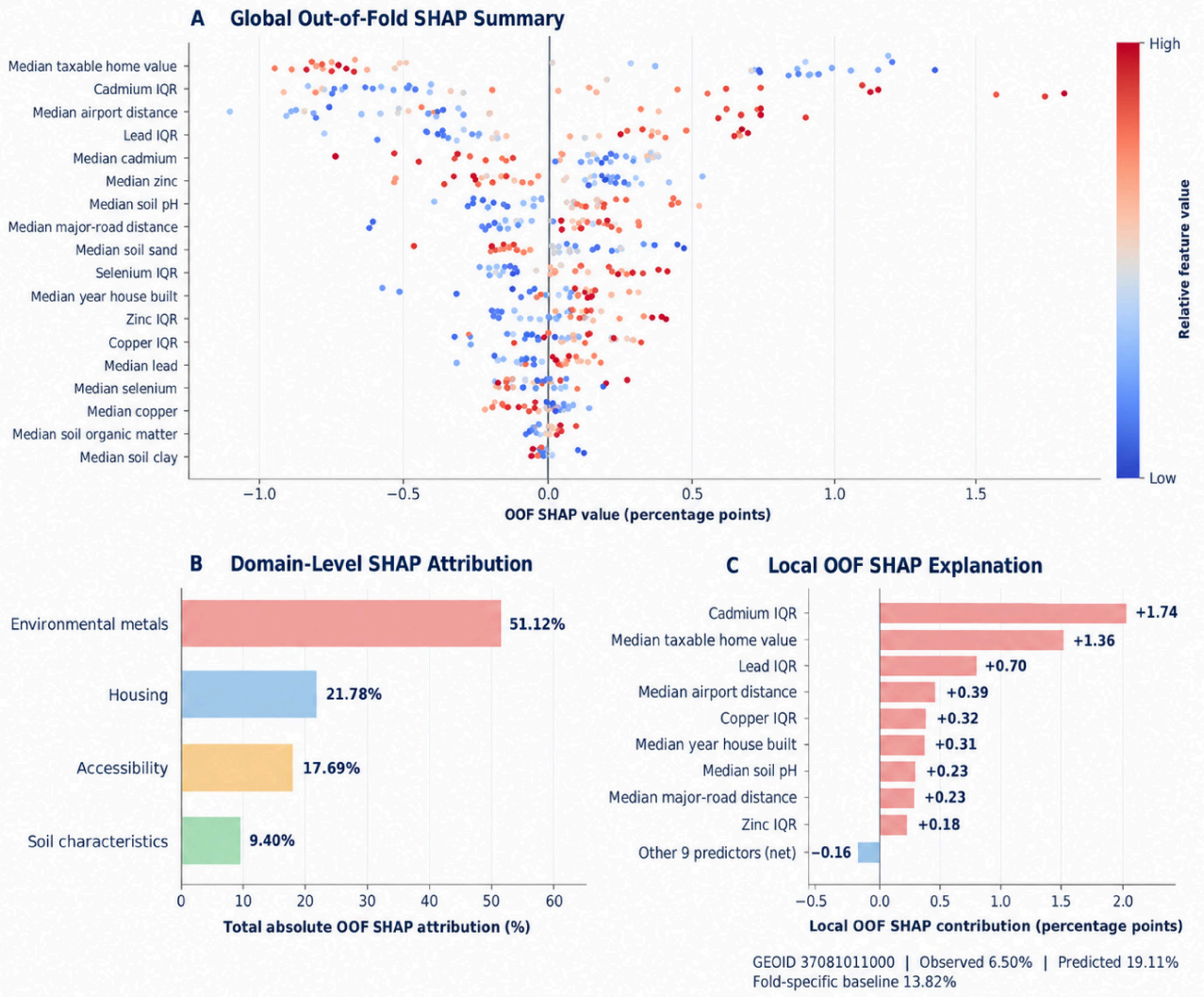


Figure 9. SHAP-based interpretation of the primary 18-predictor Random Forest model for census-tract-level modeled diabetes prevalence. (**A**) Global out-of-fold (OOF) SHAP beeswarm plot showing the magnitude and direction of predictor contributions across 37 census tracts. Each point represents one tract, and point color indicates the relative predictor value, from low (blue) to high (red). Positive and negative SHAP values indicate contributions that increase or decrease the corresponding held-out prediction relative to its fold-specific baseline. (**B**) Domain-level contributions calculated by summing mean absolute OOF SHAP values across predictors within each domain. Environmental metals accounted for 51.12% of total absolute attribution, followed by housing (21.78%), geographic accessibility (17.69%), and soil characteristics (9.40%). Domain totals are not adjusted for differences in predictor counts. (**C**) Local OOF SHAP waterfall explanation for the census tract with the highest predicted prevalence (GEOID 37081011000; observed prevalence 6.50%; predicted prevalence 19.11%). The fold-specific baseline prediction was 13.82%. Positive and negative bars indicate predictor contributions to the held-out prediction. SHAP values quantify model attribution and should not be interpreted as independent or causal effects.

3.8. Robustness and Sensitivity Analyses

Predictive performance was evaluated under alternative predictor specifications using the CDC PLACES outcome and five-block spatial cross-validation (Figure 10A,B).

The primary 18-predictor model achieved an R^2 of 0.275, RMSE of 4.213 percentage points, and MAE of 3.314 percentage points. The full 43-predictor model yielded an R^2 of 0.295, RMSE of 4.155 percentage points, and MAE of 3.324 percentage points.

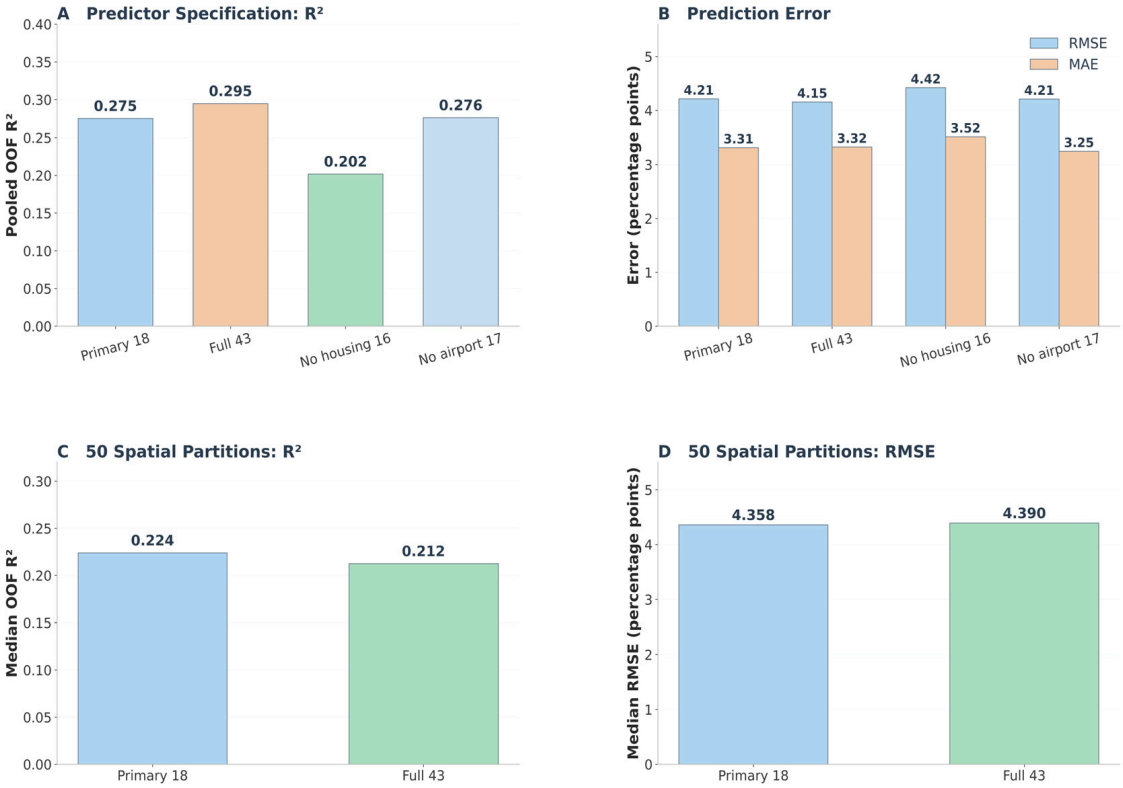


Figure 10. Robustness and sensitivity of predictive performance across predictor specifications and spatial partitions. (**A**) Overall, out-of-fold (OOF) coefficient of determination (R^2) for four Random Forest specifications: the primary 18-predictor model, full 43-predictor model, co-primary 16-predictor model excluding housing variables, and 17-predictor model excluding median airport distance. (**B**) Corresponding root-mean-square error (RMSE) and mean absolute error (MAE), expressed in percentage points of CDC PLACES 2024 modeled tract-level diabetes prevalence. (**C**) Median OOF R^2 across 50 alternative spatial partitions for the primary and full models. (**D**) Corresponding median RMSE across the 50 spatial partitions. All models were evaluated using spatially structured cross-validation across 37 census tracts. Panels (**C**,**D**) display medians rather than the full distribution of performance estimates; percentile ranges and paired comparisons are reported separately.

The housing-excluded 16-predictor co-primary model achieved an R^2 of 0.202, RMSE of 4.420 percentage points, and MAE of 3.515 percentage points. Excluding median distance to the airport produced nearly unchanged performance (R^2 = 0.276, RMSE = 4.210 percentage points, and MAE = 3.245 percentage points). These findings indicate sensitivity to housing-predictor inclusion but limited sensitivity to removal of airport distance.

Across 50 matched alternative spatial partitions, the primary 18-predictor model achieved a median pooled out-of-fold R^2 of 0.224 (5–95th percentile: 0.118–0.323), median RMSE of 4.358 percentage points (4.072–4.646), and median MAE of 3.509 percentage points (3.157–3.775) (Supplementary Table S5).

The full 43-predictor model achieved a median R^2 of 0.212 (0.086–0.322), median RMSE of 4.390 percentage points (4.072–4.730), and median MAE of 3.592 percentage points (3.249–3.859).

In paired comparisons, the primary 18-predictor model had a lower RMSE in 34 of the 50 partitions, whereas the full 43-predictor model had a lower RMSE in 16 partitions. The median paired RMSE difference (primary minus full) was −0.016 percentage points, with a 5–95th percentile range of −0.118 to 0.049 percentage points.

These results suggest a modest performance advantage for the reduced specification across the evaluated alternative spatial partitions, although predictive accuracy remained

sensitive to partition composition. The percentile ranges represent empirical variation across partition configurations rather than confidence intervals for generalization performance.

An outcome-uncertainty sensitivity analysis included 500 realizations informed by the official CDC PLACES tract-level confidence limits. Across these realizations, the median R^2 was 0.259 (2.5–97.5th percentile: 0.182–0.339), the median RMSE was 4.315 percentage points (3.982–4.668), and the median MAE was 3.371 percentage points (3.058–3.662) (Supplementary Table S6). These ranges quantify sensitivity to simulated uncertainty in the modeled prevalence estimates and should not be interpreted as confidence intervals for predictive performance.

An outcome-permutation analysis was conducted using 999 permutations. The observed pooled out-of-fold R^2 of 0.275 exceeded all 999 permuted values, yielding an empirical upper-tail p-value of 0.001. The mean permutation-based R^2 was −0.144, and the 95th percentile of the null distribution was 0.063 (Supplementary Table S6). These findings indicate predictive signal under the specified permutation scheme without establishing causal relationships.

Residual spatial autocorrelation was evaluated using global Moran's I under alternative spatial-weight definitions [69]. Under Queen contiguity, Moran's I was 0.164 (one-sided permutation $p = 0.066$; two-sided $p = 0.121$). Under k-nearest-neighbor weights, Moran's I was 0.518 (k = 2, $p = 0.001$), 0.372 (k = 3, $p = 0.002$), 0.249 (k = 4, $p = 0.006$), and 0.213 (k = 5, $p = 0.011$) (Supplementary Table S7). These results indicate that residual spatial structure depended on the neighborhood definition and was not fully eliminated under all evaluated specifications.

Additional geographic sensitivity analyses indicated that median airport distance was strongly correlated with tract-centroid easting (Pearson r = 0.927; Spearman ρ = 0.977) but only weakly correlated with northing (r = 0.112). Removing airport distance left performance essentially unchanged ($R^2 = 0.276$, RMSE = 4.210 percentage points) relative to the primary model ($R^2 = 0.275$, RMSE = 4.213 percentage points). Adding a quadratic geographic trend surface to the primary model gave $R^2 = 0.258$ and RMSE = 4.261 percentage points, while the specification excluding airport distance and including the trend surface gave $R^2 = 0.268$ and RMSE = 4.233 percentage points. These findings indicate that airport distance primarily captures broad geographic position rather than a demonstrated airport-specific exposure effect, and that neither trend-surface specification improved predictive performance (Supplementary Table S9).

Sampling-density diagnostics were evaluated for the same out-of-fold predictions. A household-count-weighted Random Forest specification achieved an R^2 of 0.233, an RMSE of 4.334 percentage points, and an MAE of 3.417 percentage points, compared with 0.275, 4.213, and 3.314 for the unweighted primary model. Absolute prediction error was positively correlated with the number of sampled households per tract (Spearman $\rho = 0.380$, $p = 0.020$), and the 13 tracts with fewer than five sampled households had lower error (MAE 2.243, RMSE 2.561 percentage points) than the 24 tracts with at least five households (MAE 3.893, RMSE 4.879 percentage points), although the two-group comparison was not statistically significant (Mann–Whitney $p = 0.131$). Prediction error was therefore not concentrated in the sparsely sampled tracts. These diagnostics do not establish that tract-level environmental summaries based on few households are reliable, and interpretation of within-tract variability predictors in sparsely sampled tracts remains limited (Supplementary Table S8).

3.9. Spatial Clustering of Diabetes Prevalence

Local spatial autocorrelation in the CDC PLACES 2024 modeled diabetes prevalence estimates was examined across the 37 study census tracts using Local Moran's I [70] with Queen contiguity spatial weights and 999 permutations (Figure 11).

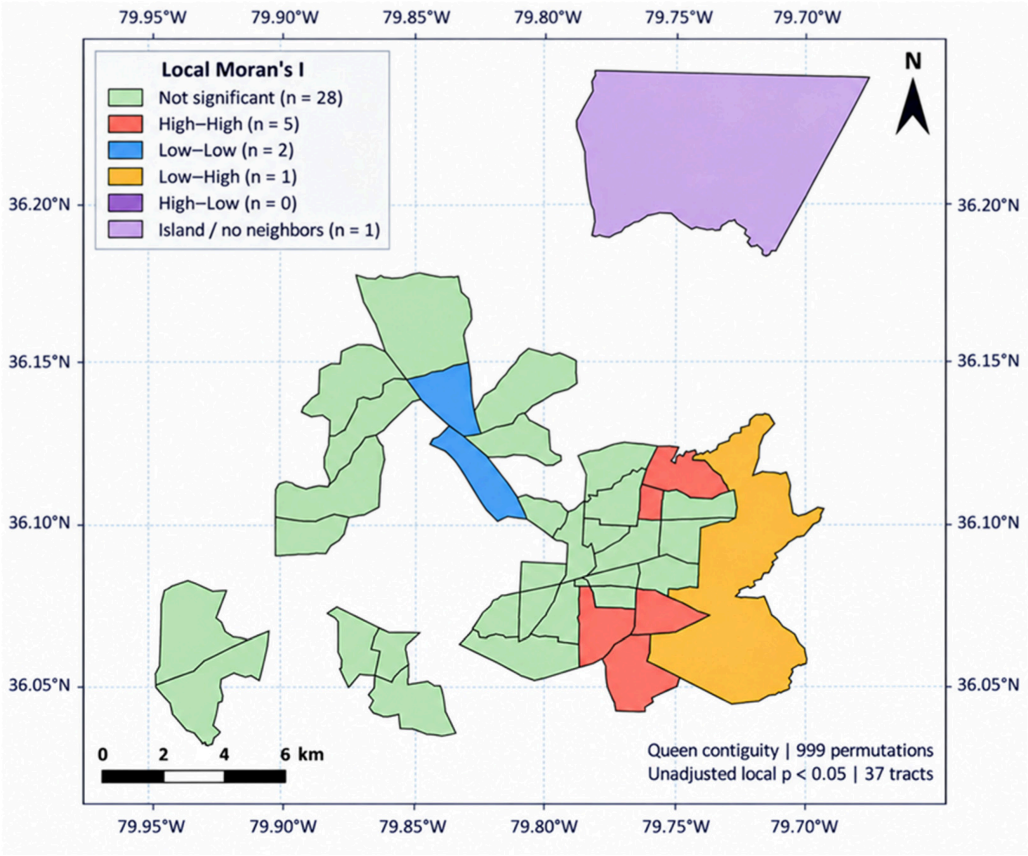


Figure 11. Local spatial clustering of CDC PLACES 2024 modeled diabetes prevalence across 37 Greensboro-area census tracts. Local Indicators of Spatial Association (LISA) were calculated using Local Moran's I with Queen contiguity spatial weights and 999 permutations. Census tracts are classified as High–High clusters (n = 5), Low–Low clusters (n = 2), Low–High spatial outliers (n = 1), High–Low spatial outliers (n = 0), nonsignificant tracts (n = 28), or an island tract without Queen-contiguous neighbors (n = 1). Cluster classifications are based on unadjusted local permutation *p*-values at the 0.05 significance threshold. Latitude and longitude coordinates, north arrow, and a scale bar are provided for geographic reference. The map displays exploratory local spatial associations in the modeled diabetes prevalence outcome; it does not represent Random Forest prediction residuals or establish causal relationships.

At an unadjusted local significance threshold of $p < 0.05$, five tracts were classified as High–High clusters, two as Low–Low clusters, and one as a Low–High spatial outlier. No High–Low spatial outliers were identified.

Among the remaining tracts, 28 showed no statistically significant local spatial association, while one tract was classified as an island because it had no neighboring tract under the Queen contiguity definition.

The High–High clusters were concentrated primarily in central-eastern and southeastern portions of the study area, whereas the two Low–Low tracts occurred farther west.

Because local significance was evaluated using unadjusted permutation *p*-values, these classifications should be interpreted as exploratory geographic patterns rather than definitive hotspots after multiple-testing correction. The clustering analysis describes the spatial distribution of the CDC PLACES modeled outcome, not Random Forest residuals. Resid-

ual spatial autocorrelation was evaluated separately using global Moran's I, as reported in Section 3.8.

3.10. Representativeness of Sampled Census Tracts

Among the 92 census tracts intersecting the Greensboro municipal boundary, 36 contained sampled households and 56 were unsampled. One unsampled tract (GEOID 37081980100) lacked valid values for all six selected American Community Survey characteristics and was excluded from the descriptive comparison. Consequently, the final analysis included 36 sampled and 55 unsampled tracts (n = 91; Supplementary Table S10).

Sampled tracts had lower median household income than unsampled tracts ($50,794 versus $73,621) and lower median owner-occupied home value ($203,450 versus $293,800). Conversely, the median percentage of renter-occupied housing was higher in sampled tracts (56.1% versus 36.3%).

Differences were also observed in racial and ethnic composition. Sampled tracts had a lower median White-alone population percentage (30.1% versus 51.4%) and a higher median Black-alone population percentage (42.3% versus 29.7%). Median Hispanic population percentages were 9.3% and 5.7%, respectively.

These descriptive differences suggest potential limitations in the socioeconomic and demographic representativeness of the sampled locations. Generalization to unsampled neighborhoods should therefore be approached cautiously.

3.11. Exploratory Comparison of Soil Cadmium and Lead Across LISA-Defined Clusters

An exploratory comparison of tract-level median soil cadmium (Cd) and lead (Pb) concentrations was conducted between the five High–High LISA clusters and the remaining 32 census tracts (Figure 12; Supplementary Table S11).

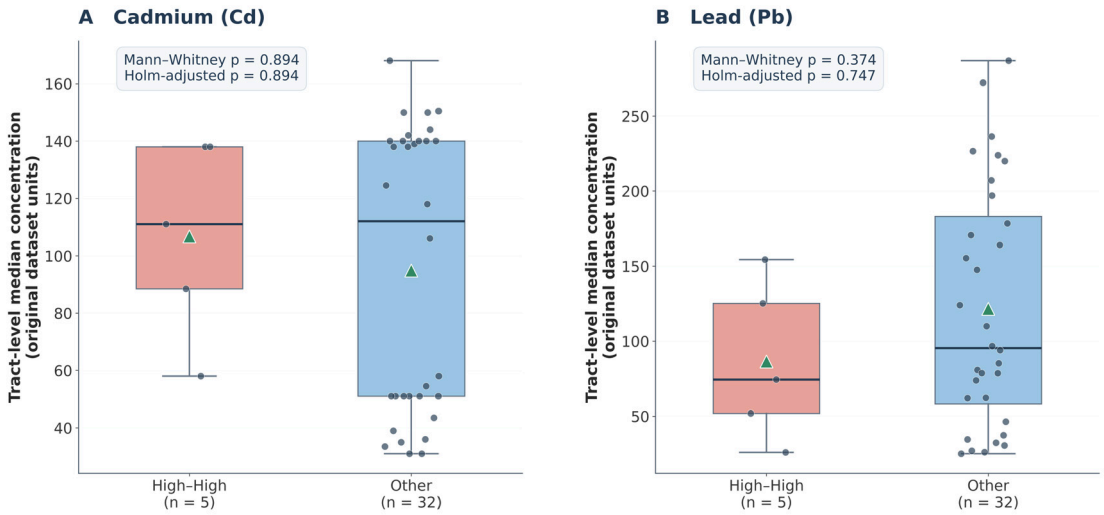


Figure 12. Soil cadmium and lead concentrations in LISA-defined diabetes clusters. Boxplots compare tract-level median concentrations of (**A**) cadmium (Cd) and (**B**) lead (Pb) between five High–High census tracts identified using Local Moran's I on CDC PLACES 2024 modeled diabetes prevalence and the remaining 32 census tracts. Boxes indicate interquartile ranges, horizontal lines represent medians, whiskers indicate non-outlier ranges, gray points represent individual census tracts, and green triangles denote group means. High–High tracts are shown in light coral and other tracts in light blue. Two-sided Mann–Whitney U tests with Holm adjustment did not identify statistically significant differences for Cd (p = 0.894; adjusted p = 0.894) or Pb (p = 0.374; adjusted p = 0.747). Concentrations are reported in the original dataset units. These comparisons are descriptive and do not establish causal associations.

Neither Cd (Mann–Whitney U = 83.5, p = 0.894; Holm-adjusted p = 0.894) nor Pb (U = 59.0, p = 0.374; Holm-adjusted p = 0.747) showed a statistically significant difference between groups.

These results indicate that the exploratory spatial clustering of modeled diabetes prevalence was not accompanied by statistically detectable marginal differences in these two soil-metal summaries. Given the small number of High–High tracts, the results should not be interpreted as evidence of no association or as supporting causal conclusions.

4. Discussion

4.1. Principal Findings

This study developed an interpretable spatial machine-learning framework to characterize geographic variation in CDC PLACES 2024 modeled diabetes prevalence across 37 census tracts in Greensboro, North Carolina. Using a primary set of 18 environmental-metal, soil, housing, and geographic-accessibility predictors, the Random Forest model achieved a pooled out-of-fold R^2 of 0.275, RMSE of 4.213 percentage points, and MAE of 3.314 percentage points under five-block spatial cross-validation. These results indicate modest predictive performance, with substantial variation in modeled diabetes prevalence remaining unexplained.

Predictive performance varied across predictor specifications. The full 43-predictor model achieved an R^2 of 0.295, whereas the housing-excluded 16-predictor co-primary model achieved an R^2 of 0.202. Removing airport distance from the primary predictor set produced nearly unchanged performance (R^2 = 0.276). Across 50 alternative spatial partitions, the primary 18-predictor model achieved a median R^2 of 0.224 and median RMSE of 4.358 percentage points, compared with 0.212 and 4.390 percentage points for the full 43-predictor specification. The primary model had a lower RMSE in 34 of the 50 matched partitions, suggesting a modest advantage for the more parsimonious specification across the evaluated partition configurations.

Out-of-fold SHAP analysis identified predictive contributions from environmental-metal, housing, geographic-accessibility, and soil variables. Environmental metals received the largest summed attribution (51.12%), whereas housing and geographic accessibility received the largest attribution per predictor (40.04% and 32.52% after feature-count normalization). That difference follows from the ten metal predictors relative to two housing and two geographic accessibility predictors. Individual predictor ranks were not stable across spatial partitions or across predictor specifications. The results therefore support cautious interpretation of domain-level predictive patterns rather than causal attribution to individual environmental exposures.

Spatial analyses identified geographic heterogeneity in modeled diabetes prevalence, including five High–High clusters under the Queen-contiguity Local Moran's I analysis using unadjusted permutation significance. Residual spatial autocorrelation depended on the spatial-weight specification, indicating that conclusions about remaining spatial structure were sensitive to the definition of neighboring census tracts.

Collectively, these findings demonstrate the potential value of combining spatially validated prediction, explainable machine learning, and geographic statistical analysis for exploratory neighborhood-level health research. However, the limited number of census tracts, modest predictive accuracy, and model-based nature of the outcome constrain the strength and generalizability of the conclusions.

4.2. Interpretation of Housing, Environmental Metals, and Geographic Predictors

Median taxable home value was the most influential individual predictor in the primary Random Forest model, accounting for 18.18% of total absolute out-of-fold SHAP

attribution. This finding suggests that housing characteristics contain useful information for predicting spatial variation in CDC PLACES modeled diabetes prevalence. At the census-tract level, taxable home value may reflect broader neighborhood characteristics, including socioeconomic conditions, patterns of residential development, access to resources, and historical investment. However, the observed attribution should not be interpreted as evidence that housing value independently influences diabetes prevalence.

Environmental-metal predictors collectively accounted for 51.12% of summed absolute SHAP attribution. Cadmium interquartile range and lead interquartile range were among the leading individual predictors, suggesting that within-tract variability in measured metal concentrations contributed to the fitted predictions. Nevertheless, the metal domain contained ten predictors, compared with two housing and two geographic-accessibility predictors. Consequently, its relatively large summed contribution partly reflects the number of variables included in that domain. After normalization by predictor count, housing accounted for 40.04% of relative attribution, geographic accessibility for 32.52%, environmental metals for 18.79%, and soil characteristics for 8.64%.

These two approaches to summarizing domain importance answer different descriptive questions and should be interpreted together. Neither approach establishes independent effects, causal relationships, or the relative biological importance of environmental exposures.

The geographic-accessibility domain also contributed to predictions, particularly through median distance to the airport. However, removing airport distance produced almost no change in predictive performance, with R^2 changing from 0.275 to 0.276. The airport-distance variable was also strongly correlated with the geographic east–west gradient, suggesting that it may represent broader spatial location rather than an airport-specific exposure pathway. Accordingly, its SHAP importance should not be interpreted as evidence that airport proximity affects diabetes prevalence.

The exploratory comparison of soil-metal concentrations between High–High diabetes prevalence clusters and other tracts further qualified the predictive findings. Neither median cadmium concentration (Mann–Whitney U = 83.5, $p = 0.894$; Holm-adjusted $p = 0.894$) nor median lead concentration (U = 59.0, $p = 0.374$; Holm-adjusted $p = 0.747$) differed significantly between these groups. These comparisons were based on only five High–High tracts and therefore had limited statistical power. The absence of statistically detectable group differences does not negate the multivariable predictive contributions identified by SHAP, but it reinforces the need to avoid interpreting feature importance as evidence of an exposure–disease association.

The strong correlation between airport distance and tract-centroid easting, together with the nearly unchanged predictive performance after removing airport distance, suggests that this variable primarily represents an east–west geographic gradient rather than an independently established airport-related exposure pathway.

4.3. Interpretation of SHAP Patterns and Potential Biological Relevance

The out-of-fold SHAP analysis provided insight into how environmental and neighborhood predictors contributed to the Random Forest predictions. The model accommodated nonlinear relationships and interactions without requiring a prespecified linear functional form. However, SHAP values characterize the behavior of the fitted prediction model rather than independent exposure–response functions. Correlations among environmental, housing, and geographic variables may redistribute attribution across predictors, and the small number of census tracts limits the stability of detailed interpretations.

The prominence of cadmium variability and lead variability in the primary specification is relevant to the broader literature on metal exposure and metabolic health [26,33]. The rank of cadmium variability was not stable across spatial partitions or across pre-

dictor specifications, however, and the analysis does not support a claim that cadmium specifically contributes to the spatial pattern in modeled prevalence. Experimental and epidemiological investigations have examined oxidative stress, inflammatory responses, altered glucose regulation, and pancreatic cellular dysfunction as possible mechanisms linking metal exposure to metabolic outcomes [25,36]. The present study did not measure individual exposure doses, biomarkers of metal uptake, metabolic biomarkers, or clinical diabetes diagnoses. The biological literature therefore provides context for interpreting the environmental predictors but does not establish that the soil-metal concentrations measured in Greensboro contributed causally to diabetes prevalence.

An additional interpretive consideration arises from the use of CDC PLACES modeled prevalence as the outcome. PLACES estimates incorporate survey and population information through a small-area estimation framework. Housing and geographic predictors may therefore capture characteristics correlated with variables involved in generating the modeled outcome. The housing-excluded 16-predictor model retained some predictive information ($R^2 = 0.202$), but its lower performance compared with the primary 18-predictor model ($R^2 = 0.275$) does not establish that housing variables have independent etiologic effects. Similarly, the 14-predictor metals-and-soil specification achieved an R^2 of 0.207, indicating modest predictive information without demonstrating a direct environmental exposure effect.

Taken together, these findings support the use of explainable spatial machine learning as an exploratory approach for characterizing neighborhood-level patterns. Stronger conclusions about environmental contributions to diabetes would require independently observed health outcomes, representative exposure measurements, longitudinal information, and analyses capable of addressing individual-level confounding and spatial dependence.

4.4. Spatial Clustering and Public Health Implications

The integration of spatial statistics and explainable machine learning provided complementary descriptions of geographic variation in modeled diabetes prevalence across Greensboro.

The Local Moran's I analysis identified five High–High clusters, two Low–Low clusters, and one Low–High spatial outlier under Queen-contiguity weights and an unadjusted local permutation significance threshold of $p < 0.05$. These findings indicate localized patterns in the modeled outcome under the selected spatial-neighborhood definition.

However, the local cluster classifications were not adjusted for multiple testing and should therefore be considered exploratory. Furthermore, the analysis was based on CDC PLACES modeled prevalence estimates rather than directly observed clinical diabetes outcomes.

The residual spatial autocorrelation analysis provided an additional qualification. Under the primary Queen-contiguity specification, global Moran's I was 0.164, with a one-sided permutation p-value of 0.066 and a two-sided p-value of 0.121. In contrast, k-nearest-neighbor specifications produced positive residual autocorrelation, with permutation p-values ranging from 0.001 to 0.011 for k = 2 through 5.

These differences indicate that conclusions about unexplained spatial structure depend on the spatial-weight definition. The primary model did not eliminate residual spatial dependence under all evaluated neighborhood specifications.

The exploratory cadmium and lead comparisons also suggest that modeled diabetes clusters should not be interpreted as straightforward indicators of elevated concentrations of the measured soil metals. Neither metal showed a statistically significant difference between the five High–High tracts and the remaining study tracts in the reported marginal comparisons.

From a public health perspective, the combined maps and predictive analyses may help identify geographic areas that warrant further investigation. Such areas could be considered for additional environmental sampling, community engagement, or collection of independently observed health information.

However, the present findings are not sufficient to establish clinical risk, identify causal environmental determinants, or determine resource allocation priorities. Any public health prioritization based on these results would require further validation, consideration of community needs, and integration with independent local health data.

More broadly, the study illustrates how spatial machine learning and explainable artificial intelligence can complement geographic epidemiology by describing complex neighborhood-level patterns while making the limitations of prediction and interpretation explicit.

4.5. Strengths and Limitations

This study has several methodological strengths. It integrates environmental measurements, neighborhood characteristics, spatially structured validation, explainable machine learning, and spatial statistical analysis within a single exploratory framework.

The use of five-block spatial cross-validation provided a geographically held-out evaluation rather than relying solely on conventional random cross-validation. The analysis also compared prespecified predictor specifications, including an 18-predictor primary model, a housing-excluded 16-predictor co-primary model, and a broader 43-predictor sensitivity model.

Robustness was further examined across 50 alternative spatial partitions, 999 outcome permutations, 500 outcome-uncertainty realizations informed by CDC PLACES confidence limits, alternative spatial-weight definitions, sampling-density diagnostics, and airport-distance ablation and geographic trend-surface specifications.

These analyses provide a more comprehensive description of predictive sensitivity than a single cross-validation result. Nevertheless, several important limitations remain.

First, the analysis was ecological and included only 37 census tracts. The small sample size relative to the number of predictors creates substantial potential for model instability and limits the precision and generalizability of predictive conclusions. The results cannot be used to infer relationships at the individual level.

Second, the household soil-sampling strategy was non-probability-based and uneven across tracts. Thirteen tracts contained fewer than five participating households, accounting for 26 household records, while the remaining 24 tracts contained 449 records. Such differences may affect the stability and representativeness of tract-level environmental summaries.

A household-count-weighted Random Forest sensitivity analysis yielded an R^2 of 0.233, an RMSE of 4.334 percentage points, and an MAE of 3.417 percentage points, compared with $R^2 = 0.275$, RMSE = 4.213, and MAE = 3.314 for the unweighted primary model (Section 3.8). These findings demonstrate some sensitivity to the treatment of sampling density but do not establish that weighting corrects geographic sampling bias.

Third, the ACS comparison identified socioeconomic and demographic differences between sampled and unsampled census tracts intersecting Greensboro. Sampled tracts had lower median household incomes and owner-occupied home values, higher renter occupancy, and different racial and ethnic compositions. These differences raise concerns about extending the findings to neighborhoods not represented by the household soil-sampling dataset.

Fourth, the diabetes outcome was obtained from CDC PLACES and therefore represents model-based small-area prevalence estimates rather than directly observed clinical diagnoses for the study tracts. The analysis did not include individual-level medical records or independent clinical outcome measurements for validation.

Because PLACES estimates incorporate demographic and census-related information, some housing and geographic predictors may share contextual structure with the outcome estimation process. Although no exact documented direct overlap was identified among the 18 primary predictors, four housing and geographic variables were classified as potential indirect proxies. The housing-excluded co-primary analysis reduced this concern but could not establish complete independence from the PLACES estimation framework. For the housing and neighborhood predictors, the structure shared with the outcome cannot be fully separated in this design; their contributions are therefore reported as markers of neighborhood structure rather than as independent predictive information.

Fifth, uncertainty in the modeled outcome contributes to uncertainty in predictive evaluation. The 500 outcome-uncertainty realizations yielded a median R^2 of 0.259, with a 2.5–97.5th percentile range of 0.182–0.339. These ranges describe sensitivity to the simulated outcome uncertainty and are not confidence intervals for external predictive performance.

Sixth, residual spatial autocorrelation depended on the spatial-neighborhood definition. Although the primary Queen-contiguity result did not meet the conventional 0.05 significance threshold, several k-nearest-neighbor specifications identified positive residual spatial dependence. This suggests that the fitted model did not fully account for spatially structured variation under all evaluated weighting schemes.

Seventh, environmental measurements were aggregated to census tracts and were not linked to individual diabetes diagnoses, personal exposure histories, or biomarkers of internal metal dose. The analysis did not establish whether measured residential environmental conditions preceded diabetes development. Residential mobility, exposure duration, occupational exposures, and previous residential environments were not available.

Individual-level body mass index, physical activity, diet, family history, genetic susceptibility, and other established diabetes-related characteristics were also unavailable as independently observed variables for the study population. Their absence limits interpretation of the environmental and neighborhood predictors. These characteristics are themselves geographically patterned and are a likely contributor to the residual spatial autocorrelation reported in Section 3.8, which limits attribution of the observed spatial pattern to the measured predictors. Modeled CDC PLACES estimates of obesity, physical inactivity, and smoking were considered as substitute predictors and were not used, because they are produced by the same multilevel regression and poststratification procedure applied to the same BRFSS and census inputs as the diabetes outcome [60]; including them would compound the shared-structure concern described above rather than resolve it.

Finally, the cross-sectional ecological design prevents causal inference. SHAP contributions describe fitted model predictions, and Local Moran's I identifies geographic patterns under selected spatial weights. Neither provides evidence that particular environmental exposures caused diabetes or that changing a predictor would change diabetes prevalence.

These limitations underscore the need to interpret the study as an exploratory analysis of modeled small-area diabetes prevalence rather than a validated clinical risk-prediction or causal exposure-assessment framework.

4.6. Future Directions

Future work could test how well this framework generalizes beyond Greensboro by applying it to larger and more diverse urban settings. Incorporating longitudinal health records, individual-level clinical data, and additional environmental exposures, such as noise, air pollution, green space, and healthcare accessibility, could improve predictive performance and deepen understanding of local diabetes determinants [9,32].

Methodological development should prioritize external validation, improved handling of spatial dependence, uncertainty propagation, and evaluation of model stability

across alternative geographic partitions. Larger datasets would also permit more reliable investigation of nonlinear relationships and interactions.

The framework could be evaluated in communities with known environmental contamination, but applying the Greensboro-trained model to other locations would require independent validation and, where appropriate, recalibration or retraining using local data.

Ultimately, combining representative environmental measurements, independently observed health outcomes, and geographically robust prediction methods may help determine whether the exploratory patterns identified here generalize to broader populations and provide actionable information for public health research.

5. Conclusions

This study developed an explainable spatial machine-learning framework to examine geographic variation in CDC PLACES 2024 modeled diabetes prevalence across 37 census tracts in Greensboro, North Carolina.

The primary 18-predictor Random Forest model achieved a pooled out-of-fold R^2 of 0.275, RMSE of 4.213 percentage points, and MAE of 3.314 percentage points under five-block spatial cross-validation. Although the model captured some variation in modeled tract-level prevalence, most variation remained unexplained.

Predictive performance was sensitive to predictor composition and geographic partitioning. Across 50 alternative spatial partitions, the primary reduced model showed a modest performance advantage over the full 43-predictor specification in paired RMSE comparisons.

SHAP analysis identified predictive contributions from environmental-metal, housing, geographic-accessibility, and soil variables. Environmental metals received the largest summed absolute attribution (51.12%), followed by housing (21.78%), geographic accessibility (17.69%), and soil characteristics (9.40%). After feature-count normalization, which accounts for the unequal number of predictors per domain, housing accounted for 40.04% of relative attribution, followed by geographic accessibility (32.52%), environmental metals (18.79%), and soil characteristics (8.64%). These contributions describe model predictions rather than independent or causal effects, and individual predictor rankings were not stable enough to support claims about specific metals.

Spatial analyses identified exploratory geographic clustering of modeled diabetes prevalence, while residual spatial autocorrelation varied with the neighborhood definition. Comparisons of tract-level median cadmium and lead concentrations between High–High clusters and other tracts did not identify statistically significant differences.

Overall, the findings demonstrate the potential utility of combining environmental data, spatial validation, and explainable machine learning for exploratory small-area health research. However, modest predictive performance, limited spatial sample size, uneven household sampling, the model-based diabetes outcome, and the cross-sectional ecological design require cautious interpretation.

Future studies should prioritize independent clinical outcomes, representative environmental sampling, longitudinal exposure information, and external validation across larger geographic populations.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ijerph23111349/s1, Table S1: Data domains and analytical roles; Table S2: Predictor specifications and pooled out-of-fold performance; Table S3: Classi-fi-cat-ion of 18 primary predictors; Table S4: Domain-level SHAP attribution in the primary 18-predictor model; Table S5: Performance across 50 spatial partitions; Table S6: Outcome permutation and outcome uncertainty sensitivity (500 realizations); Table S7: Residual Moran's I under alternative spatial weights; Table S8: Household-count weighting sensitivity; Table S9: Airport-distance and trend-

surface sensitivity; Table S10: ACS comparison of sampled and unsampled tracts; Table S11: High–High versus remaining-tract metal comparisons; Figure S1: Observed versus spatially out-of-fold predicted diabetes prevalence for the primary 18-predictor and housing-excluded 16-predictor models; Figure S2: Comparison of shared-predictor SHAP attributions for the primary and housing-excluded models; and Figure S3, Domain-level SHAP attribution comparison between primary and housing-excluded models.

Author Contributions: Conceptualization, E.O.-G.; methodology, R.T. and E.O.-G.; software, R.T. and E.O.-G.; validation, R.T. and E.O.-G.; formal analysis, R.T. and E.O.-G.; investigation, R.T. and E.O.-G.; resources, E.O.-G.; data curation, R.T. and E.O.-G.; writing—original draft preparation, R.T. and E.O.-G.; writing—review and editing, R.T. and E.O.-G.; visualization, R.T.; supervision, E.O.-G.; project administration, E.O.-G. All authors have read and agreed to the published version of the manuscript.

Funding: This research 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.

Institutional Review Board Statement: The parent household soil-sampling study was conducted under a protocol reviewed by the North Carolina A&T State University Institutional Review Board, with all participant data deidentified. The present work is a secondary analysis of those existing, deidentified data. The North Carolina A&T State University Institutional Review Board determined that the present research does not constitute human subjects research under applicable federal regulations and does not require IRB approval (Study Number HS24-0081; determination dated 23 May 2024).

Informed Consent Statement: Informed consent was obtained from participating households during the recruitment and sampling visits of the parent study. The present work is a secondary analysis of existing deidentified data that the Institutional Review Board determined does not constitute human subjects research, so no additional participant consent was required for this analysis.

Data Availability Statement: The census-tract diabetes prevalence estimates analyzed in this study are publicly available from the CDC PLACES project. The environmental, soil, housing, and accessibility data supporting the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Tasnia, R.; Obeng-Gyasi, E. Evaluating Machine Learning Models for Classifying Diabetes Using Demographic, Clinical, Lifestyle, Anthropometric, and Environmental Exposure Factors. *Toxics* **2026**, *14*, 76. [CrossRef] [PubMed]
2. Manrique, A.; Chen, A.; Wang, C.; Rohlwing, K.; Li, P.; Guo, J.S.; Shenkman, E.A.; Shao, H.; Bian, J.; Guo, Y. The intersection of place and health: Examining the association of social determinants of health and diabetes in children and adolescents in Florida. *BMC Public Health* **2026**, *26*, 1435. [CrossRef] [PubMed]
3. Egede, L.E.; Walker, R.J.; Linde, S. Association between population level social risk factors and prevalence of diabetes. *BMC Health Serv. Res.* **2025**, *25*, 1252. [CrossRef] [PubMed]
4. World Health Organization. Diabetes. Available online: https://www.who.int/news-room/fact-sheets/detail/diabetes (accessed on 18 October 2026).
5. Centers for Disease Control and Prevention. National Diabetes Statistics Report. Available online: https://www.cdc.gov/diabetes/php/data-research/index.html (accessed on 18 October 2026).
6. North Carolina Department of Health and Human Services. *2025 North Carolina State Health Assessment*; North Carolina Department of Health and Human Services: Raleigh, NC, USA, 2025.
7. Guilford County Department of Health and Human Services. *2023–2024 Guilford County Community Health Assessment*; Guilford County Department of Health and Human Services: Greensboro, NC, USA, 2025.
8. Centers for Disease Control and Prevention. PLACES: Local Data for Better Health, Census Tract Data, 2024 Release. Available online: https://www.cdc.gov/places/ (accessed on 18 October 2026).

9. Takele, W.W.; Lim, S.; Dalli, L.L.; Beare, R.; Adhinugraha, K.; Taniar, D.; Wang, S.; Boyle, J.A. Natural environment and gestational diabetes risk in Australia: A spatiotemporal ecological regression approach. *BMC Public Health* **2026**, *26*, 1206. [CrossRef] [PubMed]
10. Rose, L.H.; Burus, T.; Lacy, M.E.; Vanderford, N.L. Spatial analysis of diabetes prevalence and diabetes-related cancer mortality in the United States. *Cancer Causes Control* **2025**, *37*, 10. [CrossRef] [PubMed]
11. Gavaghan, H. Lead, unsafe at any level. *Bull. World Health Organ.* **2002**, *80*, 82. [PubMed]
12. Obeng-Gyasi, E.; Ferguson, A.C.; Stamatakis, K.A.; Province, M.A. Combined Effect of Lead Exposure and Allostatic Load on Cardiovascular Disease Mortality—A Preliminary Study. *Int. J. Environ. Res. Public Health* **2021**, *18*, 6879. [CrossRef] [PubMed]
13. Qin, F.; Shan, X.Q.; Wei, B. Effects of low-molecular-weight organic acids and residence time on desorption of Cu, Cd, and Pb from soils. *Chemosphere* **2004**, *57*, 253–263. [CrossRef] [PubMed]
14. Schachter, A.E.; Gailey, A.; Egendorf, S.P.; Mielke, H.W. Mechanisms of children's soil exposure. *Curr. Probl. Pediatr. Adolesc. Health Care* **2020**, *50*, 100742. [CrossRef] [PubMed]
15. Hansson, S.V.; Grusson, Y.; Chimienti, M.; Claustres, A.; Jean, S.; Le Roux, G. Legacy Pb pollution in the contemporary environment and its potential bioavailability in three mountain catchments. *Sci. Total Environ.* **2019**, *671*, 1227–1236. [CrossRef]
16. Mielke, H.W.; Gonzales, C.R.; Powell, E.T.; Laidlaw, M.A.S.; Berry, K.J.; Mielke, P.W., Jr.; Egendorf, S.P. The concurrent decline of soil lead and children's blood lead in New Orleans. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 22058–22064. [CrossRef] [PubMed]
17. Klaminder, J.; Bindler, R.; Emteryd, O.; Appleby, P.; Grip, H. Estimating the mean residence time of lead in the organic horizon of boreal forest soils using 210-lead, stable lead and a soil chronosequence. *Biogeochemistry* **2006**, *78*, 31–49. [CrossRef]
18. Obeng-Gyasi, E.; Roostaei, J.; Gibson, J.M. Lead Distribution in Urban Soil in a Medium-Sized City: Household-Scale Analysis. *Environ. Sci. Technol.* **2021**, *55*, 3696–3705. [CrossRef] [PubMed]
19. Brown, S.L.; Chaney, R.L.; Hettiarachchi, G.M. Lead in urban soils: A real or perceived concern for urban agriculture? *J. Environ. Qual.* **2016**, *45*, 26–36. [CrossRef] [PubMed]
20. Laidlaw, M.A.; Filippelli, G.M.; Brown, S.; Paz-Ferreiro, J.; Reichman, S.M.; Netherway, P.; Truskewycz, A.; Ball, A.S.; Mielke, H.W. Case studies and evidence-based approaches to addressing urban soil lead contamination. *Appl. Geochem.* **2017**, *83*, 14–30. [CrossRef]
21. Yan, K.; Dong, Z.; Wijayawardena, M.A.A.; Liu, Y.; Naidu, R.; Semple, K. Measurement of soil lead bioavailability and influence of soil types and properties: A review. *Chemosphere* **2017**, *184*, 27–42. [CrossRef] [PubMed]
22. Agency for Toxic Substances and Disease Registry. *Toxicological Profile for Cadmium*; U.S. Department of Health and Human Services: Atlanta, GA, USA, 2012.
23. Agency for Toxic Substances and Disease Registry. *Toxicological Profile for Lead*; U.S. Department of Health and Human Services: Atlanta, GA, USA, 2020.
24. U.S. Environmental Protection Agency. *Residential Soil Lead Directive for CERCLA Sites and RCRA Hazardous Waste Cleanup Facilities*; Office of Land and Emergency Management: Washington, DC, USA, 2025. Available online: https://www.epa.gov/superfund/residential-soil-lead-directive-cercla-sites-and-rcra-hazardous-waste-cleanup-facilities (accessed on 18 October 2026).
25. Ahamed, M.; Siddiqui, M.K.J. Low level lead exposure and oxidative stress: Current opinions. *Clin. Chim. Acta* **2007**, *383*, 57–64. [CrossRef] [PubMed]
26. Satarug, S.; Garrett, S.H.; Sens, M.A.; Sens, D.A. Cadmium, environmental exposure, and health outcomes. *Environ. Health Perspect.* **2010**, *118*, 182–190. [CrossRef] [PubMed]
27. U.S. Environmental Protection Agency. Lead; Identification of Dangerous Levels of Lead; Final Rule. 40 CFR Part 745, Subpart D (Section 745.65). *Fed. Regist.* **2001**, *66*, 1206–1240.
28. U.S. Department of Housing and Urban Development. *Requirements for Notification, Evaluation and Reduction of Lead-Based Paint Hazards in Federally Owned Residential Property and Housing Receiving Federal Assistance (Lead Safe Housing Rule)*; 24 CFR Part 35; U.S. Department of Housing and Urban Development: Washington, DC, USA, 2000.
29. U.S. Environmental Protection Agency. Regional Screening Levels (RSLs)—Generic Tables. Available online: https://www.epa.gov/risk/regional-screening-levels-rsls-generic-tables (accessed on 18 October 2026).
30. North Carolina General Assembly. Section 130A-131.9C: Abatement and remediation. In *North Carolina General Statutes*; North Carolina General Assembly: Raleigh, NC, USA, 1997.
31. North Carolina Department of Environmental Quality. Risk Evaluation Resources. Available online: https://www.deq.nc.gov/ (accessed on 18 October 2026).
32. Liu, J.; Liu, C.; Liu, Z.; Zhou, Y.; Li, X.; Yang, Y. Spatial analysis of air pollutant exposure and its association with metabolic diseases using machine learning. *BMC Public Health* **2025**, *25*, 831. [CrossRef] [PubMed]
33. Adokwe, J.B.; Pouyfung, P.; Kuraeiad, S.; Wongrith, P.; Inchai, P.; Yimthiang, S.; Satarug, S.; Khamphaya, T. Concurrent Lead and Cadmium Exposure Among Diabetics: A Case-Control Study of Socio-Demographic and Consumption Behaviors. *Nutrients* **2025**, *17*, 710. [CrossRef] [PubMed]

34. Javaid, A.; Akbar, I.; Javed, H.; Khan, U.; Iftikhar, H.; Zahra, D.; Rashid, F.; Ashfaq, U.A. Role of heavy metals in diabetes: Mechanisms and treatment strategies. *Crit. Rev. Eukaryot. Gene Expr.* **2021**, *31*, 65–80. [CrossRef] [PubMed]
35. Nguyen, H.D.; Oh, H.; Jo, W.H.; Hoang, N.H.M.; Kim, M.S. Mixtures modeling identifies heavy metals and pyrethroid insecticide metabolites associated with obesity. *Environ. Sci. Pollut. Res. Int.* **2022**, *29*, 20379–20397. [CrossRef] [PubMed]
36. Hong, H.; Xu, Y.; Xu, J.; Zhang, J.; Xi, Y.; Pi, H.; Yang, L.; Yu, Z.; Wu, Q.; Meng, Z. Cadmium exposure impairs pancreatic β-cell function and exaggerates diabetes by disrupting lipid metabolism. *Environ. Int.* **2021**, *149*, 106406. [CrossRef] [PubMed]
37. Wang, B.; Zhang, W.; Chen, C.; Chen, Y.; Xia, F.; Wang, N.; Lu, Y. Lead exposure and impaired glucose homeostasis in Chinese adults: A repeated measures study with 5 years of follow-up. *Ecotoxicol. Environ. Saf.* **2022**, *243*, 113953. [CrossRef] [PubMed]
38. Hou, D.; Jia, X.; Wang, L.; McGrath, S.P.; Zhu, Y.-G.; Hu, Q.; Zhao, F.-J.; Bank, M.S.; O'Connor, D.; Nriagu, J. Global soil pollution by toxic metals threatens agriculture and human health. *Science* **2025**, *388*, 316–321. [CrossRef] [PubMed]
39. Alrashdi, M.M.; Ragazzon-Smith, A.; Strashnov, I.; Polya, D.A. Chemical analysis of toxic elements: Total cadmium, lead, mercury, arsenic and inorganic arsenic in local and imported rice consumed in the Kingdom of Saudi Arabia. *Environ. Geochem. Health* **2024**, *46*, 518. [CrossRef] [PubMed]
40. Vinceti, M.; Filippini, T.; Wise, L.A.; Rothman, K.J. A systematic review and dose-response meta-analysis of exposure to environmental selenium and the risk of type 2 diabetes in nonexperimental studies. *Environ. Res.* **2021**, *197*, 111210. [CrossRef] [PubMed]
41. Vermeulen, R.; Schymanski, E.L.; Barabási, A.-L.; Miller, G.W. The exposome and health: Where chemistry meets biology. *Science* **2020**, *367*, 392–396. [CrossRef] [PubMed]
42. Vineis, P.; Robinson, O.; Chadeau-Hyam, M.; Dehghan, A.; Mudway, I.; Dagnino, S. What is new in the exposome? *Environ. Int.* **2020**, *143*, 105887. [CrossRef] [PubMed]
43. Salerno, P.R.; Qian, A.; Dong, W.; Deo, S.; Nasir, K.; Rajagopalan, S.; Al-Kindi, S. County-level socio-environmental factors and obesity prevalence in the United States. *Diabetes Obes. Metab.* **2024**, *26*, 1766–1774. [CrossRef] [PubMed]
44. Motairek, I.; Makhlouf, M.H.; Rajagopalan, S.; Al-Kindi, S. The exposome and cardiovascular health. *Can. J. Cardiol.* **2023**, *39*, 1191–1203. [CrossRef] [PubMed]
45. Harville, E.; Grady, S.; Langston, M.; Juarez, P.; Vilda, D.; Wallace, M. The public health exposome and pregnancy-related mortality in the United States: A high-dimensional computational analysis. *BMC Public Health* **2022**, *22*, 2097. [CrossRef] [PubMed]
46. Cleal, J.K.; Poore, K.R.; Lewis, R.M. The placental exposome, placental epigenetic adaptations and lifelong cardio-metabolic health. *Mol. Asp. Med.* **2022**, *87*, 101095. [CrossRef] [PubMed]
47. Luo, H.M.; Hu, W.B.; Xu, Y.J.; Zheng, X.Y.; He, Q.; Lyu, L.; Meng, R.L.; Xu, X.J.; Zou, F. Identifying High-Risk Areas for Type 2 Diabetes Mellitus Mortality in Guangdong, China: Spatiotemporal Clustering and Socioenvironmental Determinants. *Biomed. Environ. Sci.* **2025**, *38*, 585–597. [CrossRef] [PubMed]
48. Hill, J.O.; Galloway, J.M.; Goley, A.; Marrero, D.G.; Minners, R.; Montgomery, B.; Peterson, G.E.; Ratner, R.E.; Sanchez, E.; Aroda, V.R. Scientific statement: Socioecological determinants of prediabetes and type 2 diabetes. *Diabetes Care* **2013**, *36*, 2430–2439. [CrossRef] [PubMed]
49. Spanakis, E.K.; Golden, S.H. Race/ethnic difference in diabetes and diabetic complications. *Curr. Diab. Rep.* **2013**, *13*, 814–823. [CrossRef] [PubMed]
50. Dugani, S.B.; Wood-Wentz, C.M.; Mielke, M.M.; Bailey, K.R.; Vella, A. Assessment of disparities in diabetes mortality in adults in US rural vs nonrural counties, 1999–2018. *JAMA Netw. Open* **2022**, *5*, e2232318. [CrossRef] [PubMed]
51. Li, W.; Huang, G.; Tang, N.; Lu, P.; Jiang, L.; Lv, J.; Qin, Y.; Lin, Y.; Xu, F.; Lei, D. Effects of heavy metal exposure on hypertension: A machine learning modeling approach. *Chemosphere* **2023**, *337*, 139435. [CrossRef] [PubMed]
52. Zou, Q.; Qu, K.; Luo, Y.; Yin, D.; Ju, Y.; Tang, H. Predicting Diabetes Mellitus with Machine Learning Techniques. *Front. Genet.* **2018**, *9*, 515. [CrossRef] [PubMed]
53. Shen, M.; Zhang, Y.; Zhan, R.; Du, T.; Shen, P.; Lu, X.; Liu, S.; Guo, R.; Shen, X. Predicting the risk of cardiovascular disease in adults exposed to heavy metals: Interpretable machine learning. *Ecotoxicol. Environ. Saf.* **2025**, *290*, 117570. [CrossRef] [PubMed]
54. Ogunpola, A.; Saeed, F.; Basurra, S.; Albarrak, A.M.; Qasem, S.N. Machine Learning-Based Predictive Models for Detection of Cardiovascular Diseases. *Diagnostics* **2024**, *14*, 144. [CrossRef] [PubMed]
55. Pasambo, Y.; Efendi, F.; Pradipta, R.O.; Murti, F.A.K.; Syahri, A.; Alam, S. Socioeconomic Disparities Among People with Diabetes Mellitus (DM) in Indonesia. *Sage Open* **2025**, *15*, 21582440251385984. [CrossRef]
56. U.S. Environmental Protection Agency. *Method 6200: Field Portable X-Ray Fluorescence Spectrometry for the Determination of Elemental Concentrations in Soil and Sediment*; U.S. Environmental Protection Agency: Washington, DC, USA, 2007. Available online: https://www.epa.gov/sites/default/files/2015-12/documents/6200.pdf (accessed on 18 October 2026).
57. Soil Survey Staff. *Soil Survey Geographic (SSURGO) Database*; U.S. Department of Agriculture, Natural Resources Conservation Service. Available online: https://websoilsurvey.nrcs.usda.gov/ (accessed on 18 October 2026).
58. U.S. Census Bureau. TIGER/Line Shapefiles, 2020 Census Tracts. Available online: https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.html (accessed on 18 October 2026).

59. U.S. Census Bureau. American Community Survey 5-Year Estimates. Available online: https://www.census.gov/programs-surveys/acs (accessed on 18 October 2026).
60. Greenlund, K.J.; Lu, H.; Wang, Y.; Matthews, K.A.; LeClercq, J.M.; Lee, B.; Carlson, S.A. PLACES: Local data for better health. *Prev. Chronic Dis.* **2022**, *19*, E31. [CrossRef] [PubMed]
61. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
62. Roberts, D.R.; Bahn, V.; Ciuti, S.; Boyce, M.S.; Elith, J.; Guillera-Arroita, G.; Hauenstein, S.; Lahoz-Monfort, J.J.; Schröder, B.; Thuiller, W.; et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **2017**, *40*, 913–929. [CrossRef]
63. McKinney, W. Data structures for statistical computing in Python. In Proceedings of the 9th Python in Science Conference, Austin, TX, USA, 28 June–3 July 2010; pp. 56–61. [CrossRef]
64. Harris, C.R.; Millman, K.J.; van der Walt, S.J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N.J.; et al. Array programming with NumPy. *Nature* **2020**, *585*, 357–362. [CrossRef] [PubMed]
65. Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; et al. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat. Methods* **2020**, *17*, 261–272. [CrossRef] [PubMed]
66. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
67. Esri. *ArcGIS Pro, version 3.6.0*; Esri: Redlands, CA, USA, 2025.
68. Lundberg, S.M.; Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*; Curran Associates: Red Hook, NY, USA, 2017; pp. 4765–4774.
69. Moran, P.A.P. Notes on continuous stochastic phenomena. *Biometrika* **1950**, *37*, 17–23. [CrossRef]
70. Anselin, L. Local Indicators of Spatial Association—LISA. *Geogr. Anal.* **1995**, *27*, 93–115. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.