Project 2: Analyzing Patterns of Asthma Prevalence in the DMV

Alonzo Finch
Jiatong Peng
Wei Zhang

---

1. Introduction

Asthma is a health condition that has continued to increase in prevalence over time. The project goal was to quantify the spatial relationship between environmental and socioeconomic factors and asthma prevalence, focusing on the District of Columbia, Maryland, and Virginia (DMV region). Ideally, the analysis could identify if there is a disproportionate burden on communities affected by other socioeconomic and/or environmental factors.

This project combined multiple data sources, including data from the Allergy and Asthma Network, the U.S. Centers for Disease Control and Prevention, and the Federal Emergency Management Agency. After utilizing Importance Ranking based on Random Forest modelling and the multicollinearity checks, 17 variables were selected covering environmental factors, social vulnerability indicators and flood risk exposure.

---

2. Data

The project utilized data from five sources. All of the data is presented as tabular areal data with the area of aggregation being at the census tract level. The combined dataset contains a collection of a variety of social, economic, environmental, demographic, and health factors that could influence asthma prevalence across the DMV. All the datasets below were derived from 2022 data for consistency.

2.1 Asthma Equity Dataset

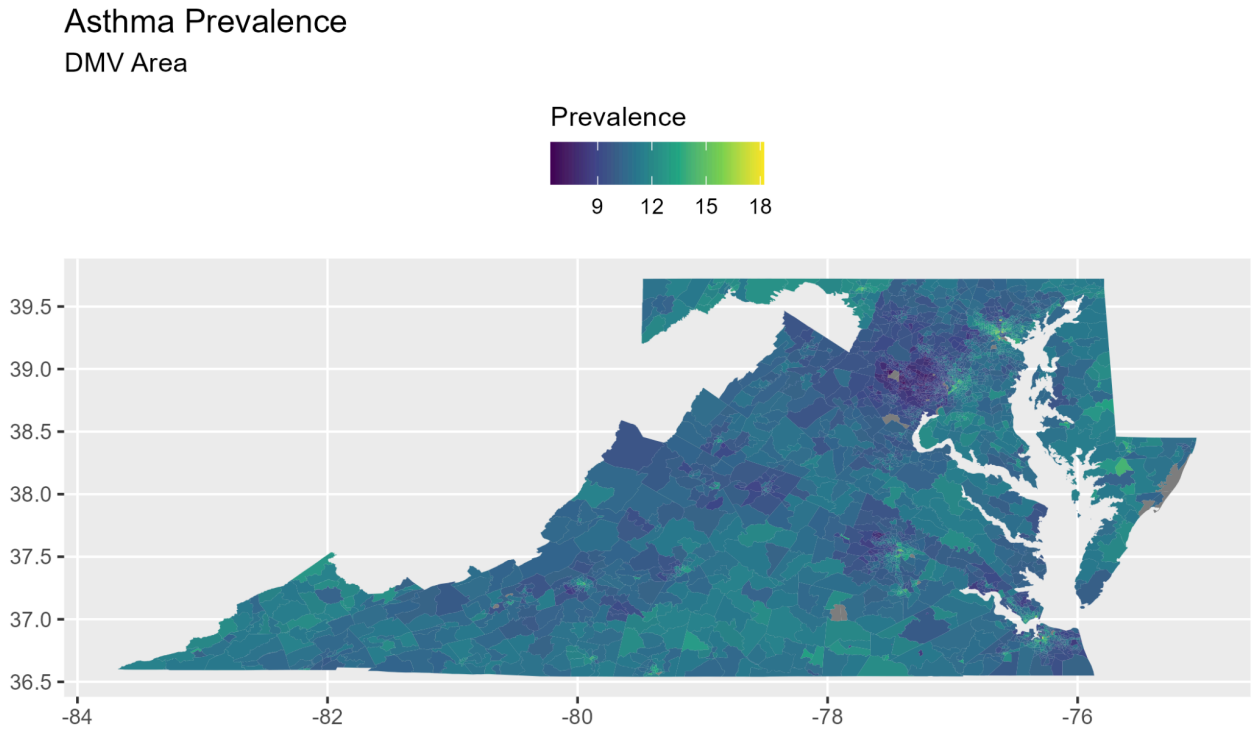

Figure 2.1.1 Geographic Plot of Asthma Prevalence in the DMV

The Allergy and Asthma Network is a nonprofit organization focused on outreach, education, advocacy, and research on allergies and asthma. This dataset contains the main variable of interest, asthma prevalence (Figure 2.1.1). It also contains other aggregate covariates such as particulate matter (a combined form of PM 2.5 and O3 concentration; Figure 2.1.2), redlining, labor participation rate, housing stress, and other variables. This Asthma Equity dataset was the basis of our combined dataset due to the low number of tracts without data.

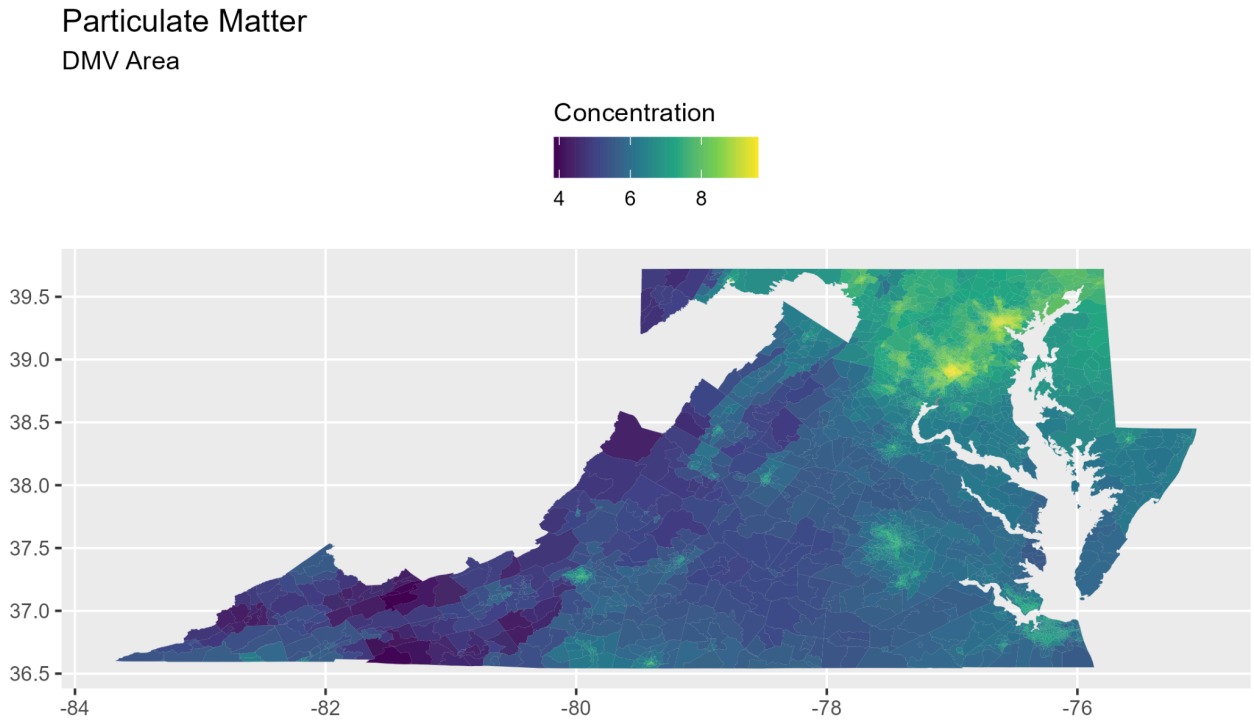

Figure 2.1.2 Geographic Plot of Particulate Matter Concentration in the DMV

2.2 CDC Places

The U.S. Centers for Disease Control and Prevention (CDC) provides three databases that were used for this project. The first of these is the CDC PLACES database. CDC PLACES provides health and health-related areal data across the United States (Centers for Disease Control and Prevention, 2024). It groups its variables as health outcomes, prevention, health risk behaviors, disabilities, health status, health-related social need, and social determinants of health. For the purposes of our analysis we extracted the crude estimates of these variables, such as crude estimates of adult asthma prevalence, adult obesity prevalence, smoking prevalence, COPD prevalence, etc.

2.3 CDC Environmental Justice Index

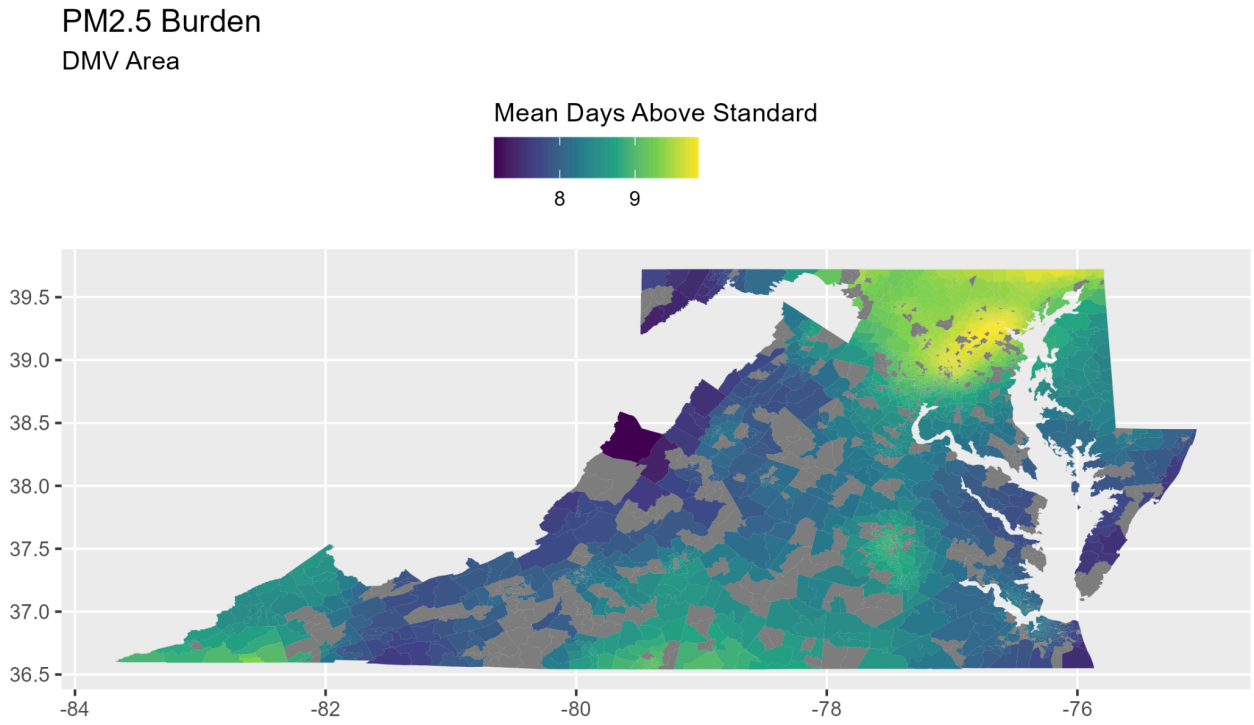

Figure 2.3.1 Geographic Plot of PM 2.5 Burden in the DMV

The CDC Environmental Justice Index (EJI) was used in place of a similar dataset from the U.S. Environmental Protection Agency (EPA). The EJI contains estimates for 36 environmental, social, and health factors (Centers for Disease Control and Prevention & Agency for Toxic Substances and Disease Registry, 2024), as well as percentile flags and binary representations of these factors based on a threshold. The EJI also groups these variables into ten domains and three overarching modules. The EJI dataset contains both variables directly estimated for the purposes of this index database, as well as estimates from other federal databases, such as CDC PLACES, the U.S. Census Bureau, CDC Social Vulnerability Index (SVI), the EPA, and others. For our project, our usage of the EJI dataset was for the environmental variables such as mean days above standard PM 2.5 (Figure 2.3.1) or O3 concentration and proximity to highways and roadways, (Figure 2.3.2), railways, and airports.

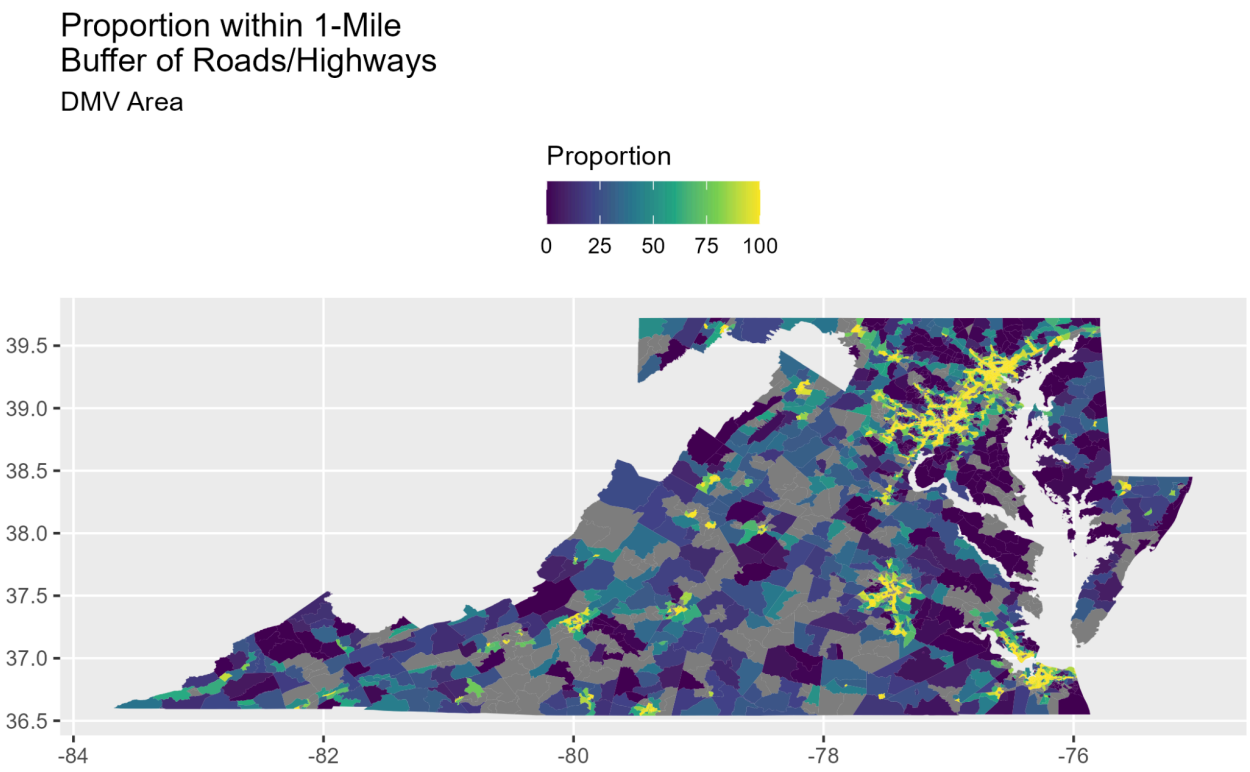

Figure 2.3.2 Geographic Plot of Proximity to Roads and Highways in the DMV

2.4 CDC Social Vulnerability Index

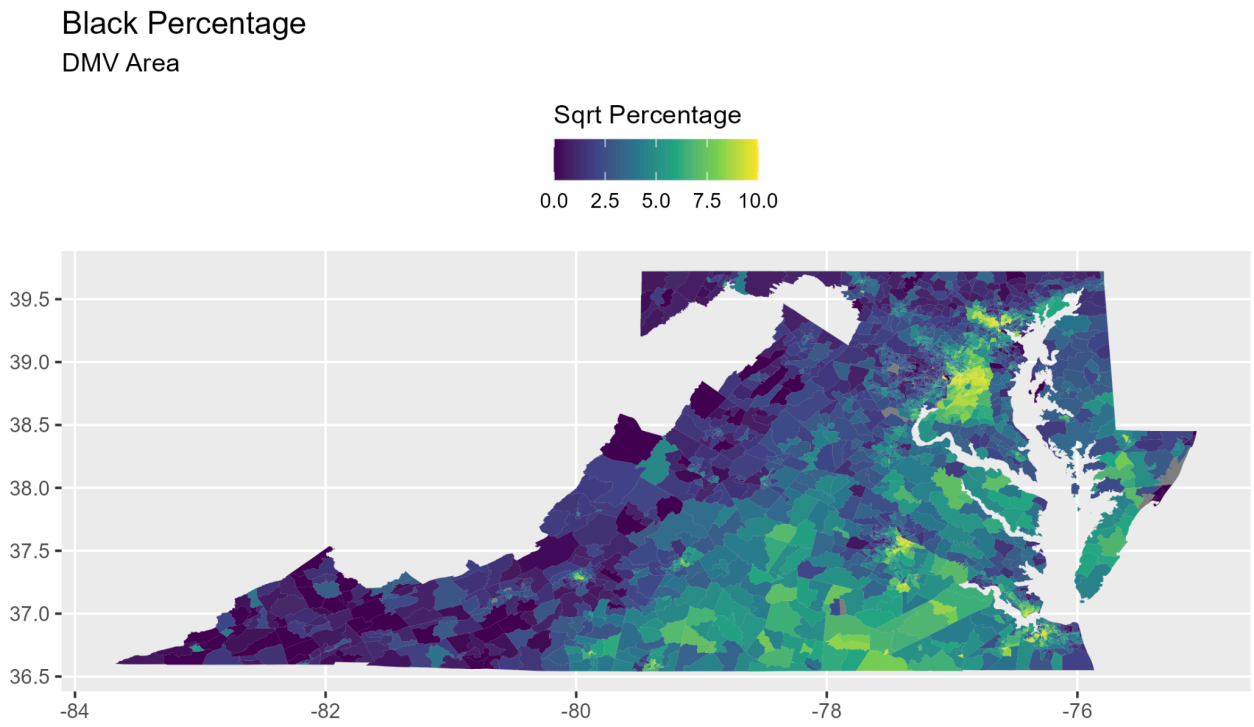

Figure 2.4.1 Geographic Plot of Percentage of Black Individuals in the DMV

The CDC Social Vulnerability Index (SVI) focuses on socioeconomic and demographic factors. Much like the earlier CDC EJI dataset, the dataset contains estimates, percentile flags, and binary indicators based on a threshold. Some of these variables were estimates of the number of minorities (such as black, asian, etc.), percentage of individuals above 65, number of households with more people than rooms, number of housing structures with 10 or more units, and others.

2.5 FEMA National Flood Hazard Layers

The Federal Emergency Management Agency National Flood Hazard Layer (FEMA NFHL) is a geospatial database that contains current effective flood hazard data (Federal Emergency Management Agency, 2025). A key element of this project was to understand what, if any, impact flood risk had on asthma prevalence throughout the DMV. Flood risk is not directly available from the FEMA NFHL, so flood risk had to be estimated by computing the intersection between flood hazard zones and census tracts, plotted below in Figure 2.5.1.

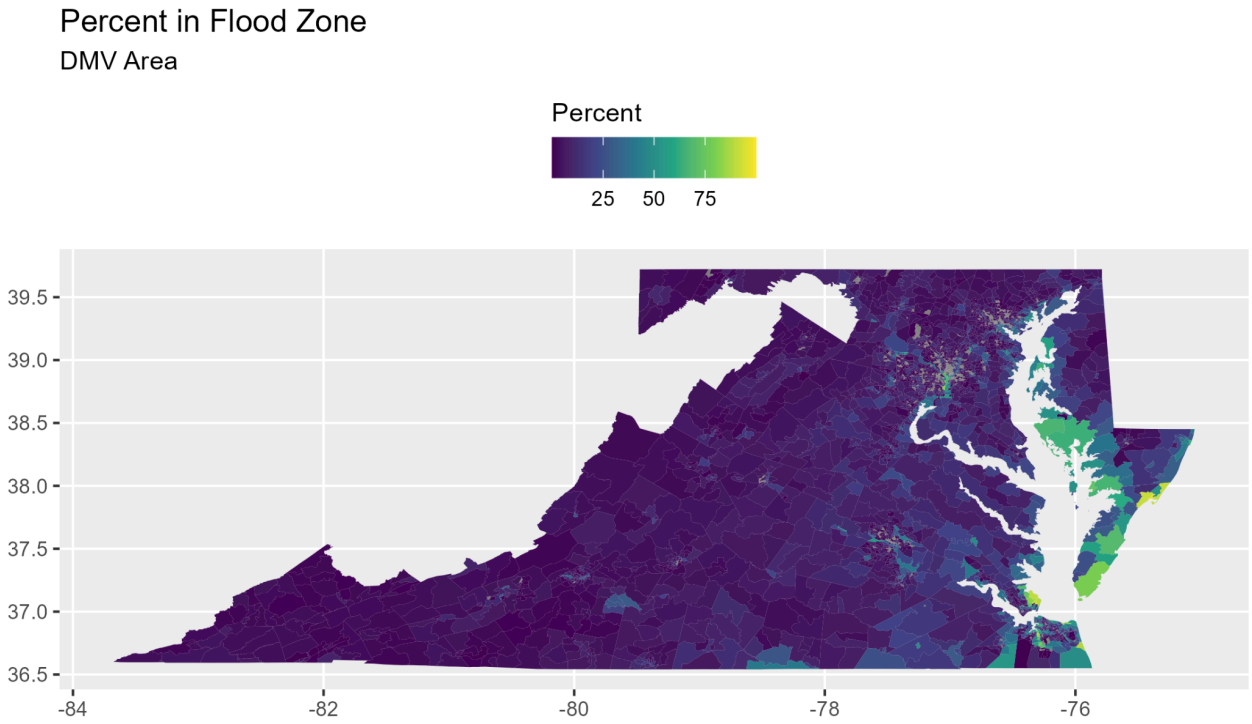

Figure 2.5.1 Geographic Plot of Flood Risk in the DMV

---

3. Variable Selection

3.1 Overview

In this part, we wanted to identify a subset of variables that could effectively explain the variation in asthma prevalence across census tracts in DMV region, at the same time avoiding multicollinearity.

Before this final version, we have tried to manually select several variables and solve the multicollinearity, but it led to unexplainable results in subsequent modeling. Therefore, we tried

the following steps to get the final selection, using Random Forest model to rank all variable importance and using multi-stages VIF screening to get the final selection.

3.2 Importance Ranking Based on Random Forest

We first used two versions of Random Forest models to get a ranking of all variable importance.

The dataset was randomly splitted into a training set (80%) and a testing set (20%). Then the Random Forest models were fitted on the training set, and variable importance was evaluated using the percentage increase in mean squared error (%IncMSE) to quantify how much the prediction error increases when a particular variable is permuted. The higher %IncMSE indicates greater importance.

Considering the high proportion of missing values in the dataset, we tried two random forest models. In the first model, our missing data strategy is omitting rows with NA values entirely from training and prediction. In the second model, our missing data strategy is imputing missing values using median (for numeric data) or mode (for factor data). Both models used 500 trees (ntree = 500) and the default number of predictors per split (mtry). A random seed was set for reproducibility (set.seed(6289)).

For both models, we wanted to see the model performance. Therefore, we evaluated the models through Train Percentage of Variance Explained, Test RMSE, Test R-squared, OOB Error Curve and Observation Coverage.

Train Percentage of Variance Explained represents how well the Random Forest models fit the training data. A higher percentage suggests a better fit to the training data.

Test RMSE represents the root mean squared error on the testing set, measuring the average magnitude of prediction error. A lower RMSE indicates more accurate predictions on unseen data.

Test R-squared represents the proportion of variance in the testing set that is explained by the models' predictions. A higher Test R-squared indicates better prediction.

OOB Error Curve, displayed in Figure 3.2.1 and Figure 3.2.2, shows how the models' out-of-bag error decreases as the number of trees increases, visualizing model stability.

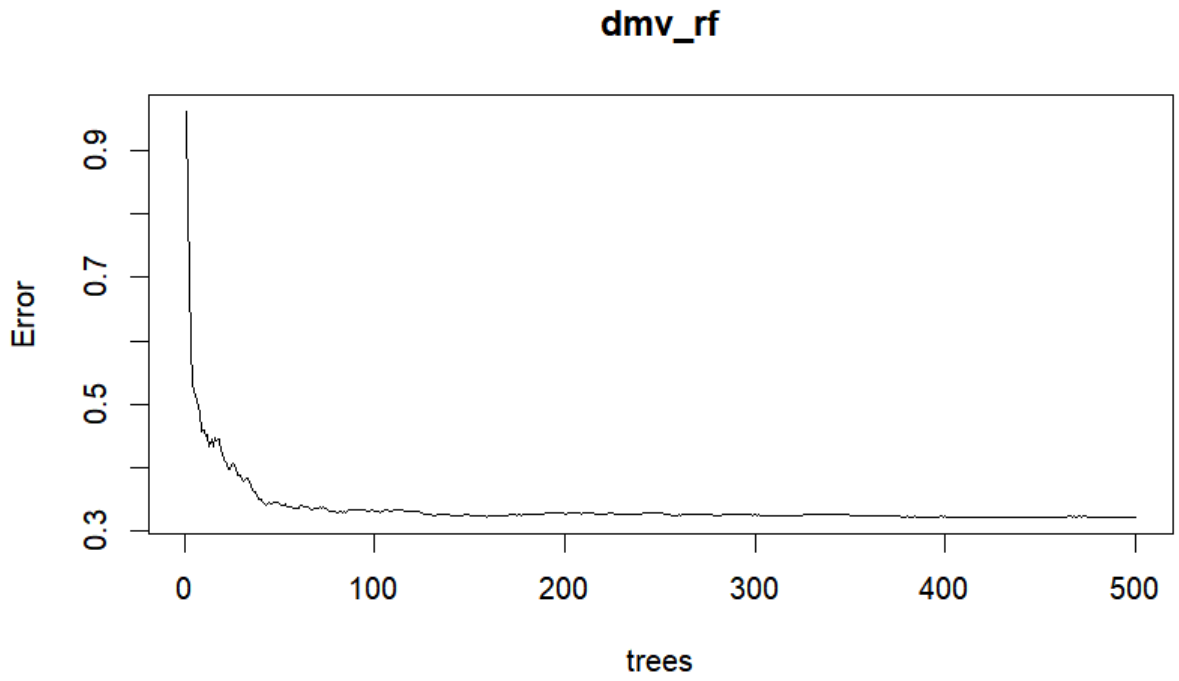

Figure 3.2.1 OBB Error Curve of Random Forest Model 1

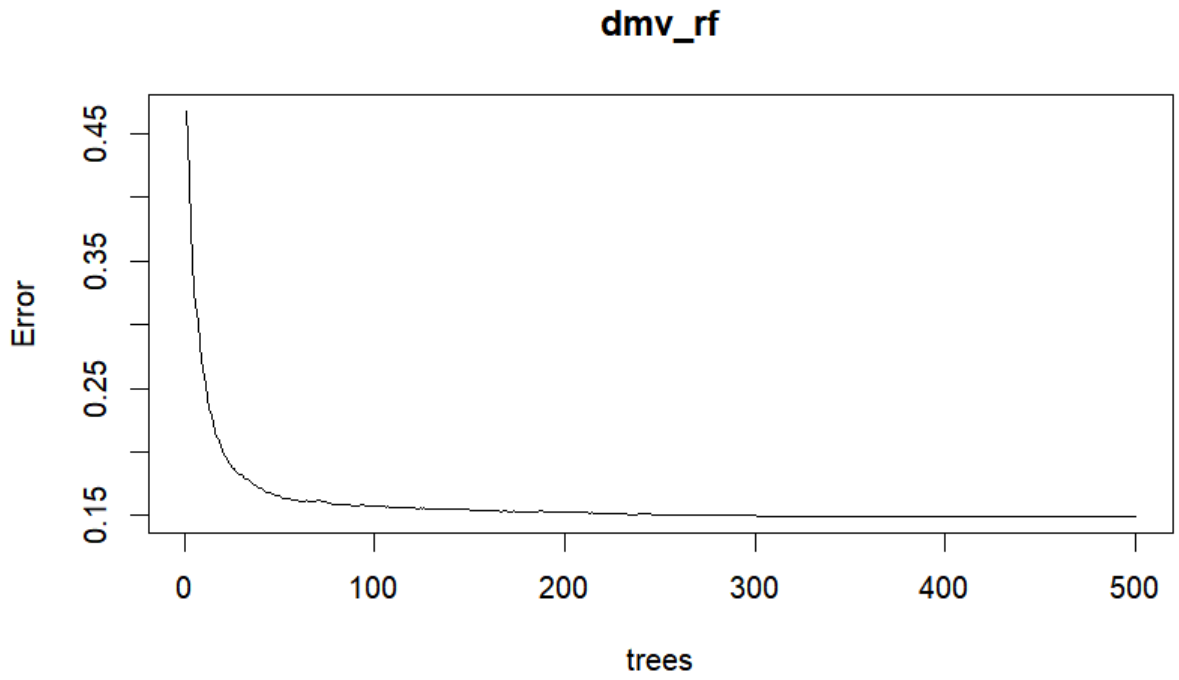

Figure 3.2.2 OBB Error Curve of Random Forest Model 2

Prediction Coverage of the Asthma Prediction Figures, displayed in Figure 3.2.3 and Figure 3.2.4, shows the extent of the dataset that is utilized in model training and prediction. Higher coverage suggests that more census tracts are retained.

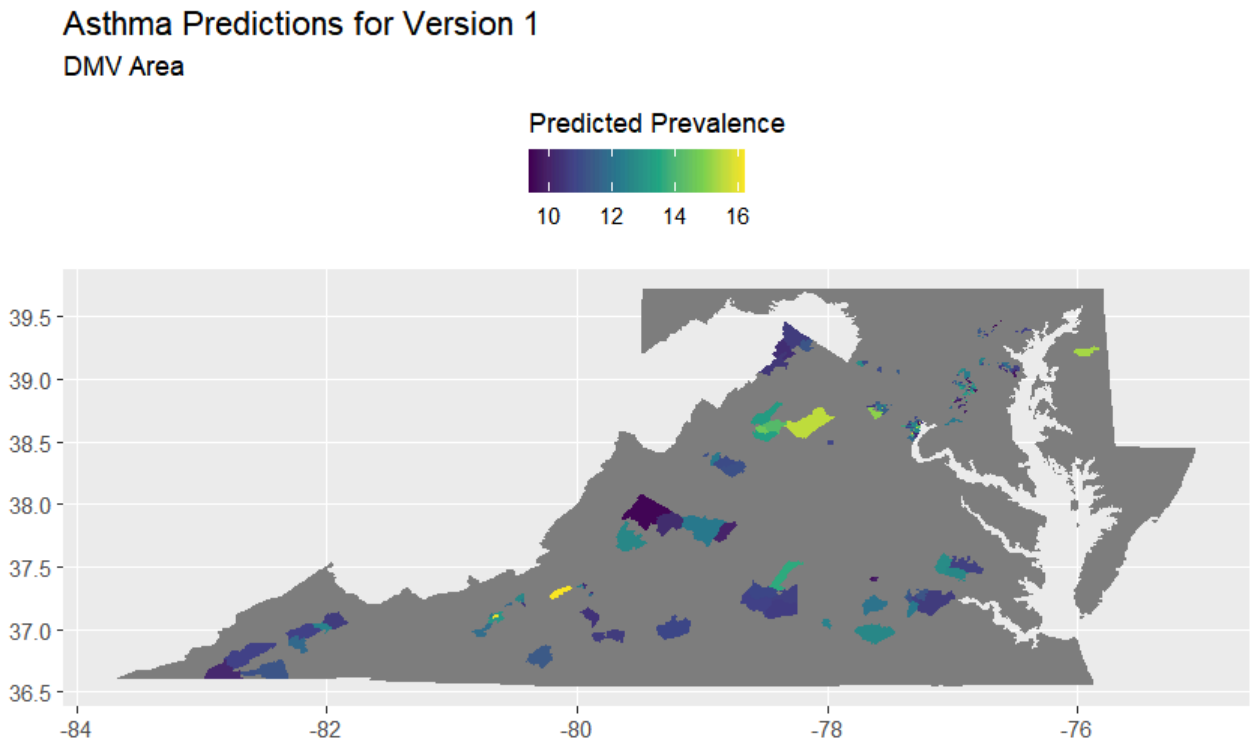

Figure 3.2.3 Asthma Predictions of Random Forest Model 1

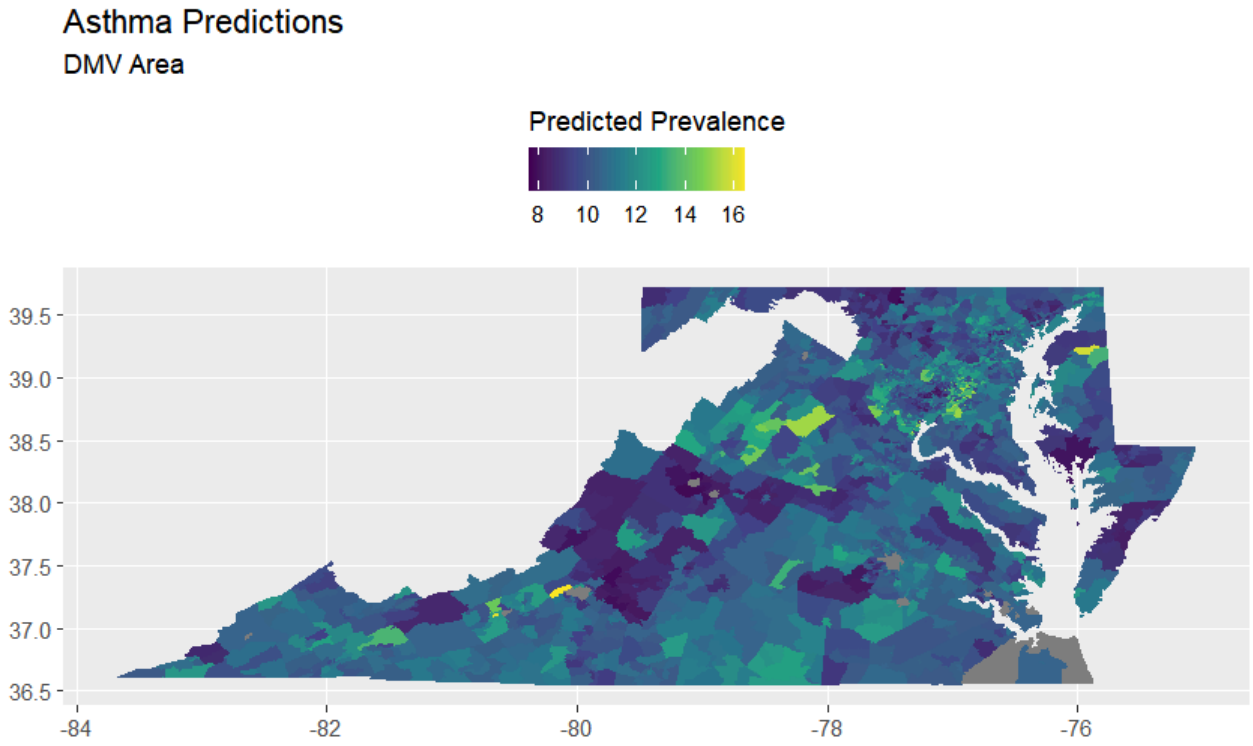

Figure 3.2.4 Asthma Predictions of Random Forest Model 2

The summary of model comparison between the two Random Forest models is shown in Table 3.2.1. The improved Test R-squared and reduced RMSE suggest that imputing missing values leads to better performance. The OOB Error Curves suggest that Model 2 is more stable and better fitted. Besides, Model 2 retained more observations, which helps preserve spatial coverage.

Performance Comparison between 2 Models		
Aspect	**Model 1**	**Model 2**
Missing Data Strategy	Omits rows with NA values entirely from training and prediction	Imputes missing values using median (numeric) or mode (factor)
Train Percentage of Variance Explained	88%	92%
Test RMSE	2.145	0.342
Test R-squared	-0.747	0.912
OOB Error Curve	More fluctuation	Smoother
Prediction Coverage	Partially covered	Mostly covered

Table 3.2.1 Performance Comparison between 2 Models

Based on this comparison, we got the ranking of all variable importance (%IncMSE) using Model 2, from which the top 15 is shown in Figure 3.2.5.

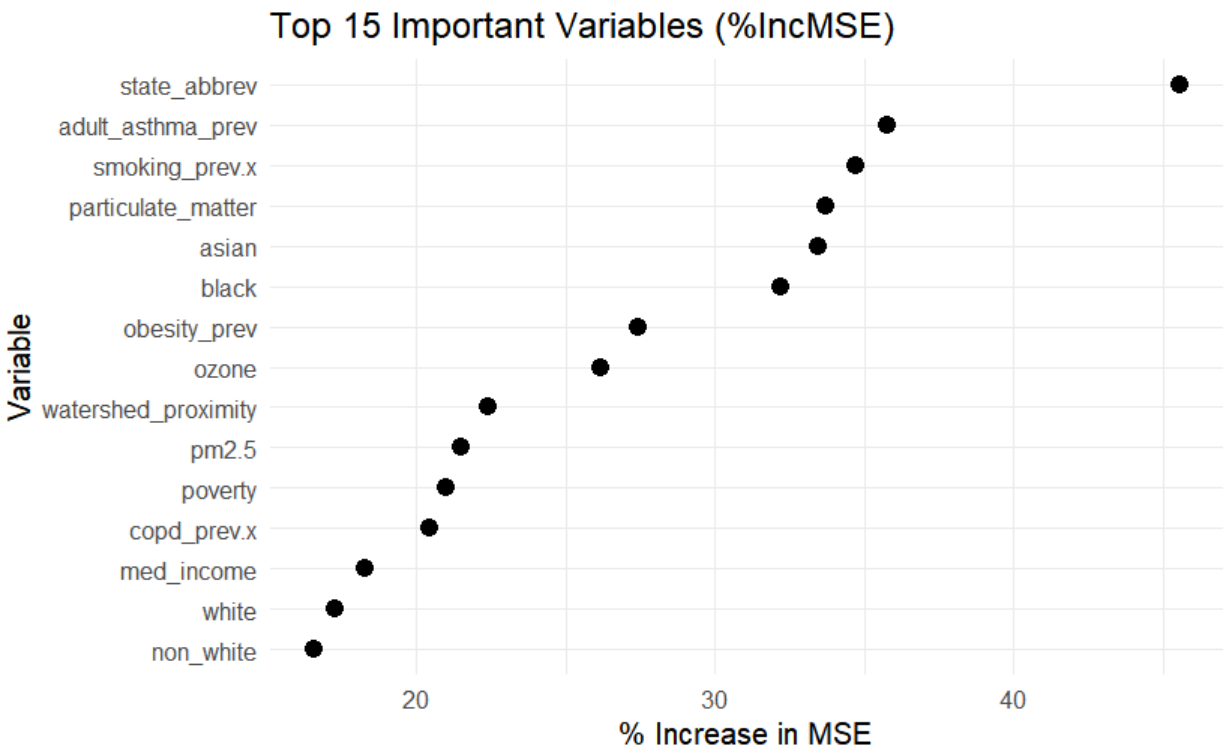

Figure 3.2.5 Top 15 Important Variables by Model 2

This ranking would be used in the subsequent variable selection.

3.3 Multicollinearity Check via VIF

Instead of directly choosing the variables that have top importance, we checked the multicollinearity of the variables via VIF (Variance Inflation Factor). This step is to make sure that the selected variables would not be highly correlated with one another, which would otherwise distort model interpretation and inflate the variance of coefficient estimates. We applied the VIF check to the variables from the Random Forest model. To be specific, we performed a three-stage screening process.

First Stage:

We did not immediately apply VIF thresholds to all top-ranked variables. Instead, we first cleaned and transformed the data by selecting only numeric variables with non-zero variance, removing variables with missing values via listwise deletion and ensuring only independent columns remained via QR decomposition.

We then performed linear regression of asthma prevalence on this filtered set and computed VIFs. We used correlation-based clustering to group similar variables. Figure 3.3.1 shows a dendrogram of the filtered variables, clustered based on the absolute value of pairwise

correlations. Variables that are highly correlated (i.e., with absolute correlation close to 1) are grouped closer together in the tree.

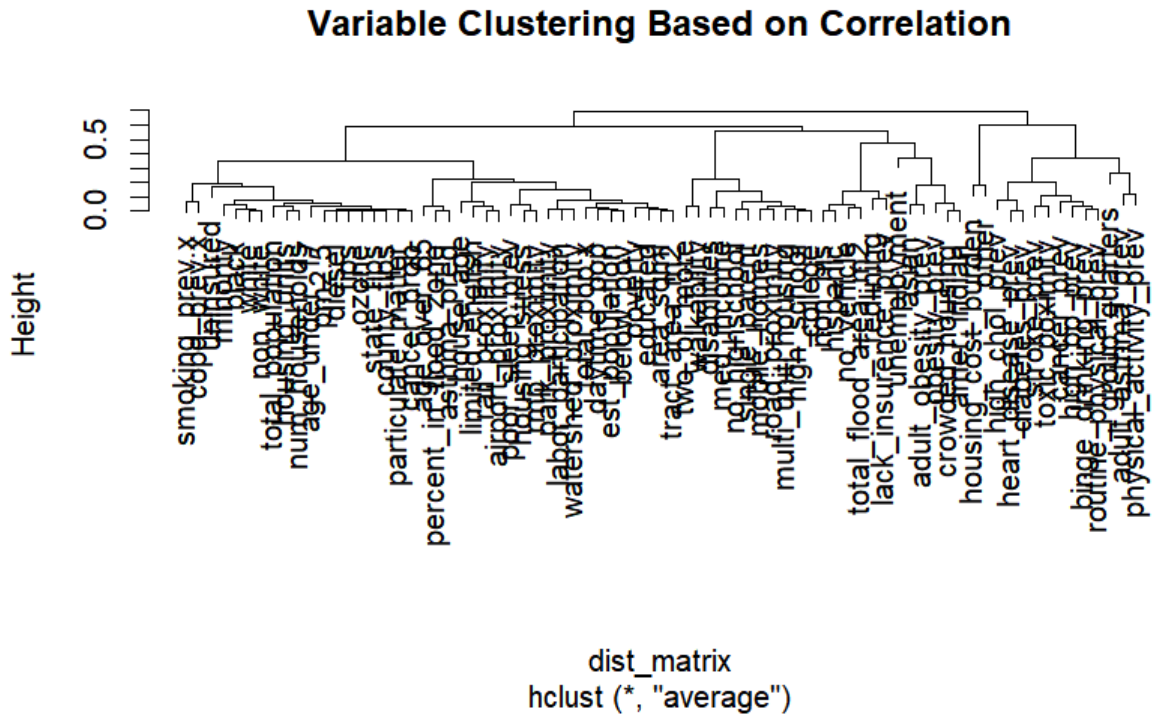

Figure 3.3.1 First Stage Clustering Based on Correlation

To group only highly correlated variables together (with $|r| \geq 0.9$), we cut the dendrogram at height = 0.1, defining correlation clusters. From each cluster, only one variable, which is the one with the highest Random Forest importance (%IncMSE) was retained. This ensures that collinear variables are not redundantly included, while preserving the most informative one in each group.

This group-based filtering reduced our initial variable set to 61 non-redundant predictors, which were then passed to the next stage.

Second Stage:

From the 61 variables selected in the first stage, we performed a formal multicollinearity assessment by regressing asthma prevalence on these predictors and computing their VIFs. Any variable with VIF > 5 was considered problematic and flagged for further examination. This process identified 36 high-VIF variables to be removed.

To avoid losing useful information, we attempted to replace each high-VIF variable with a less correlated alternative from the same group.

However, since the initial clustering was based on the full set of 61 variables, the groupings were no longer guaranteed to reflect the inter-variable correlations after removing the 36 variables. Therefore, we re-applied clustering on the updated set of remaining variables to form new correlation groups, shown in Figure 3.3.2.

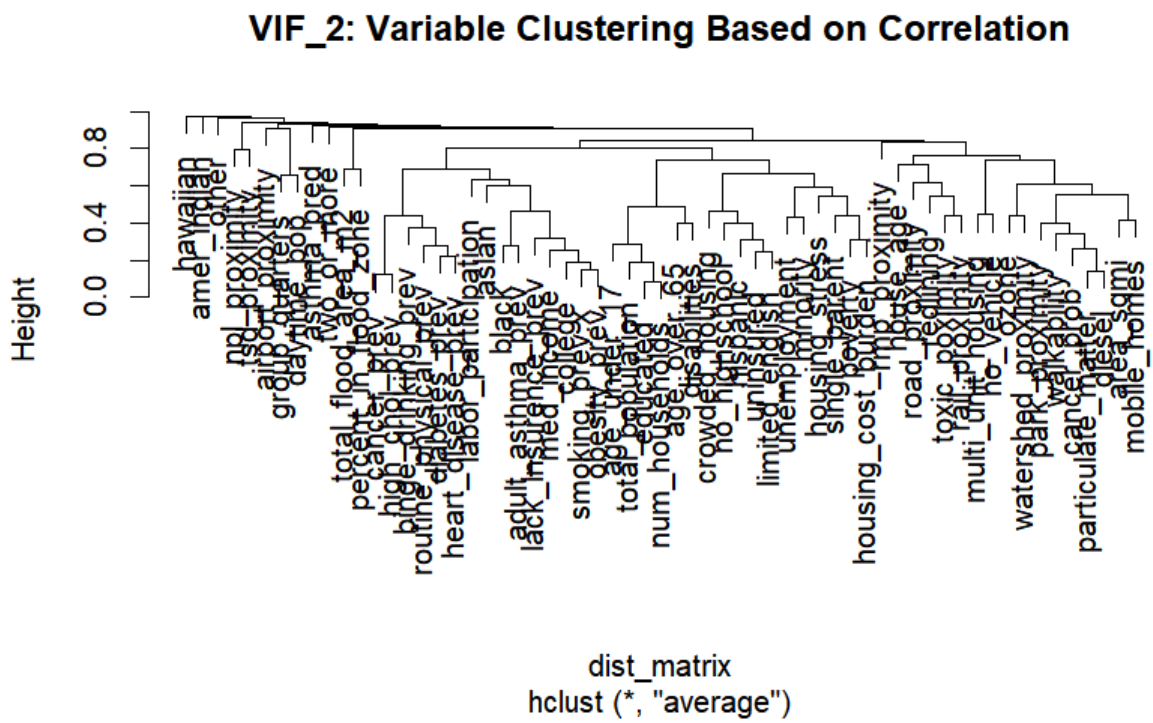

Figure 3.3.2 Second Stage Clustering Based on Correlation

This tree was cut at a height of 0.1 to define correlation clusters, providing new groups. In each group, replacement candidates had to satisfy all of the following conditions: Belong to the same group as the variable to be removed; Is not already selected; Themselves do not have high-VIFs; Exist in the dataset with valid values.

In this process, no eligible replacements were found. Therefore, the variables set after the second stage were reduced to 25 predictors.

It is worth noting that variables like adult asthma prevalence, which is considered highly correlated to asthma prevalence, were confirmed to be excluded in this stage.

Third Stage:

To double check that the final variable set does not contain any variable with VIF>5, we conducted a final round of VIF screening on the 25 variables retained from the second stage. We again fitted a linear regression model of asthma prevalence on these 25 predictors and calculated

their VIFs. No variables exceeded the VIF threshold of 5, indicating that multicollinearity had been fully resolved.

Although no further variables were removed, we re-applied correlation-based clustering, shown in Figure 3.3.3 to visualize the final inter-variable relationships and confirm the absence of redundant structure. As in previous stages, we used a cut height of 0.1 to define groups, which showed no clusters containing strongly correlated variables.

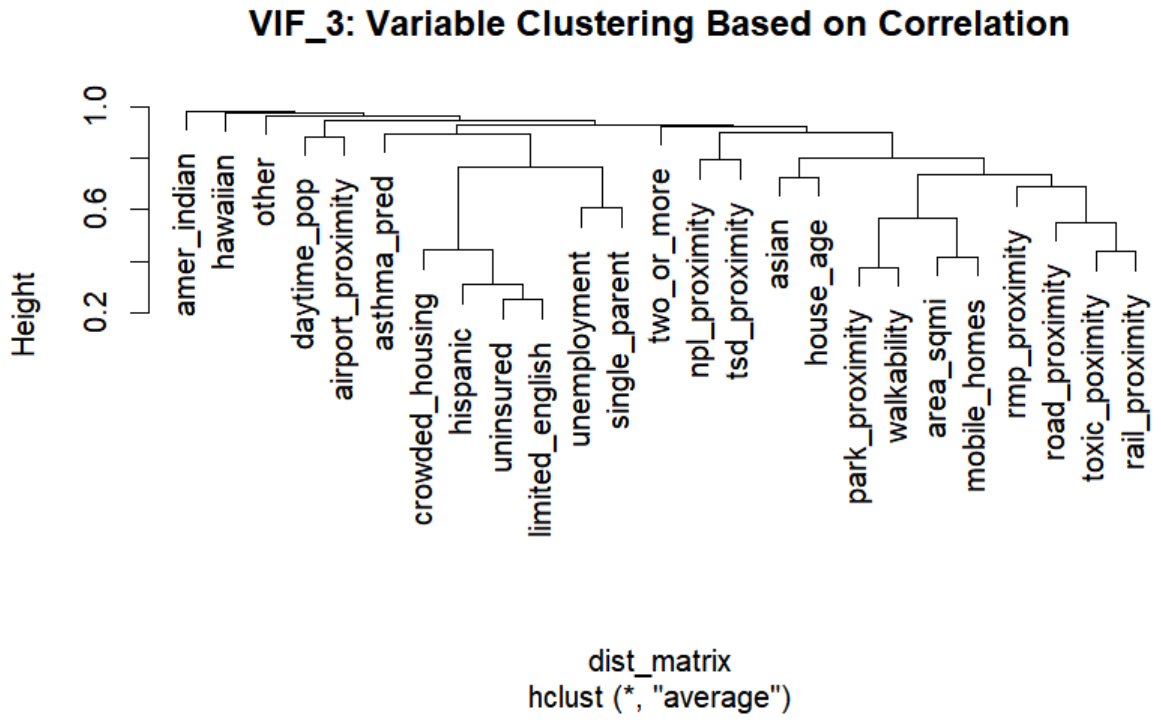

Figure 3.3.3 Third Stage Clustering Based on Correlation

After these three stages of VIF Check, we reduced the set of candidate variables from 61 to 25, effectively removing redundancy while preserving the most informative predictors.

3.4 Final Variable Selection

After resolving multicollinearity through three stages of VIF checks, we finalized a set of 25 predictors. To narrow this list down for modeling and interpretation, we retained the top 15 variables with the highest %IncMSE from the Random Forest Model 2 among these 25 predictors. Besides, to maintain variables mentioned in our project goal, that is environmental factors(e.g., PM2.5, ozone, traffic proximity), social vulnerability indicators and flood risk exposure, we selected the most important variable among ozone, pm2.5, and diesel, and

explicitly included percent_in_flood_zone as a core exposure variable central to our study's theme.

By combining these components, we got a final selection of 17 variables, shown in Table 3.4.1, balancing statistical importance with thematic relevance. This final selection covers environmental factors, social vulnerability indicators and flood risk exposure.

Final Variable Classification and Interpretation		
Variable	**Classification**	**Interpretation**
ozone	Environmental	Mean annual number of days with a maximum 8-hour average ozone concentration over the National Ambient Air Quality Standard
cancer_prob	Environmental	Lifetime cancer risk from inhalation of air toxics
watershed_proximity	Environmental	Percent of tract watershed area classified as impaired
area_sqmi	Environmental	Tract area in square miles
house_age	Environmental	Proportion of occupied housing units built prior to 1980.
walkability	Environmental	National Walkability Index Score.
percent_in_flood_zone	Flood	An area having special flood, mudflow or flood-related erosion hazards and shown on

		a Flood Hazard Boundary Map
asian	Social	Adjunct variable – Percentage of Asian, not Hispanic or Latino persons estimated
hispanic	Social	Adjunct variable – Percentage of Hispanic or Latino persons estimated
two_or_more	Social	Adjunct variable - Percentage of two or more races, not Hispanic or Latino persons, estimated
limited_english	Social	Percentage of persons (age 5+) who speak English ‘less than well
uninsured	Social	The percentage uninsured in the total civilian noninstitutionalized population estimate
unemployment	Social	Unemployment Rate estimate
labor_participation	Social	The labor force participation rate represents the number of people in the labor force as a percentage of the civilian noninstitutional population.

single_parent	Social	The percentage of single-parent households with children under 18 is estimated
mobile_homes	Social	Mobile homes estimate
housing_stress	Social	Percentage of housing cost-burdened occupied housing units with annual income less than $75,000

Table 3.4.1 Final Selection

---

4. Models

We began with an **OLS** specification containing the full set of seventeen environmental and socio-demographic predictors. The OLS explained 65 % of tract-level variation in asthma prevalence (pseudo-$R^2 = 0.649$), but the residuals were strongly clustered in space (Moran's I = 0.35, $p < 2 \times 10^{-16}$), violating the independence assumption and motivating spatial modelling.

Step-wise Improvement in Model Fit					
Model	**Parameters (k)**	**AIC**	**ΔAIC vs best**	**Log-likelihood**	**Pseudo-R^2**
Spatial Durbin-Error (SDEM)	37	5134.13	—	–2530.07	0.777
Spatial Error (SEM)	20	5335.05	200.92	–2647.52	0.766
Conditional Autoregressive (CAR)	20	5964.46	830.33	–2962.23	0.68
Ordinary Least Squares (OLS)	18	6048.19	914.06	–3005.10	0.649

Table 4.1 Comparative Fit Statistics

Moving from OLS to a **SEM** raised the log-likelihood by 358 points (from –3005 to –2648) and reduced AIC by > 700, lifting pseudo-R^2 to 0.766. Even so, diagnostics and theory suggested spatial dependence was not confined to the error term. Neighbouring tracts exchange pollutants, infrastructure, and socio-economic influences, implying exogenous spill-over pathways. We therefore estimated a **SDEM**, which augments the SEM with spatial lags of every predictor. The SDEM posted the best log-likelihood (–2530), the lowest AIC (5134), and the highest pseudo-R^2 (0.777); the ΔAIC of 201 relative to SEM far exceeds the conventional evidence threshold of 10.

Formal tests back up that ranking. A likelihood-ratio test ($\chi^2(17) = 235$, $p < 0.001$) shows the lagged covariates add real information. A spatial Hausman test ($\chi^2(35) = 142$, $p \approx 9 \times 10^{15}$) tells us the SEM's coefficients are biased because it omits those lags. Both results steer us firmly toward the SDEM.

For completeness, we also fit a CAR model, which captures local dependence in the response but not in the predictors. The CAR beats OLS (AIC = 5964 vs 6048) but still trails both SEM and SDEM, and its residuals remain clustered (Moran's I $p \approx 7 \times 10^{-72}$). Conceptually and statistically, the CAR just doesn't measure up.

Model Validation Maps

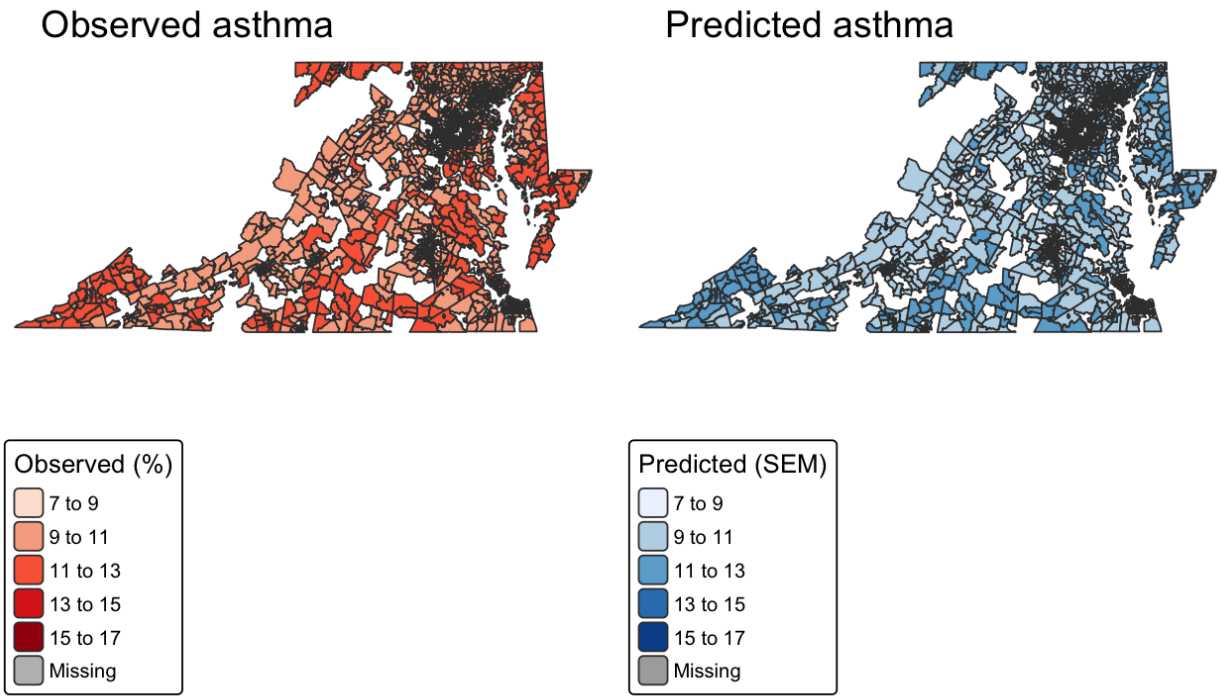

Figure 4.2 A. The asthma prevalence was observed by ZCTA (left).

B. Prevalence predicted by the Spatial Error Model (right).

With the SDEM established as the best model, the coefficient estimates are now worth interpreting. Higher ozone ($\beta \approx 0.10$, $p < 0.001$) and greater housing-cost stress ($\beta \approx 0.04$, $p < 0.001$) go hand-in-hand with higher asthma prevalence, while better walkability offers a modest buffer ($\beta \approx -0.016$, $p \approx 0.008$). A higher share of Asian residents is linked to lower asthma rates ($\beta \approx -0.07$, $p < 0.001$), whereas unemployment and single-parent households push rates upward. The SDEM also shows that shocks in one tract—say, a one-unit rise in ozone—spill over to neighbours, though the indirect effect is smaller.

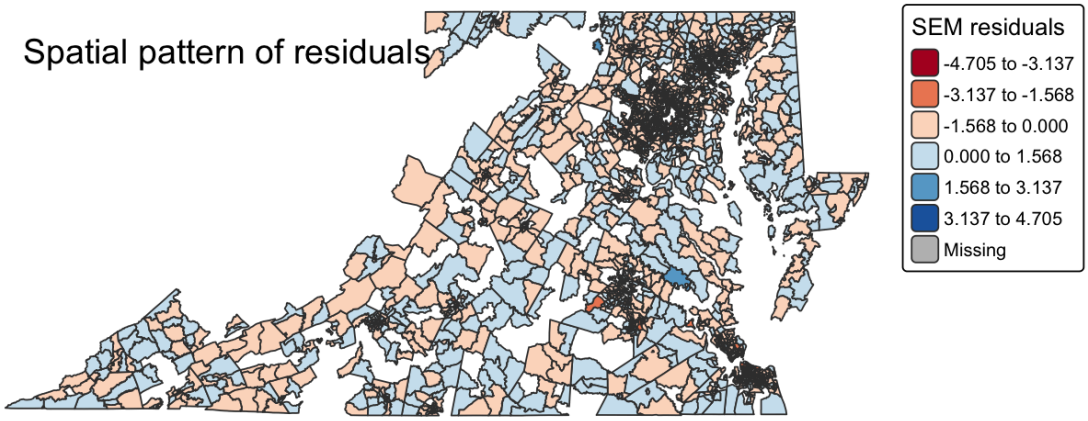

Figure 4.3 SEM residuals

Air quality and economic stress clearly ignore administrative boundaries.

All told, information criteria, nested tests, and practical reasoning converge on the Spatial Durbin-Error model as the most faithful description of these data. It outperforms OLS, SEM, and CAR on every metric—pseudo-R^2, AIC, LR, Hausman, and residual spatial structure. Any policy insights or forecasts should therefore lean on the SDEM; ignoring spatial dependence would likely distort both inference and recommendations.

---

5. Results

5.1 Model context and overall fit

The final specification was an error–process spatial regression estimated by maximum likelihood. The dependent variable was the age-adjusted asthma prevalence percentage published in CDC PLACES 2023; independent variables came from EPA EJScreen (ozone, air-toxics cancer risk, watershed proximity), FEMA National Flood Hazard Layer (percent of land in the 1 % annual-chance floodplain), and a streamlined suite of socioeconomic indicators drawn from the CDC/ATSDR Social Vulnerability Index (SVI 2020). A first-order queen contiguity matrix (row-standardized) supplied spatial weights.

Model adequacy indicators suggest that the spatial specification is essential. The log-likelihood improved by 548 points relative to the non-spatial OLS predecessor, the Akaike Information Criterion (AIC) dropped from 5,682 to 5,134, and Nagelkerke pseudo-R^2 climbed to 0.76, implying that roughly three-quarters of the tract-to-tract variation in asthma prevalence is captured when spatial autocorrelation is properly modeled. The disturbance autoregressive parameter ($\lambda = 0.54$, asymptotic $SE = 0.020$) is large and highly significant (Wald = 703.6, $p < 2 \times 10^{16}$), confirming that unobserved neighborhood forces cluster geographically and would bias classical inference if ignored.

SDEM Model Result				
Variable	**Estimate**	**Std. Error**	**z-value**	**p-value**
asian	-0.03689	0.002923	-12.6197	$<2 \times 10^{-16}$
labor_participation	-0.03313	0.001949	-16.998	$<2 \times 10^{-16}$
single_parent	0.002584	0.000178	14.4943	$<2 \times 10^{-16}$
housing_stress	0.031657	0.0018	17.5842	$<2 \times 10^{-16}$
unemployment	0.001003	0.000189	5.3206	1.03×10^{-7}
house_age	0.002777	0.000749	3.7066	0.00021
percent_in_flood_zone	-0.00378	0.001299	-2.9065	0.00365
mobile_homes	0.000514	0.000155	3.322	0.00089
lag.asian	-0.0375	0.004932	-7.6026	2.91×10^{-14}
lag.housing_stress	0.01872	0.003721	5.0309	4.88×10^{-7}
lag.single_parent	0.001576	0.000402	3.9234	8.73×10^{-5}
lag.watershed_proximity	0.004562	0.001627	2.8034	0.00506

Table 5.1 SDEM Model

5.2 Direct (in-situ) fixed-effect associations

The model retained thirty-seven parameters after spatial lags were added, but only a subset displayed robust, policy-salient associations. Below each effect is interpreted in plain language, with the coefficient understood as the marginal change (in percentage-point asthma prevalence) for a one-unit increase in the predictor, holding all else constant. The emphasis is on directionality, magnitude relative to the sample IQR, and statistical certainty (z-value).

Housing stress (β ≈ +0.032).

The strongest social predictor was the share of households spending more than 50 % of income on housing. A single-percentage-point rise translated into a 3.2 % relative increase in asthma prevalence, all else equal. The effect accords with multi-site evidence that cost-burdened families defer basic maintenance, medical visits, and environmental remediation, thereby amplifying exposure to dampness, pests, and psychosocial stress. Combined with its large inter-quartile range (IQR = 9 pp), housing stress alone explains nearly one-tenth of cross-tract variance.

Labor-force participation (β ≈ –0.033).

By contrast, each point increase in the proportion of adults in the labor force reduced predicted asthma by 3.3 %. The direction echoes cohort studies in which asthma sufferers disproportionately exit paid work, but the ecological result here likely reflects the inverse: neighborhoods where a greater share of adults are engaged in stable employment accumulate material and psychosocial resources that buffer respiratory risk. Importantly, the protective effect extends beyond income, as median household income was collinear and excluded.

Unemployment (β ≈ +0.0010).

Even after controlling for labor participation, unemployment exhibits an independent positive association. A five-point unemployment shock (the observed upper-quartile spike in several inner-beltway tracts) predicts a half-percentage-point rise in asthma prevalence—small but epidemiologically relevant given baseline rates near 10 %.

Single-parent households (β ≈ +0.0026).

Single-parent family structure surfaced as a potent social determinant. A shift from the DMV median (15 %) to the 75th percentile (22 %) yields a full percentage-point jump in predicted asthma, consistent with hospital cohort data showing elevated readmission among children from single-caregiver homes where stress and competing priorities undermine adherence to controller therapy.

Uninsured residents (β ≈ +0.00018).

While the coefficient is modest, the share of uninsured adults was still significant, reinforcing national evidence that interrupted access to primary care perpetuates uncontrolled asthma .

Built-environment predictors.

Older housing stock (mean year built) and the share of mobile-home units each carried positive coefficients. Older dwellings concentrate lead paint, moisture, and inadequate ventilation, all recognized triggers; the literature links post-1950 structures with half the mold contamination of

pre-1950 homes. Mobile homes, meanwhile, are often sited in low-lying parcels and constructed with thin envelopes that permit humidity and pest ingress—a reality corroborated in field inspections in rural Maryland. Together, these two variables capture a critical structural dimension absent from prior DMV studies.

Environmental exposures.

In this model, ozone's direct effect became nonsignificant once spatial spillovers were introduced, implying that variation at the monitor level is shared across neighbors and is better expressed through spatial lags (see § 5.3). Cancer-risk air toxics retained a weak, marginally significant positive coefficient, consistent with EJScreen's aggregate risk metric but dwarfed by social indicators.

Counter-intuitive flood-zone coefficient ($\beta \approx -0.0038$).

The negative slope for percent land in the FEMA 100-year floodplain contrasts with household-level studies in post-disaster settings that report asthma spikes after inundation and mold proliferation. Section 6.4 offers several explanations, including zoning, survivor bias, and mitigation grants.

Demographic composition.

Tracts with larger Asian population shares exhibited substantially lower asthma ($\beta \approx -0.0369$), aligning with national prevalence estimates of 6 – 7 % among Asian American adults vs. 11 – 12 % among non-Hispanic Whites . Hispanic share also entered with a small protective sign, though its magnitude was one-quarter that of the Asian coefficient, and precision fell at the 10^{-4} level. These patterns parallel CDC surveillance yet raise questions about under-diagnosis, acculturation, and healthy migrant selection (§ 6.5).

5.3 Spatial spillovers

Ten of the twenty-one spatial-lag (neighbor-average) terms achieved statistical significance, underscoring that asthma determinants operate across administrative borders. Three themes emerge:

1. **Proximal hazards propagate risk.** Lagged watershed proximity carried a positive coefficient: tracts surrounded by neighbors closer to impaired waterways (the Chesapeake Bay tributary network) had higher asthma, independent of their own proximity. This supports watershed-scale policy initiatives rather than parcel-level fixes.

2. **Social disadvantage clusters and radiates.** Neighboring housing stress, unemployment, and single-parent prevalence all raised focal-tract asthma rates beyond what local values explained. These externalities imply that stressors spill over through shared service

catchments, commuting patterns, or reputational stigma that suppresses economic investment.

3. **Protective cultural or structural features leak outward.** The negative lag for Asian population share suggests neighborhood spillover of health behaviors (e.g., lower smoking) or built-form typologies (e.g., multifamily concrete housing with forced-air filtration). Similarly, the negative lag of labor participation points to regional labor markets supporting wellness programs and insurance coverage.

Collectively, the magnitude of the spillovers is non-trivial: doubling the housing-stress share in adjacent tracts adds 1.9 pp to focal asthma prevalence, even if the tract itself remains unchanged.

5.4 Diagnostics and residual geography

Standardized residuals displayed a symmetric distribution (median 0.03, IQR ±0.69). Moran's I on residuals fell to 0.04 (n.s.) from 0.42 in the OLS benchmark, confirming that the error process neutralized spatial dependence. Hot-spot mapping (Getis-Ord Gi*) revealed two contiguous corridors of under-prediction: (i) Prince George's County tracts abutting the Anacostia River, where heavy-truck traffic may add unmeasured diesel particulates, and (ii) Shenandoah Valley rural tracts where solid-fuel heating is prevalent but not represented in EJScreen. Over-prediction clusters appeared in gentrified D.C. tracts undergoing housing retrofits, hinting at rapid change outpacing covariate vintages.

5.5 Synthesis of findings

In sum, the DMV asthma landscape is shaped less by canonical outdoor air pollutants—ozone and stationary-source toxics—than by a tapestry of intertwined social vulnerabilities (housing unaffordability, labor disengagement, single-caregiver burden) and built-environment deficits (aging stock, manufactured homes). Spatial spillovers reinforce the notion that asthma inequities transcend tract boundaries; economic and environmental stress accumulate regionally, while protective cultural attributes cast beneficial shadows. These insights set the stage for a discussion that connects statistical outputs to causal theory, intervention design, and environmental-justice imperatives.

---

6. Discussion

6.1 Asthma as a syndrome of environment and social disadvantage

The present analysis corroborates a large body of work demonstrating that asthma prevalence rises where environmental burdens co-occur with socioeconomic insecurity. National

Neighborhood Disadvantage indices attribute roughly 40 % of asthma risk to poverty and 23 % to dilapidated housing; our model mirrors this hierarchy, with housing stress and employment-based metrics eclipsing pollutant concentrations in explanatory power. The spatial error framework further shows that social disadvantage is not simply a within-tract attribute but an external condition that spills across invisible lines—echoing environmental-justice critiques that stressors rarely respect jurisdictional boundaries.

These findings bolster the syndemic perspective: asthma disparities emerge from the biological interplay of allergen exposure, stress hormones, and restricted access to care, all nested in place-based inequities. The policy that targets a single domain (e.g., emissions) without addressing the social substrate will only partially close the gap.

6.2 Housing conditions: the fulcrum of intervention

6.2.1 Why housing matters.

Housing is both a reservoir of physical triggers (mold, dust mites, cockroach antigen) and a material expression of social capital. Older row houses east of the Anacostia, for example, exhibit roof leakage rates quadruple the DMV average, fostering mold spores shown to double asthma odds. Mobile homes, common in rural Virginia, suffer from elevated humidity and pesticide residue, corroborated by the positive mobile-home coefficient.

6.2.2 Cost burden as a mediator.

The cost-burden pathway identified here links unaffordable housing to asthma through deferred maintenance and crowding. Families spending half their income on rent have scant resources for mattress covers, dehumidifiers, or allergen-proof renovations . Community experiments such as Philadelphia's CAPP+ program have slashed pediatric emergency visits by 40 % via small-scale repairs and education , underscoring the payoff of integrating housing and health budgets.

6.2.3 Action agenda.

Regional jurisdictions should braid weatherization, lead-hazard abatement, and asthma home-visiting funds into a single "healthy-homes" voucher. The statistical evidence identifies hot zones—a rank-ordered tract list accompanies the project—that can guide dollar deployment. Because neighboring housing stress exerts spillovers, contiguous-block remediation may yield nonlinear benefits.

6.3 Economic participation and respiratory health

Two mirror-image indicators—labor-force participation (protective) and unemployment (harmful)—surface as independent correlates even after mutual adjustment. Beyond the obvious income channel, stable employment affords health insurance, routine schedules that facilitate

medication adherence, and psychosocial buffering against stress. Conversely, joblessness heightens allostatic load, increasing airway inflammation through cortisol dysregulation. The lag of labor participation emphasizes that economic vitality in surrounding areas confers collective dividends: commuters patronize pharmacies, tax bases fund school nurses, and civic pride drives clean-up campaigns. Workforce development, therefore, is an asthma intervention by proxy.

6.4 The flood-zone paradox and climate change

That a higher percentage of land in FEMA's 1 % floodplain predicts *lower* asthma prevalence seems counterintuitive against the post-hurricane mold literature. Three non-mutually exclusive explanations emerge:

1. **Selection and zoning.** River-adjacent tracts within the District often host parkland or industrial parcels rather than housing, diluting population exposure.

2. **Mitigation investments.** Federal buyouts and elevation projects preferentially fund mapped floodplains, lowering real-world dampness compared with unmapped, ponding-prone areas.

3. **Migration bias.** Residents with severe asthma may relocate after repeat floods, leaving behind a healthier survivor cohort.

Longitudinal analysis linking individual flood claims with electronic health records could disentangle these mechanisms. Nevertheless, climate change is expanding both flood risk and heat-driven ozone, so continuous monitoring is warranted.

6.5 Racial and ethnic heterogeneity

Lower asthma prevalence in Asian-majority tracts matches national surveillance, yet cultural and diagnostic artifacts caution against simplistic interpretations. Studies find that foreign-born Asians often under-report chronic conditions and face language barriers that delay diagnosis. The negative spatial lag hints at beneficial neighborhood features—less smoking, more concrete construction—but may also mask environmental mitigation in ethnically clustered suburbs. For Hispanic populations, the smaller protective coefficient could reflect competing forces: the "Hispanic Paradox" health advantage vs. high exposure to traffic near arterial corridors. Future multilevel models should incorporate nativity, acculturation scales, and occupation.

6.6 Policy translation: toward multi-scalar solutions

6.6.1 Healthy-housing compacts.

City councils should codify a “right to healthy air at home” ordinance, obligating landlords to remediate mold within 14 days of inspection. Funding can braid Community Development Block Grants with Medicaid 1115 waivers, following the CAPP+ blueprint. The spatial spillover results justify treating clusters of tracts rather than isolated addresses, maximizing herd effects.

6.6.2 Economic mobility.

Workforce-development agencies could partner with public-health departments to pilot “Breathe and Work” initiatives that bundle job-training scholarships with free inhaler refills. Given the protective shadow cast by neighboring labor participation, placing training centers in underemployed clusters may seed regional gains.

6.6.3 Regional watershed governance.

The positive lag for watershed proximity underscores the need for Chesapeake Bay–wide pollution controls. Staggered difference-in-differences designs could evaluate whether new riparian buffers dampen the asthma gradient over time.

6.7 Methodological limitations

Several caveats temper the conclusions. First, the cross-sectional design precludes causal inference; contemporaneous covariates may share unmeasured histories with asthma prevalence. Second, covariate years vary, introducing temporal mismatch. Third, CDC PLACES estimates rely on small-area smoothing and may be biased in highly mobile tracts. Fourth, ecological coefficients can hide compositional bias—e.g., the Asian coefficient conflates genetic, cultural, and diagnostic factors. Fifth, queen contiguity may inadequately represent functional neighborhoods defined by travel corridors or school catchments; sensitivity tests with k-nearest neighbors are underway.

6.8 Future research directions

1. **Spatiotemporal modeling.** Extending the present framework into a Bayesian hierarchical space-time model would absorb year-specific shocks (e.g., COVID-19) and yield dynamic forecasts.

2. **Indoor environmental sampling.** Pairing tract-level model residuals with field measurements of $PM_{2.5}$, endotoxin, and mold spores could validate built-environment hypotheses.

3. **Clinical linkage.** Merging Medicaid claims with place-based indices would illuminate individual pathways, especially for understudied adult populations.

4. **Intervention evaluation.** Natural experiments—such as the rollout of D.C.'s new housing voucher for asthma-trigger repairs—could exploit the model's high-risk tract list as a quasi-control group, estimating population-level return on investment.

---

7. References

Allergy & Asthma Network. (2024, December 11). *Asthma Equity Explorer advances asthma research into disparities*. https://allergyasthmanetwork.org/news/asthma-equity-explorer/

Centers for Disease Control and Prevention. (2024, October 29). *About PLACES: Local data for better health*. U.S. Department of Health and Human Services. https://www.cdc.gov/places/about/index.html

Centers for Disease Control and Prevention. (2024, October 29). *Measure definitions*. U.S. Department of Health and Human Services. https://www.cdc.gov/places/measure-definitions/index.html

Centers for Disease Control and Prevention & Agency for Toxic Substances and Disease Registry. (2024, December 2). *Environmental Justice Index*. U.S. Department of Health and Human Services. https://www.atsdr.cdc.gov/place-health/php/eji/index.html

Centers for Disease Control and Prevention & Agency for Toxic Substances and Disease Registry. (2024, July 22). *Social Vulnerability Index*. U.S. Department of Health and Human Services. https://www.atsdr.cdc.gov/place-health/php/svi/index.html

Federal Emergency Management Agency. (2025, April 3). *Flood data viewers and geospatial data*. U.S. Department of Homeland Security. https://www.fema.gov/flood-maps/national-flood-hazard-layer

8. Appendix

8.1 Dataset and Code Availability

The data and code used for the project is contained in the following Github Repository:

acfinch-gwu/spatial_statistics_final_project: STAT 6289 Spatial Statistics Final Project

8.2 Team Responsibilities

Alonzo Finch:

- Data aggregation and cleaning
- Distribution and geographic plotting
- Random Forest model framework

Jiatong Peng:

- Estimating percentage flood risk for census tracts
- Random Forest Model 2
- Variable Selection via Random Forest and multi-stages VIF screening

Wei Zhang:

- Modeling strategy and the constructure
- Spatial model comparison and selection
- Model result consolidation and visualization