Assessing environmental injustice in suburban industrial sectors:
A case study of Northwest Indiana and New Jersey

Abstract

This study attempts to evaluate whether environmental injustice is occurring by examining environmental and demographic factors' effects on health outcomes in Northwest Indiana (NWI) and New Jersey (NJ) using data from the CDC’s Environmental Justice Index. Four health conditions—high blood pressure, asthma, diabetes, and cancer—were modeled at the census-tract–level using stepwise regression and refined for interpretability. Results show that socioeconomic factors, including minority population percentage, poverty rates, and age, are the strongest predictors of health outcomes, while environmental exposure variables are less significant. Models retain robust predictive power when tested on neighboring regions, and comparisons with null models containing only pollution variables confirm that socioeconomic factors do play a critical role in driving health disparities. While this study underscores the need for targeted interventions in vulnerable communities, limitations in census data complicate the assessment for environmental injustice.

Introduction

Environmental injustice considers that marginalized communities are disproportionately affected by environmental harms such as pollution and industrial waste. While occurring in a wide variety of areas, this study focuses on two suburban industrial areas: Northwest Indiana and New Jersey. Northwest Indiana (NWI) comprises Lake, Porter, and LaPorte counties, and is an overflow of the Chicago metropolitan area into the Hoosier state. Proximity to the city and Lake Michigan has made NWI a hub for industry, including ⅜ of the integrated steel mills in the US, the largest oil refinery in the Midwest, two ports, and a plethora of manufacturing. These industries greatly pollute the region, causing adverse health effects in the population. New Jersey faces similar challenges. It faces severe environmental challenges, with industrial pollution affecting vulnerable communities like the Ramapough Lenape Nation in Ringwood, who suffer from toxic waste dumping, and Newark's Ironbound District, where air and water pollution harm immigrant populations. This project attempts to characterize whether environmental injustice is statistically detectable by examining how environmental and demographic factors influence health outcomes in NWI and NJ using census tract data from the CDC Environmental Justice Index dataset. Four health metrics—high blood pressure, asthma, diabetes, and cancer rates—are modeled in both regions using stepwise regression and refined into smaller, more interpretable models. These smaller models retain strong predictive power while reducing complexity, even when applied to other regions.

Methods: Data and Analysis

The Center for Disease Control and Prevention (CDC) Environmental Justice Index (EJI) is a national geospatial tool for evaluating cumulative impacts of environmental factors on health and equity in the United States. The dataset categorizes data into three general categories: Social Vulnerability, Environmental Burden, and Health Vulnerability, which are used to calculate an EJI for each census tract within the United States. The dataset contains 118 variables, and its observational unit is a census tract, totalling about 80,000 across the US.

Four sections of the dataset were used for each region studied herein: Northwest Indiana (nwi_data) contains Lake, Porter, and LaPorte counties in Indiana, totaling 177 census tracts; New Jersey (NJ_data) includes the entire state, containing 2004 tracts; Cook County, Illinois (chi_data) includes the city of Chiago and a few surrounding suburbs, containing 1317 census tracts; and the Philadelphia metropolitan area (phil_data) includes Lehigh, Northampton, Bucks, Montgomery, Chester, and Philadelphia counties, totaling at 989 tracts. Both "nwi_data" and "NJ_data" will train our models and be used to evaluate environmental injustice. The datasets "chi_data" and "phil_data" serve as test datasets which will be used to determine the accuracy of our models. For each region, the full dataset was parsed to remove unwanted variables. Within the dataset, data collected is represented in raw (E), percentage (EP), and percentage rank (EPL) formats. Due to this repetition, only raw and percentage forms of the data were kept. This was performed for all datasets before the analysis began.

We recognize that this dataset is collected as census data, and therefore our findings are highly biased via the protocols used to collect census data. For example, census tracts within our dataset did not report data for a large majority of the variables, likely due to no permanent population existing in these tracts. For this reason, these tracts were removed from the datasets to avoid coding errors. Interestingly, these tracts could have a particular influence on our project if data was collected. For example, two tracts from "nwi_data" correspond to the Port of Indiana and the surrounding steel plants, which have a significant daytime population who are exposed to harmful pollutants.

Models for four response variables (EP_BPHIGH, EP_ASTHMA, EP_DIABETES, and EP_CANCER) were first generated using automated stepwise linear regression in both directions using the Bayesian Information Criterion (BIC) for selection. When each model was generated, the remaining response variables were excluded to remove the influence of comorbid conditions from our models. Despite high predictive accuracy, these *stepwise models* were minimized to simplify and improve interpretation by removing their least impactful variables (smallest magnitude), creating *final models*. Final models were evaluated and compared to stepwise models using residual plots, added variable plots, variance inflation factors, and spatial residual maps to ensure predictive accuracy was conserved while improving interpretation.

A *null model* was developed which predicts the response variable via four pollution variables (E_OZONE, E_PM, E_DSLPM, E_TOTCR). This null hypothesis states that there is no correlation between health outcomes and any socioeconomic variables in our dataset, only direct pollution exposure. The Akaike Information Criterion (AIC) and BIC of each final model were compared to those of the corresponding null model to ensure the final model improved from the null. Both the AIC and BIC were compared to ensure heavy penalization of complex models when using BIC did not inflate the more complex model (increasing the likelihood of type I and II errors).

Finally, the predictor variables from each final model were used to predict the response variable in a new region, creating *test models*. Final models trained on “nwi_data” were remodeled on “chi_data” and models trained on “NJ_data” were remodeled on “phil_data”. Methods for comparison were repeated. Residual plots, added variable plots, variance inflation factors, and spatial residual maps of the final model were compared between the training dataset (NWI or NJ) and the test dataset (Chicago or Philadelphia, respectively). Finally, these models were also compared to the null model (using chi_data or phil_data) using AIC and BIC.

Results

In total, eight models were produced to predict the prevalence of four health conditions in two separate locations: New Jersey and Northwest Indiana (Fig 1A). Surprisingly, all final models only rely on seven socioeconomic variables and no environmental variables for all predictions (Fig 1B). No R^2 values were lower than 0.6 after model reduction (the majority of the models had $R^2 > 0.75$), suggesting our models are able to predict adverse health outcomes. (Fig 1C). These models were simplified from initial stepwise selections by selecting those with the largest coefficients, while ensuring diagnostic assumptions, such as linearity and homoscedasticity, were met. Spatial residual maps were utilized to confirm that there is an even distribution of the residuals across geographic areas (see Appendix 1).

For example, the model predicting diabetes prevalence in New Jersey identified three significant predictors: the percentage of minority populations, the percentage of individuals aged 65 and older, and the percentage of individuals living below 200% of the poverty level (Equation 1). The model equation highlighted the influence of socioeconomic and demographic factors on diabetes prevalence, with higher rates observed in areas with more minorities, elderly residents, and economic hardship (see Appendix 2).

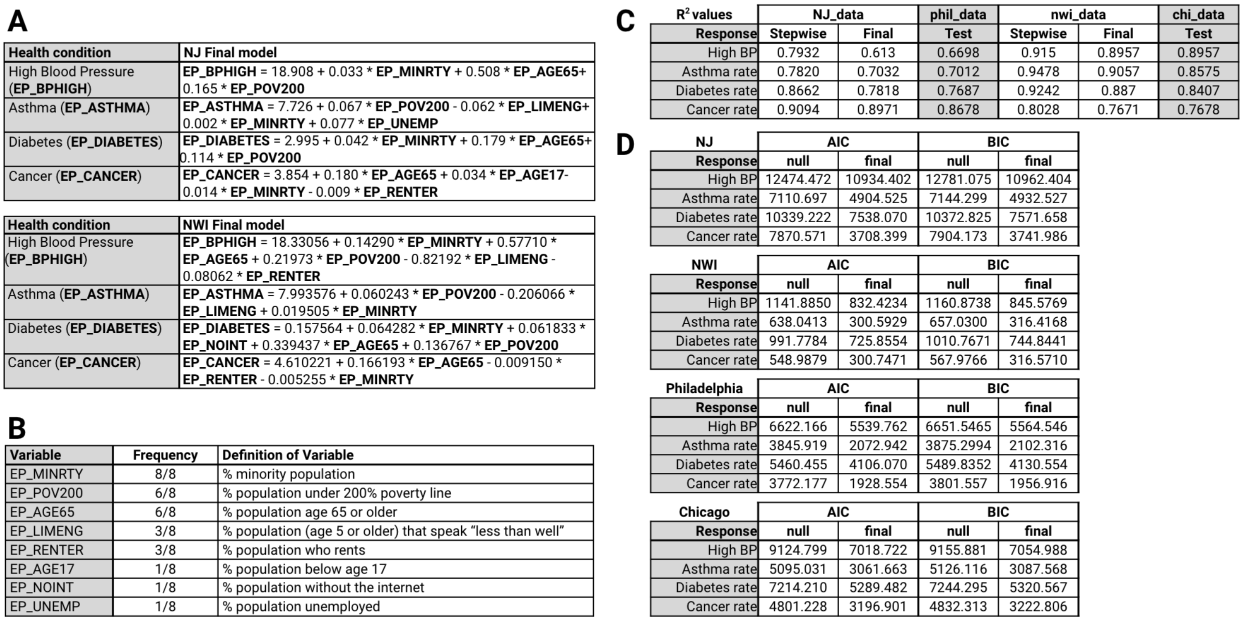
$$EP\_DIABETES = 2.995 + 0.042 * EP\_MINRTY + 0.179 * EP\_AGE65 + 0.114 * EP\_POV200 \qquad \textbf{(Eqn 1)}$$

When compared to null models based solely on pollution variables, the final models consistently had lower AIC and BIC scores, suggesting superior performance in predicting health condition prevalence and providing evidence of environmental injustice (Fig 1D). The variables of these models were also tested on neighboring metropolitan datasets, showing only minor reductions in R^2 values, indicating their generalizability as predictors in a region where environmental injustice is present. Residual analyses demonstrated uniform distributions across areas with significant minority populations, ensuring fairness and accuracy in predictions. Additionally, lower AIC and BIC scores were reported for models of neighboring datasets when compared a null model for that dataset, recapitulating superior predictive performance (Fig 1D).

Discussion

The study demonstrates that social and demographic factors can predict adverse health outcomes better than pollution variables, particularly in areas experiencing environmental injustice. Surprisingly, socioeconomic factors such as minority population percentage, poverty rates, and age demographics consistently emerged as key determinants. These findings may suggest that environmental injustice is rooted in systemic social inequities rather than solely environmental exposures.

While many predictors throughout our models had correlations that supported a hypothesis that environmental injustice was occuring, others were less intuitive, highlighting possible problems



A

Health condition	NJ Final model
High Blood Pressure (**EP_BPHIGH**)	**EP_BPHIGH** = 18.908 + 0.033 * **EP_MINRTY** + 0.508 * **EP_AGE65**+ 0.165 * **EP_POV200**
Asthma (**EP_ASTHMA**)	**EP_ASTHMA** = 7.726 + 0.067 * **EP_POV200** - 0.062 * **EP_LIMENG**+ 0.002 * **EP_MINRTY** + 0.077 * **EP_UNEMP**
Diabetes (**EP_DIABETES**)	**EP_DIABETES** = 2.995 + 0.042 * **EP_MINRTY** + 0.179 * **EP_AGE65**+ 0.114 * **EP_POV200**
Cancer (**EP_CANCER**)	**EP_CANCER** = 3.854 + 0.180 * **EP_AGE65** + 0.034 * **EP_AGE17**- 0.014 * **EP_MINRTY** - 0.009 * **EP_RENTER**

Health condition	NWI Final model
High Blood Pressure (**EP_BPHIGH**)	**EP_BPHIGH** = 18.33056 + 0.14290 * **EP_MINRTY** + 0.57710 * **EP_AGE65** + 0.21973 * **EP_POV200** - 0.82192 * **EP_LIMENG** - 0.08062 * **EP_RENTER**
Asthma (**EP_ASTHMA**)	**EP_ASTHMA** = 7.993576 + 0.060243 * **EP_POV200** - 0.206066 * **EP_LIMENG** + 0.019505 * **EP_MINRTY**
Diabetes (**EP_DIABETES**)	**EP_DIABETES** = 0.157564 + 0.064282 * **EP_MINRTY** + 0.061833 * **EP_NOINT** + 0.339437 * **EP_AGE65** + 0.136767 * **EP_POV200**
Cancer (**EP_CANCER**)	**EP_CANCER** = 4.610221 + 0.166193 * **EP_AGE65** - 0.009150 * **EP_RENTER** - 0.005255 * **EP_MINRTY**

B

Variable	Frequency	Definition of Variable
EP_MINRTY	8/8	% minority population
EP_POV200	6/8	% population under 200% poverty line
EP_AGE65	6/8	% population age 65 or older
EP_LIMENG	3/8	% population (age 5 or older) that speak "less than well"
EP_RENTER	3/8	% population who rents
EP_AGE17	1/8	% population below age 17
EP_NOINT	1/8	% population without the internet
EP_UNEMP	1/8	% population unemployed

C

R^2 values	NJ_data		phil_data	nwi_data		chi_data
Response	Stepwise	Final	Test	Stepwise	Final	Test
High BP	0.7932	0.613	0.6698	0.915	0.8957	0.8957
Asthma rate	0.7820	0.7032	0.7012	0.9478	0.9057	0.8575
Diabetes rate	0.8662	0.7818	0.7687	0.9242	0.887	0.8407
Cancer rate	0.9094	0.8971	0.8678	0.8028	0.7671	0.7678

D

NJ	AIC		BIC	
Response	null	final	null	final
High BP	12474.472	10934.402	12781.075	10962.404
Asthma rate	7110.697	4904.525	7144.299	4932.527
Diabetes rate	10339.222	7538.070	10372.825	7571.658
Cancer rate	7870.571	3708.399	7904.173	3741.986

NWI	AIC		BIC	
Response	null	final	null	final
High BP	1141.8850	832.4234	1160.8738	845.5769
Asthma rate	638.0413	300.5929	657.0300	316.4168
Diabetes rate	991.7784	725.8554	1010.7671	744.8441
Cancer rate	548.9879	300.7471	567.9766	316.5710

Philadelphia	AIC		BIC	
Response	null	final	null	final
High BP	6622.166	5539.762	6651.5465	5564.546
Asthma rate	3845.919	2072.942	3875.2994	2102.316
Diabetes rate	5460.455	4106.070	5489.8352	4130.554
Cancer rate	3772.177	1928.554	3801.557	1956.916

Chicago	AIC		BIC	
Response	null	final	null	final
High BP	9124.799	7018.722	9155.881	7054.988
Asthma rate	5095.031	3061.663	5126.116	3087.568
Diabetes rate	7214.210	5289.482	7244.295	5320.567
Cancer rate	4801.228	3196.901	4832.313	3222.806

Figure 1. Minimized models of socioeconomic predictors better predict adverse health outcomes in comparison to environmental predictors. A. Final minimized models for four adverse health outcomes in NJ and NWI. **B.** Only socioeconomic variables were included in final models, with several appearing in the majority of models in NJ and NWI. **C.** R^2 values for minimized final models are slightly reduced to stepwise counterparts. Test models report equivalent or improved R^2 values compared to final models. **D.** AIC and BIC values for final models are less than null model values for all adverse health outcomes and regions.

within our dataset. For example, all three models which contain English proficiency showed a negative correlation, suggesting tracts with higher proportions of poor English proficiency have lower rates of adverse health outcomes. This correlation likely reflects barriers to healthcare access causing underreporting in census data, rather than direct causal relationships. Similarly, while not included in the final models, diesel particulate matter also reported a negative correlation with health outcomes, yet it was determined higher levels of diesel particulate matter are correlated with poor English proficiency. Together, these data emphasize the need for improved census data collection methods to produce more equitable datasets.

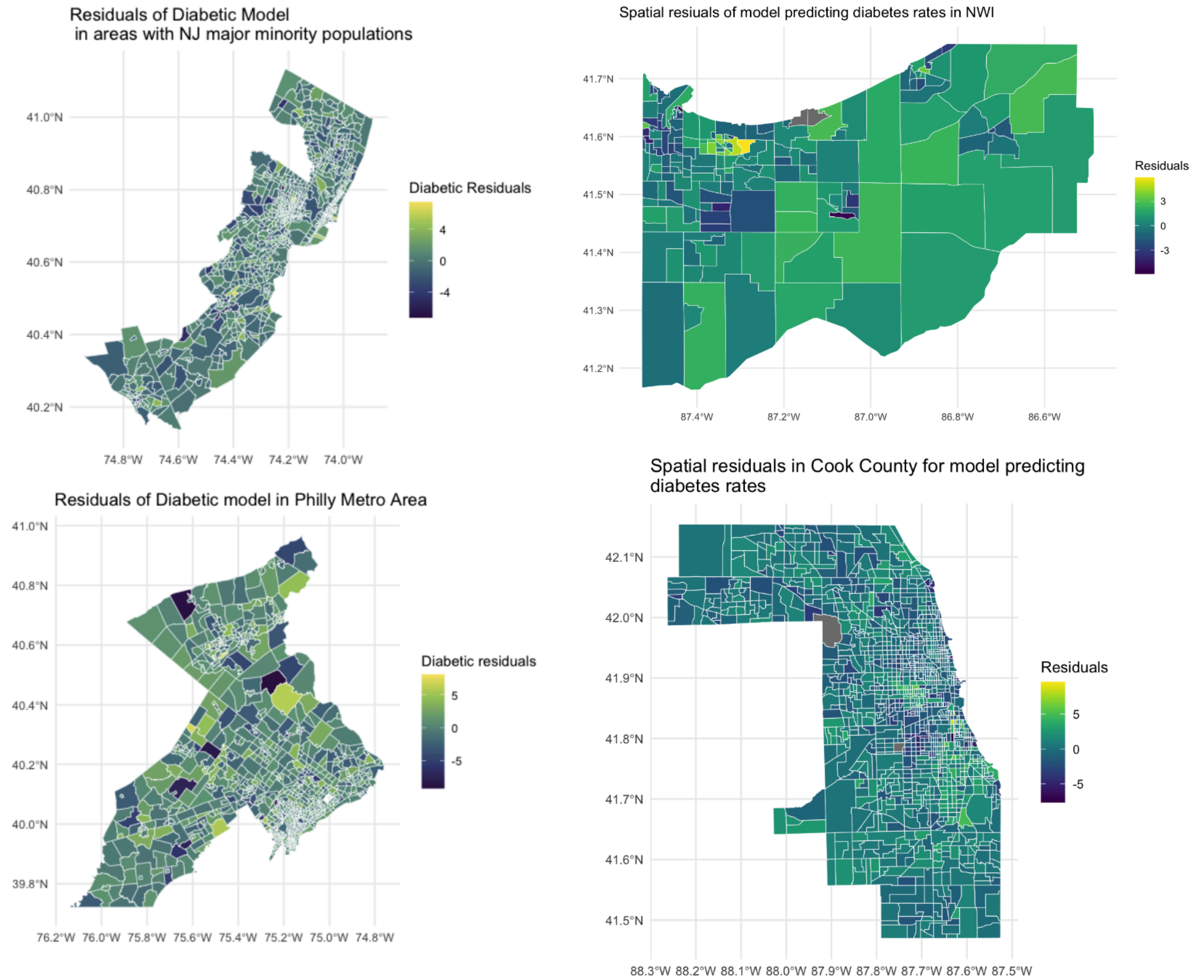
Separately, the null hypothesis used was not optimized for proper evaluation; only air pollution variables were included due to the complexity of other environmental variables in the dataset. This possibly decreased the accuracy of our null model, causing some final models to artificially perform better than in reality. Development of a more representative null hypothesis could greatly improve the significance of our findings.

Suburban areas like Newark and Trenton in New Jersey, as well as the industrial Northwest Indiana, exemplify how vulnerable populations—low-income, minority, and elderly—face disproportionate health disparities. While industrial and environmental conditions contribute to health outcomes, our study shows social inequities are more significant predictors. Enhancing census data quality could further refine and validate these findings, making the case for targeted public health interventions in these communities.

References:

1. Capital B News. "Gary's Steel Industry Is Linked to Increased Health Risks and Lower Quality of Life for Residents." https://gary.capitalbnews.org/gary-health-crisis-steel-industry-pollution/
2. U.S. Environmental Protection Agency. "Ringwood Mines Landfill Site." https://www.epa.gov/superfund/ringwood-mines-landfill
3. CDC. “EJI Indicators.” Place and Health - Geospatial Research, Analysis, and Services Program (GRASP), 3 Dec. 2024, www.atsdr.cdc.gov/place-health/php/eji/eji-indicators.html. Accessed 10 Dec. 2024.
4. Centers for Disease Control and Prevention and Agency for Toxic Substances Disease Registry. 2022 Environmental Justice Index. Accessed 16 October 2024. https://atsdr.cdc.gov/place-health/php/eji/eji-data-download.html

Appendices



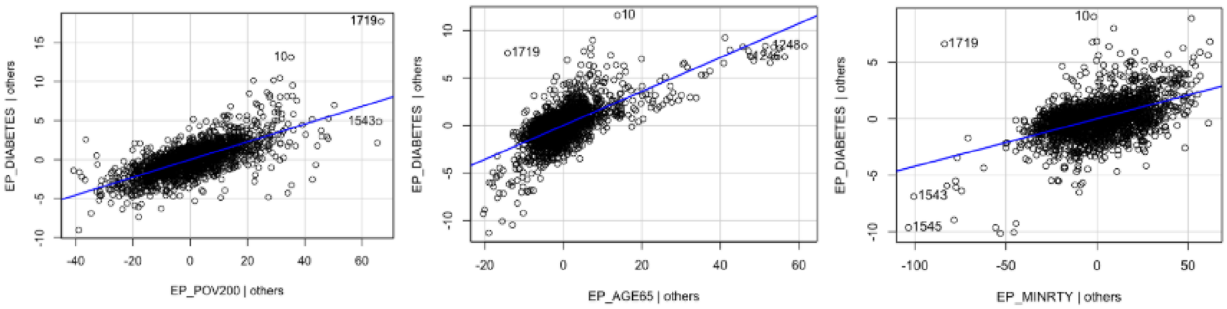
Residuals of Diabetic Model in areas with NJ major minority populations

Spatial resiuals of model predicting diabetes rates in NWI

Residuals of Diabetic model in Philly Metro Area

Spatial residuals in Cook County for model predicting diabetes rates

Appendix 1. Spatial residual maps for diabetes rate in all regions studied. Spatial residual maps allow us to determine if a specific geographic region (several proximal tracts) are unrepresented within the models. The optimal residual map would contain an even distribution of the color associated with zero, suggesting little to no variance across the region. The spatial residual map of NJ does not contain the entirety of the state in order to observe extremely small census tracts where dense populations exist; these regions were our focus region within the entire state. However, the entire state of NJ was evaluated in the study. These four maps serve as an example; this process was repeated for all models for all four adverse health income responses.



Appendix 2. Added variable plots for the final model predicting diabetes rate in New Jersey. All three socioeconomic variables included in the dataset have positive correlations for predicting diabetes prevalence. Added variable plots for other models were performed for analysis, where linearity and homoscedasticity were observed; this serves as an example.