# Supplemental Material: Environmental Injustices of Leaks from Urban Natural Gas Distribution Systems: Patterns Among and Between 13 U.S. Metro Areas

Zachary D. Weller[*Δ], Seongwon Im [Δ], Virginia Palacios[^], Emily Stuchiner[#], Joseph C. von Fischer[#]

Δ Department of Statistics, Colorado State University, 200 W. Lake St, 1877 Campus Delivery, Fort Collins, CO 80523-1877
^ Commission Shift 212 Flores Avenue, Laredo, Texas 78040212 Flores Avenue, Laredo, Texas 78040
# Department of Biology, Colorado State University, 200 W. Lake St, 1878 Campus Delivery, Fort Collins, CO 80523-1878
*Corresponding author: zachary.weller@pnnl.gov

12 pages
9 figures
0 tables

# 1. Statistical Methodology

We used a Bayesian, simple linear regression model to examine the relationship between leak density and the percentage of leak-prone pipe from the 13 metro areas where we conducted our ALD surveys. We fit this model using Rstanarm[27,28]. The results of this analysis are included in Section 3.1.

We used a Bayesian, negative binomial regression model to estimate the relationships between leak density and the EJ predictors, and leak density and median housing age, across the 13 metro areas. For this analysis, we combined data from each census tract in each metro area. In these models, we included a metro area-specific random effect. Before combining the data from each metro area, we standardized the EJ predictor variables and median household income to z-scores within each metro area. We used this standardization to account for differences in demographic and socioeconomic characteristics across metro areas. For example, in one metro area, the largest value for percent PoC by census tract was 80%, while in other metro areas it was 100%. Our standardization accounts for these relative differences. We do not standardize median housing age because median housing age is a proxy for pipeline infrastructure age, and we anticipate relative differences between metro areas (e.g., unlike income).

We used spatial, Bayesian, negative binomial regression model to characterize the relationship between leak density and the EJ predictors, median household income, and median housing age within nine metro areas. Metro areas with a small number of leak indications and/or little sampling effort (as indicated by the number of tracts surveyed) were not included in these analyses due to model convergence issues. Our model is given as,

$$Y_i \sim NegBin(E_i\lambda_i, \phi), \lambda_i = \exp\,(x_i\beta + \psi_i).$$

In this model, $Y_i$ is the observed count of leak indications for each census tract *i*. The offset variable, $E_i$, controls for variation in the miles of roadway surveyed across tracts and is the expected number of leak indications under the assumption of a constant leak density across tracts. The parameter $\lambda_i$ is the mean for census tract *i* and $\phi$ is the negative binomial dispersion parameter. We model the mean $\lambda_i$ as a function of tract-specific covariates (specified in the vector $x_i$) and a tract-specific random effect $\psi_i$, where these random effects are spatially structured using a conditional autoregressive (CAR) model. The parameters in the vector $\beta$ include an intercept term and the covariate effects on the average leak density. We estimated model parameters using rstan[29]. Additional details of this model, including estimation and sensitivity analyses, are provided in the Supplemental Material (Section 1). Hereafter, we refer to this model as the spatial model. Due to very low counts of leak indications and/or low numbers of census tracts surveyed, we excluded Birmingham, Burlington, Indianapolis, Mesa from further analysis.  The paucity of data from these cities prevented us from developing a meaningful spatial regression model that fit the data well.

We fit the spatial model separately for each metro area to estimate the relationship between leak density and each of the EJ predictors, median household income, and median housing age. The results of these models are shown in Sections 3.3 and reveal patterns in leak density as a function of each of the predictor variables within each metro area. We hypothesized that an association between leak density and the EJ predictors or household income in the single-variable models may be explained or reduced by also including median housing age in the model. To test this hypothesis, we estimated the spatial model for each pairwise combination of median housing age and one of the EJ predictors or median income. The results of these two-predictor spatial models are shown in Section 3.4.

## 2. Statistical Model Estimation

We used rstan[29] to fit a Bayesian simple linear regression model with leak density as the response and the percentage of leak prone pipe as the predictor variable (Figure 1i). We used non-informative uniform prior distributions for the intercept and slope parameters. We used an exponential distribution with parameter one as the prior distribution for the residual variance (the default in rstan). We ran four chains for 5,000 iterations each and discarded the first 1,000 iterations as burn-in. We examined the trace plots as well as the Gelman-Rubin diagnostics to ensure that the chains had converged.

We used rstan to fit a Bayesian negative binomial regression model for all of the metro areas combined, which we refer to as the combined analysis (Section 3.2, Figure 2 and Figure SM2). We used negative binomial models because diagnostics indicated that Poisson and quasiPoisson models did not adequately model overdispersion[51,52]. We used these models to estimate trends across all metro areas. In these models, we estimated a metro area specific random effect. We fit a separate model for each EJ predictor and for housing age with leak density as the response variable.

For the combined analysis, we used a normal prior distribution centered at 0 with a variance of 2.5 for both the intercept and slope parameters. We used an exponential prior distribution with a rate of 1 for the dispersion parameter for the negative binomial regression. Like the simple linear regression model above, we ran four chains for 5,000 iterations each with the first 1,000 iterations as burn-in. The MCMC trace plots suggested that the chains mixed well indicating adequate convergence.

We fit a Bayesian negative binomial spatial regression model for each metro area, which we refer to as the metro-by-metro analysis (Main Text Figures 3-6, SM Figures SM3-SM9). We fit this model using Stan[28] through R[29]. We used these models to estimate trends within each metro area. In these models, we estimated a census tract specific random effect. These random effects were assumed to be spatially correlated. We modeled the spatial correlation using a conditional autoregressive (CAR) model. The CAR model requires that all census tracts be connected via a graph structure. Census tracts that share a border with one another (i.e., adjacent tracts) are considered neighbors in the CAR model. In some metro areas, there were nonconnected clusters of census tracts, which we refer to as islands. To estimate the CAR model, we manually constructed a minimum number of connections between islands by connecting two census tracts that were closest to each other from each island. For example, there were four islands in Los Angeles, and we created a graph structure by including three manual connections.

We fit a separate spatial model for each EJ predictor and for housing age with leak density as the response variable for each metro area. We also fit models with multiple predictors. Of primary interest were two predictor models where each EJ predictor was paired with median housing age to estimate the relationship between the EJ predictor and leak density while controlling for housing age (a proxy for infrastructure age).

The spatial regression model is specified by regression parameters, a negative binomial dispersion parameter, spatial correlation parameter, and variance of the spatial random effects. We denote the regression parameters as beta's, the dispersion parameter as phi, the spatial correlation as rho, and the variance of the random effects as sigma^2. To impose identifiability of the variance of the negative binomial distribution, we fixed the dispersion parameter phi. We estimated rho and sigma using a uniform(0, 1) prior for rho and a gamma(2,2) prior for 1/sigma^2. For the regression parameters, we used normal prior distribution centered at 0 with a variance of 2.5 for both the intercept and slope parameters. We estimated model parameters using 3 chains, each containing 10,000 iterations with 5,000 iterations as burn-in. We performed a sensitivity analysis of our prior distributions to verify that they had little effect on the posterior distribution.

We fixed the negative binomial dispersion parameter for the spatial regression models, and we performed a sensitivity analysis for this parameter. For each metro area and predictor variable(s), we set the dispersion parameter equal to the dispersion parameter estimated by a frequentist non-spatial negative binomial regression model using the same predictor(s). Within a metro area, the estimated dispersion parameter was similar across models with different predictors. Across metro areas and predictors, the typical value for the dispersion parameter was 4.5. As a sensitivity analysis, we re-estimated each model using a set of different dispersion parameter values, which included 2, 4, 8, and 13. Regardless of this choice of the dispersion parameter, our conclusions about the effects of the different predictors changed very little.

## 3. Survey Coverage

The boxplots in Figure SM1 show the distribution of each of the explanatory variables from surveyed census tracts. In each metro area, we designed our ALD survey to cover a diverse set of areas with regards to socioeconomic features and varying types of infrastructure (e.g., residential vs industrial). For most variables in most metro areas, our survey included census tracts having a wide range of sociodemographic features.

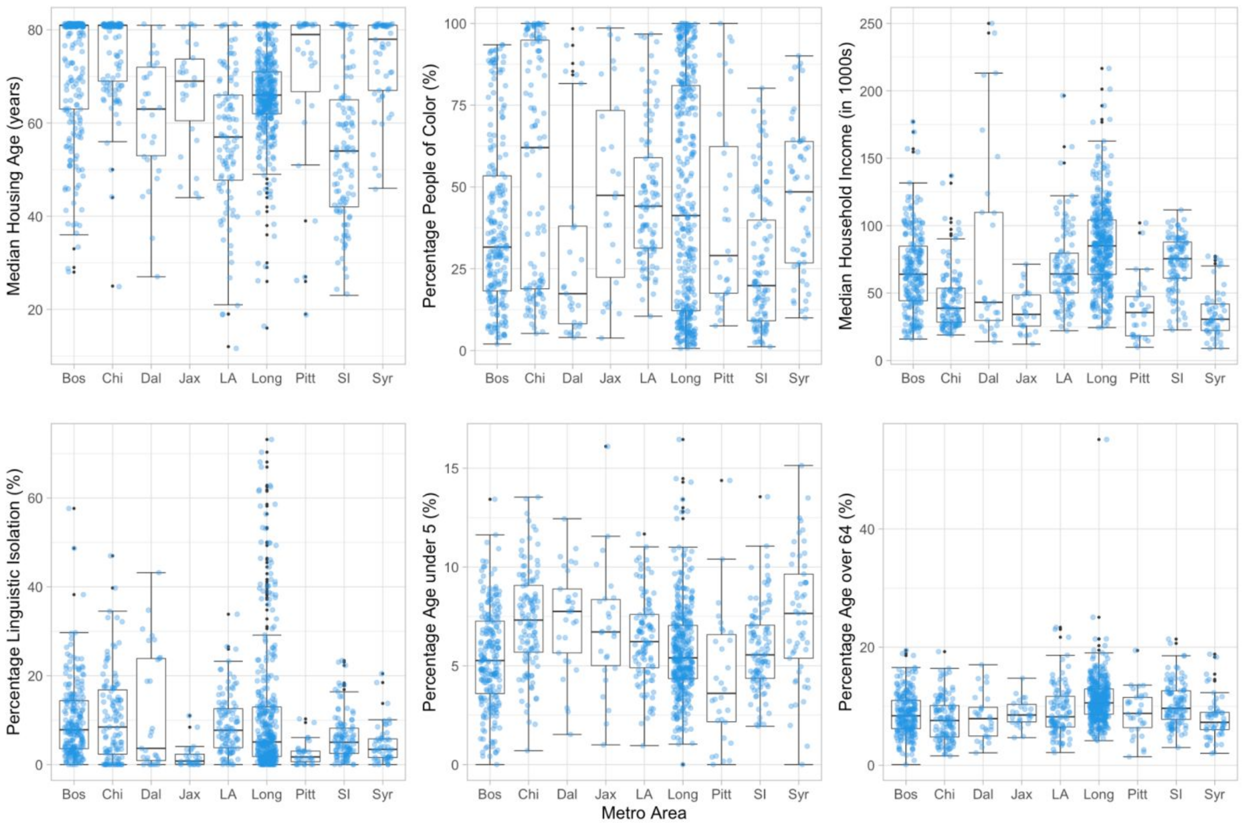

**Figure SM1**: *Boxplots of sociodemographic predictors and median housing age for surveyed census tracts in nine metro areas. Each dot represents a surveyed census tract. In each metro area, we designed our survey efforts to cover regions that varied across their sociodemographic and socioeconomic characteristics.*

# 4. Additional Results of the Analysis of All Metropolitan Areas Combined

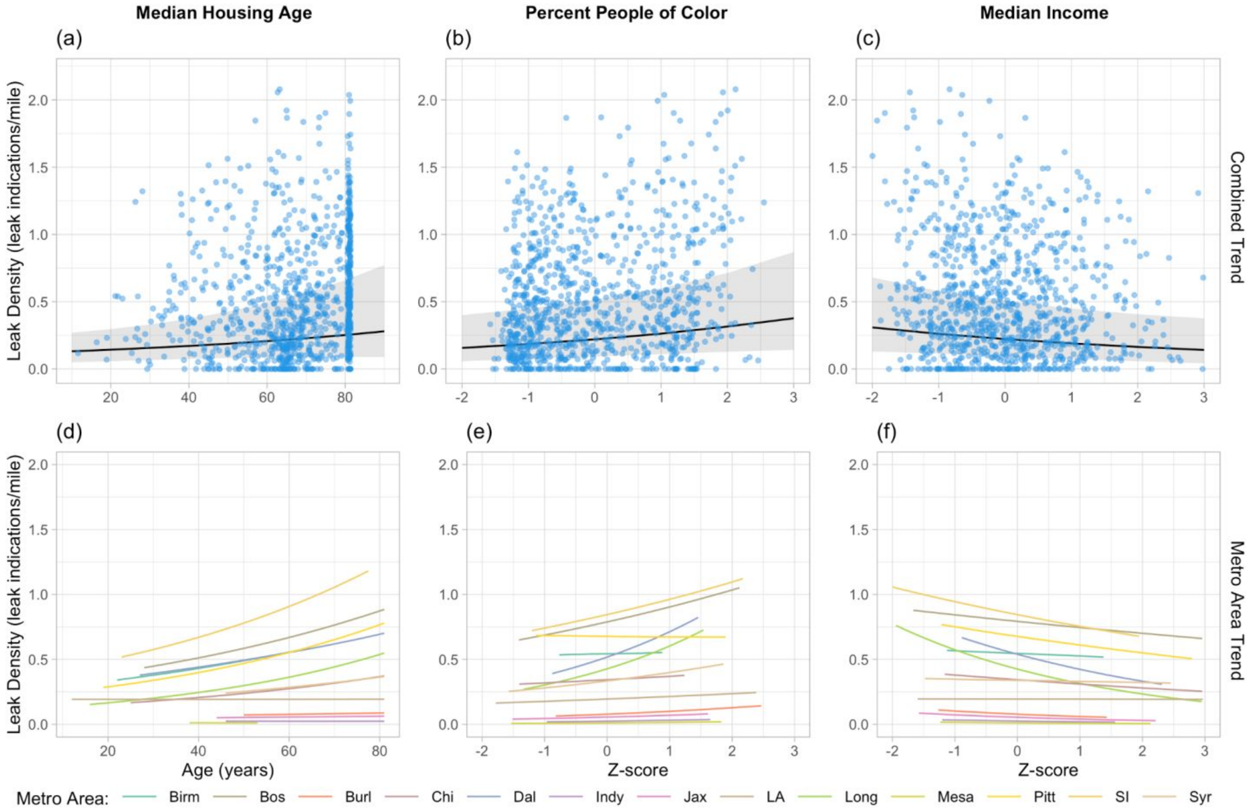

**Figure SM2**: *The combined relationship between median housing age (a), percent PoC (b), median income (c), and leak density. The same relationships within each metro area are shown in panels (d - f). Each point in the figures represents a single census tract, where observations have been combined across all metro areas. There is statistical evidence of increasing leak density with both older housing and a greater percentage of people of color. The y-axis is truncated at 2.2 leaks/mile to facilitate comparisons and visualization of the trend lines. There are 14 census tracts (1.3% of the total observations) that are not shown due to this truncation.*

We performed the same combined analyses for three other EJ predictors: percent linguistic isolation, percentage of population under age 5, and percentage of population over age 64. There was little statistical evidence of a relationship between leak density and these three other predictors. The largest estimated effect was for percent linguistic isolation. Leak density is estimated to increase by 8% (-8%, 22%) for each one standard deviation additional difference in percent linguistic isolation. For each of the age variables, the effect estimates were near zero, with estimated increases of approximately 1.5% associated with an additional one standard deviation difference for both age under 5 (-9%, 11%) and age over 64 (-6%, 11%).

# 5. Additional Results of the Within Metro Area Analysis

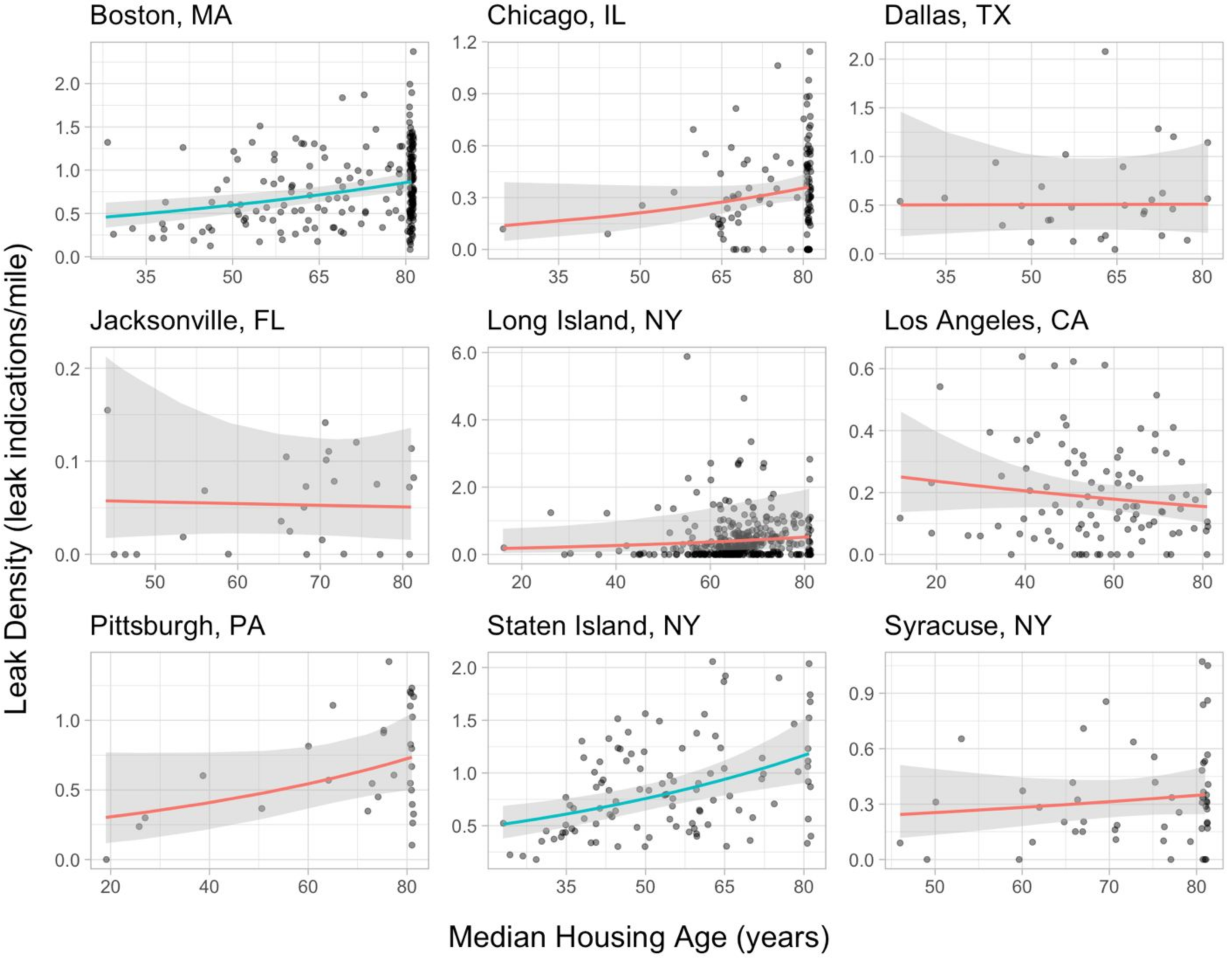

**Figure SM3:** *The relationship between median housing age and leak density from surveyed census tracts in each of nine metro areas. In eight of the nine metro areas, leak density is estimated to increase with increased housing age. Both the magnitude of this association and the overall leaks/mile vary among metro areas. Blue lines indicate that there is statistical evidence to conclude that leaks/mile increase with increasing housing age. Note that the x and y axes vary across metro areas.*

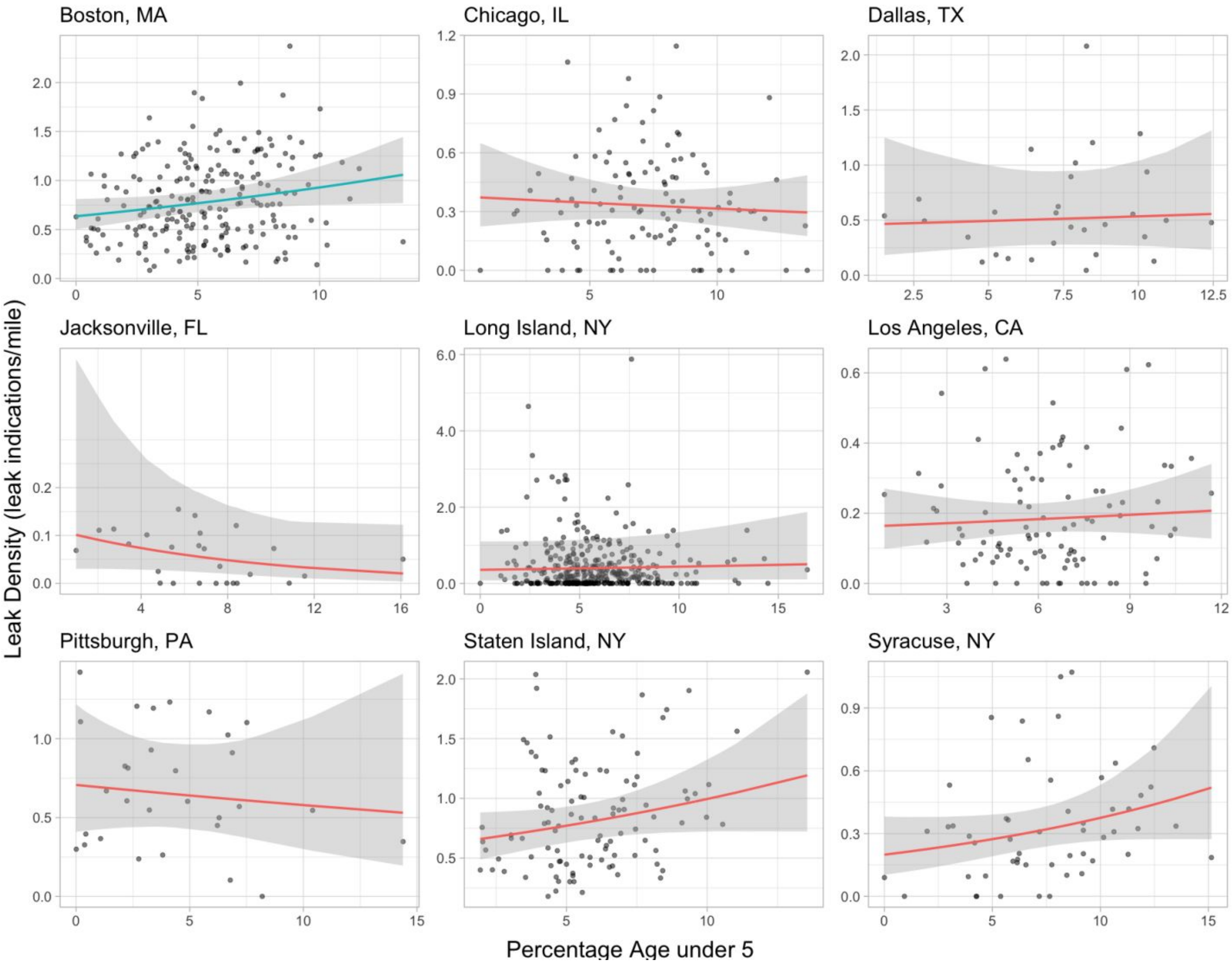

**Figure SM4**: *The relationship between percent under age 5 and leak density from surveyed census tracts in nine metro areas. Both the magnitude of this association and overall leaks/mile vary among metro areas. Blue lines indicate that there is statistical evidence to conclude that leaks/mile increase with increasing percentage age under 5. Note that the x and y axes vary across metro areas.*

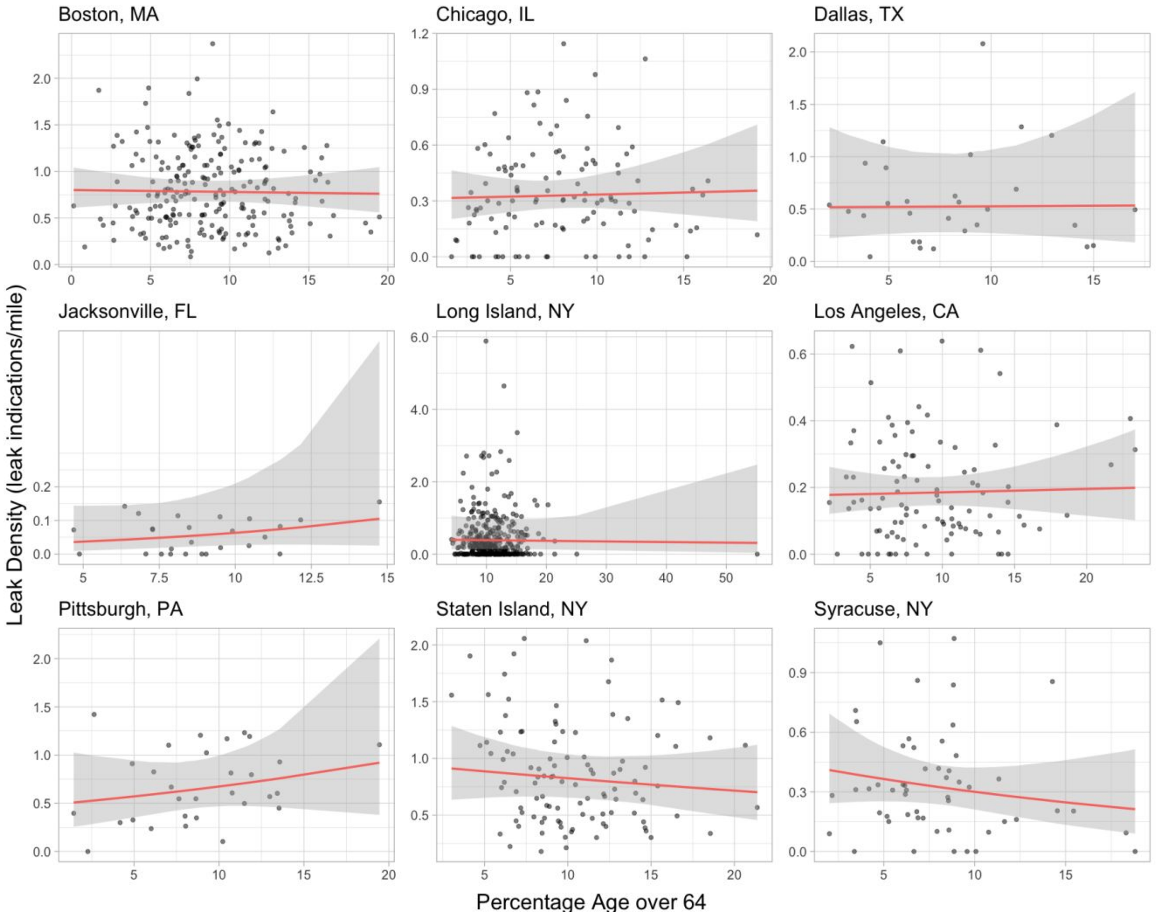

**Figure SM5:** *The relationship between percent over age 64 and leak density from surveyed census tracts in nine metro areas. Across the nine metro areas, there was not statistical evidence of a relationship between percentage age over 64 and leaks per mile. Note that the x and y axes vary across metro areas.*

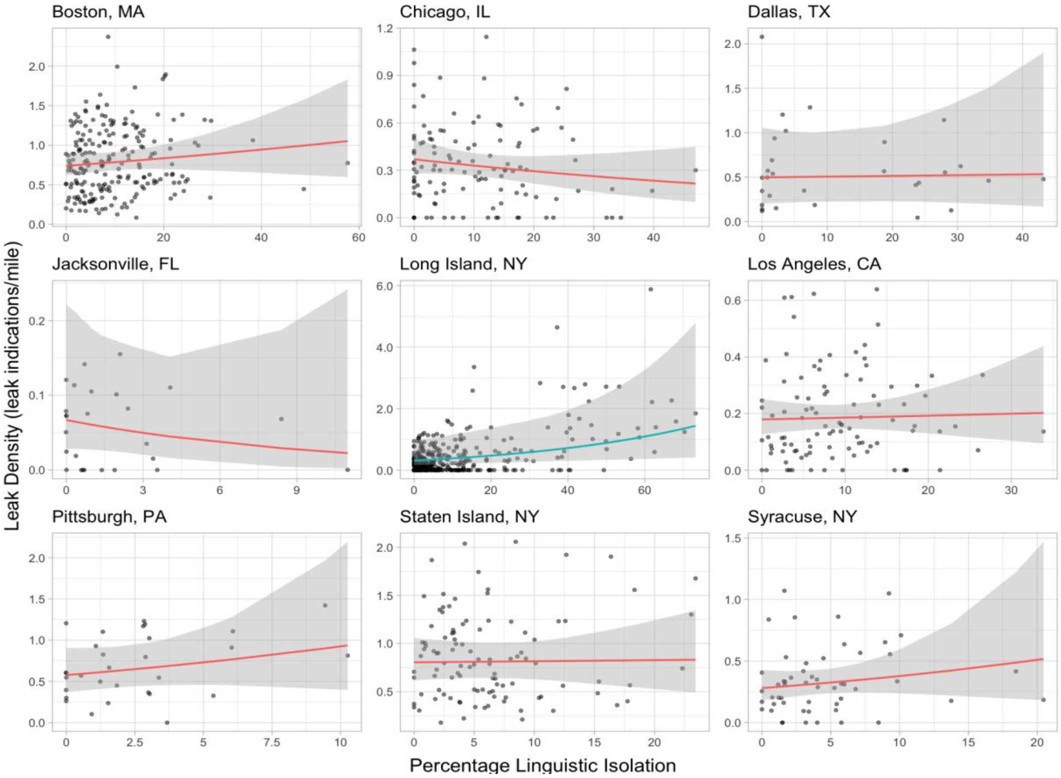

**Figure SM6**: *The relationship between percent linguistic isolation and leak density from surveyed census tracts in nine metro areas. Both the magnitude of this association and overall leaks/mile vary among metro areas. Blue lines indicate that there is statistical evidence that leaks/mile increase with increasing linguistic isolation. Note that the x and y axes vary across metro areas.*

There was no clear association between leak density and population age characteristics (percent under age 5, percent over age 64) across the nine metro areas. In six metro areas (Boston, Dallas, Long Island, Los Angeles, Staten Island, and Syracuse), leak density was estimated to increase with increasing percent under age five. In Boston (estimate of 10%, 95% CrI: 1%, 21%), there was statistical evidence of an association between leak density and percent under age 5. There were five metro areas (Chicago, Dallas, Jacksonville, Los Angeles, and Pittsburgh) where leak density was estimated to increase with increasing percent over age 64. In many cases, these estimates were near zero and had large uncertainty.

# 6. Additional Results of Controlling for Median Housing Age

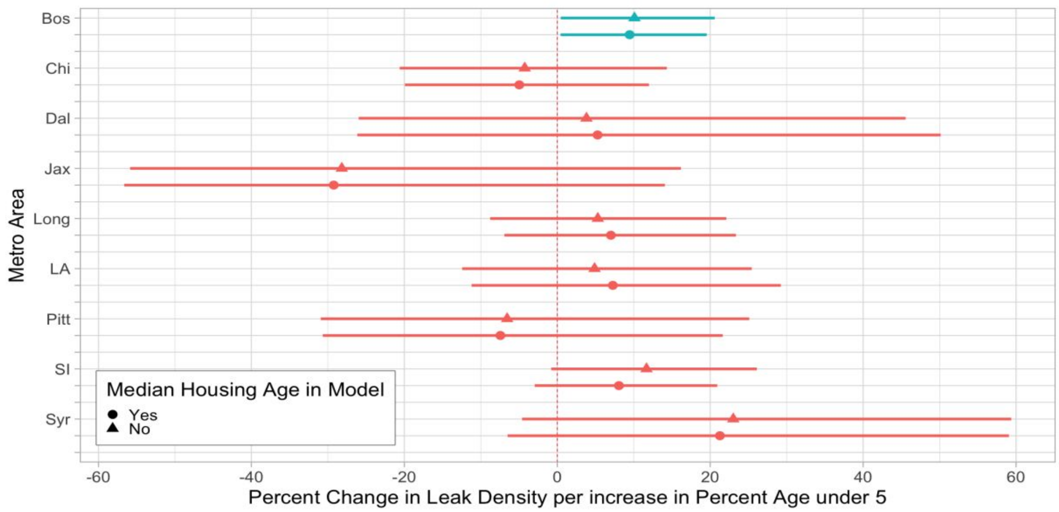

**Figure SM7**: *The estimated relationship between percent under age 5 and leak density with and without median housing age in the model. The x-axis shows the estimated percent change in leak density associated with a one-standard deviation additional difference in percent under age 5. Blue credible intervals do not contain zero, indicating that there is statistical evidence of a relationship between percentage age under 5 and leak density in the corresponding metro area.*

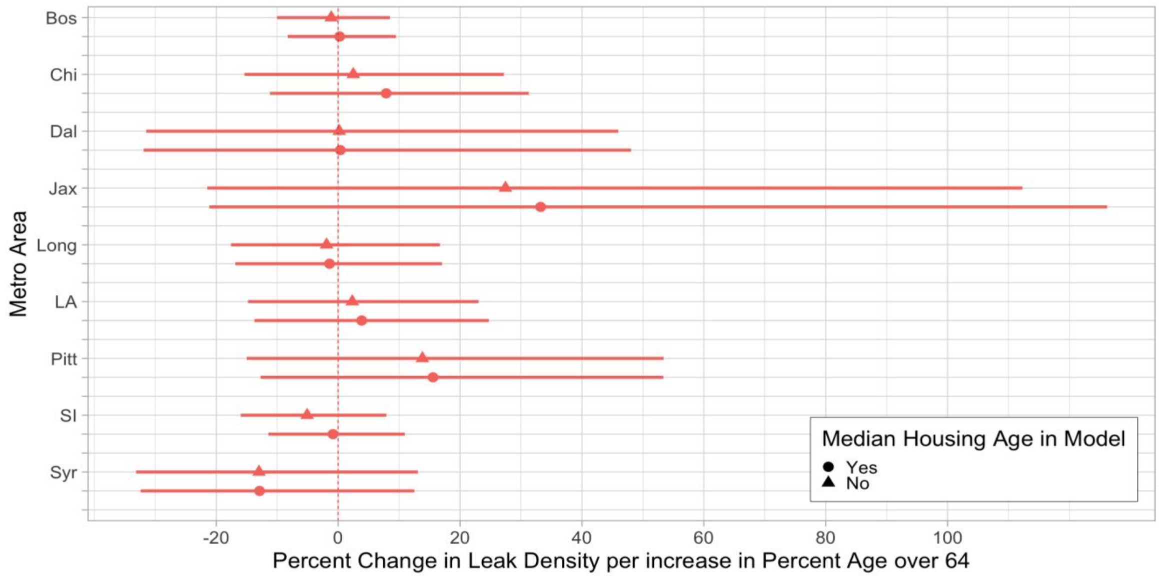

**Figure SM8**: *The estimated relationship between percentage age over 64 and leak density with and without median housing age in the model. The x-axis shows the estimated percent change in leak density associated with a one-standard deviation additional difference in percentage age over 64. Metro areas in blue have credible intervals that do not include 0.*

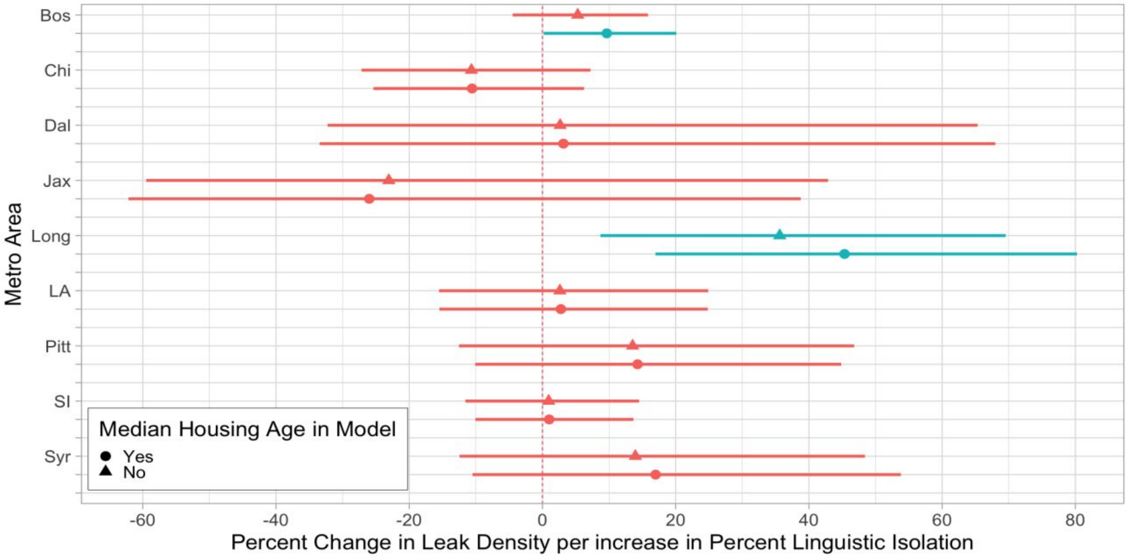

**Figure SM9**: *The estimated relationship between percent linguistic isolation and leak density with and without median housing age in the model. The x-axis shows the estimated percent change in leak density associated with a one-standard deviation additional difference in percent linguistic isolation. Blue credible intervals do not contain zero, indicating there is statistical evidence of a relationship between percent linguistic isolation and leak density in the corresponding metro area.*

We also examined the relationship between leak density and each of the three other EJ predictors (linguistic isolation, age under 5, and age over 64) in models that also included median housing age. Estimates and credible intervals of the effects of these predictors, with and without median housing age, are shown in Figures SM7-SM9. For most metro areas, the effects of linguistic isolation, age under 5, and age over 64 on leak density were similar with and without median housing age. There were a few notable results among these analyses. First, in Long Island there was statistical evidence that leak density increases with linguistic isolation (Figure SM9) and in Boston the effect of linguistic isolation increased from an estimated 5% to 10% when median housing age was included in the model. When median housing age was included in the model for Boston, there was statistical evidence (estimate of 10%, 95% CrI: 0.2%, 20%) of an association between linguistic isolation and leak density. Second, in Boston there was statistical evidence to conclude that leak density increases with increasing percentage under age 5 (estimate of 9%, 95% CrI: 0%, 19%).