Examining the relationship between socioeconomic variables and air pollution in Toronto *

Farah Chin *400229991*
Talat Hakim *400315290*
Nathan Nadeau *400342430*
Cindia Dao-Vu *400319161*
Ifra Awan *400261667*
Oliver Lawless *400271127*

---

abstract goes here; some extra spiel added here

Keywords: pollution, demographics, spatial analysis, transportation

---

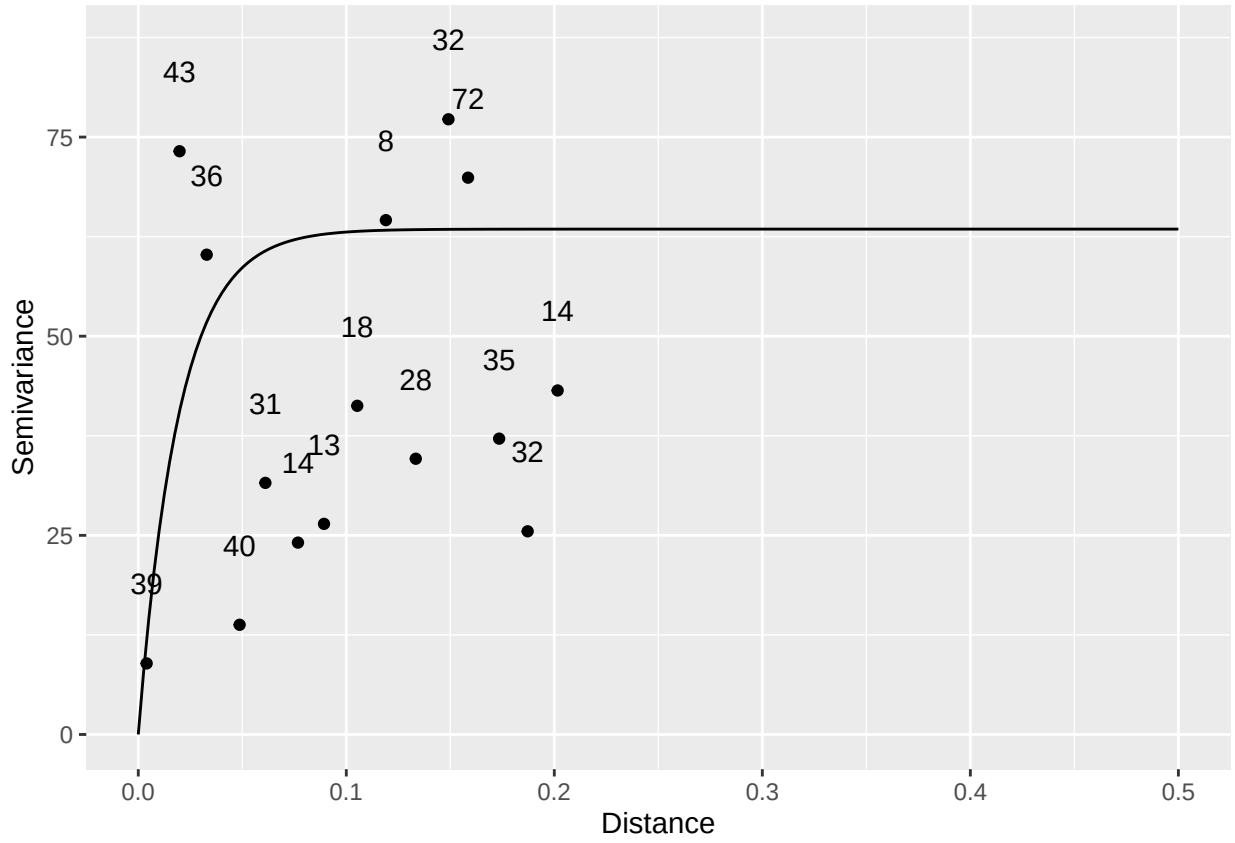
Introduction

This study analyzes the correlations between emission-based air pollution and socioeconomic spatial phenomena within the City of Toronto. It examines the predominant air quality factors responsible for the emission of fine particulate matter (PM2.5). The aim of this study is to address variations in air pollution statistics throughout the city. Generally, air quality tends to be worse in areas of lower socioeconomic status (SES) where people are forced to live near major sources of contamination such as major roads and factories. This paper contains a literature review to set a foundation for the analysis. Using air quality data from several sensors, we interpolated fine particulate matter concentrations across Toronto. Then, using data collected from the Canadian census we performed a linear regression analysis to determine the correlation between the air pollution of an area and the socioeconomic factors. Next, the statistics were summarized and interpreted to determine which socioeconomic factors are the most correlated with poor air quality. Finally, the paper concludes with some recommendations for future correlation analysis and steps the City of Toronto should take to manage air pollution.

*Paper submitted to complete the requirements of ENVSOCTY 4GA3 Applied Spatial Statistics; with additional edits by Antonio Paez for this version.

Background

Data



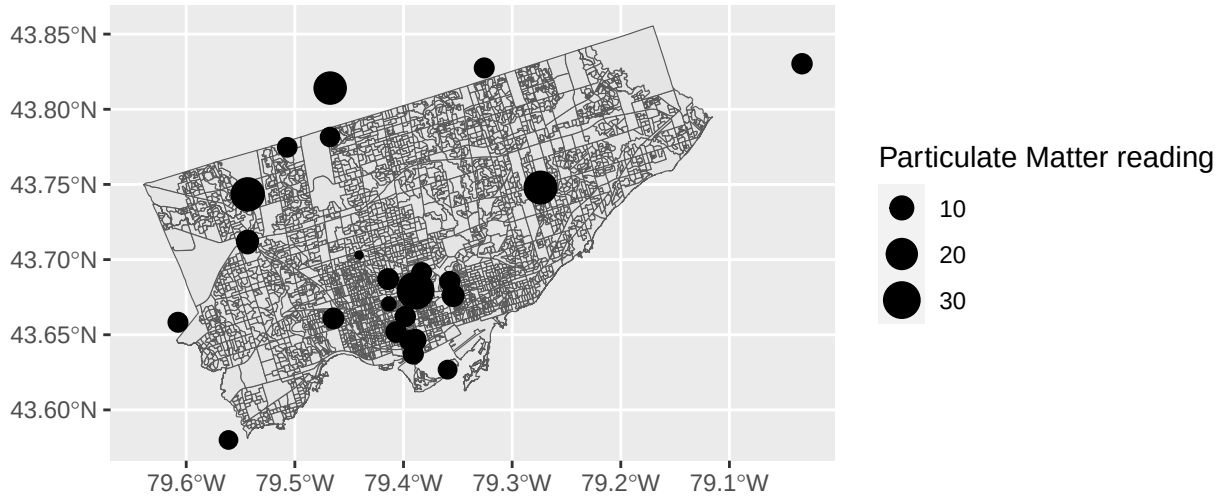
Toronto Census Data

Go to https://mountainmath.github.io/cancensus/index.html and follow instructions for setting up CensusMapping account and generating API key to retrieve data.

More info: https://mountainmath.github.io/cancensus/articles/cancensus.html

Set your cache to the data folder:

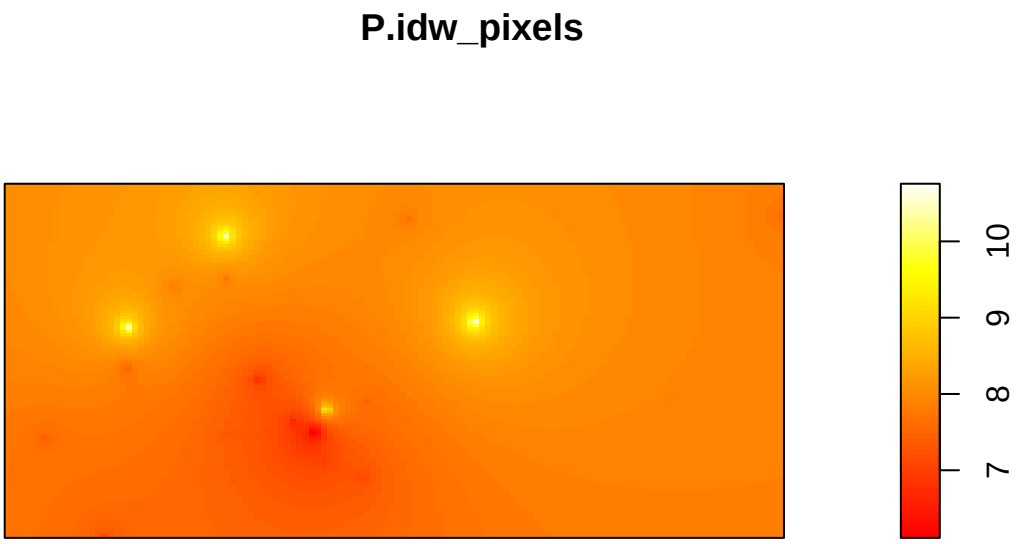
```
## [1] "C:/Users/Farah Chin/Documents/McMaster/ENVSOCTY 4GA3/Air-Pollution-Correlates-4GA3/data"
```

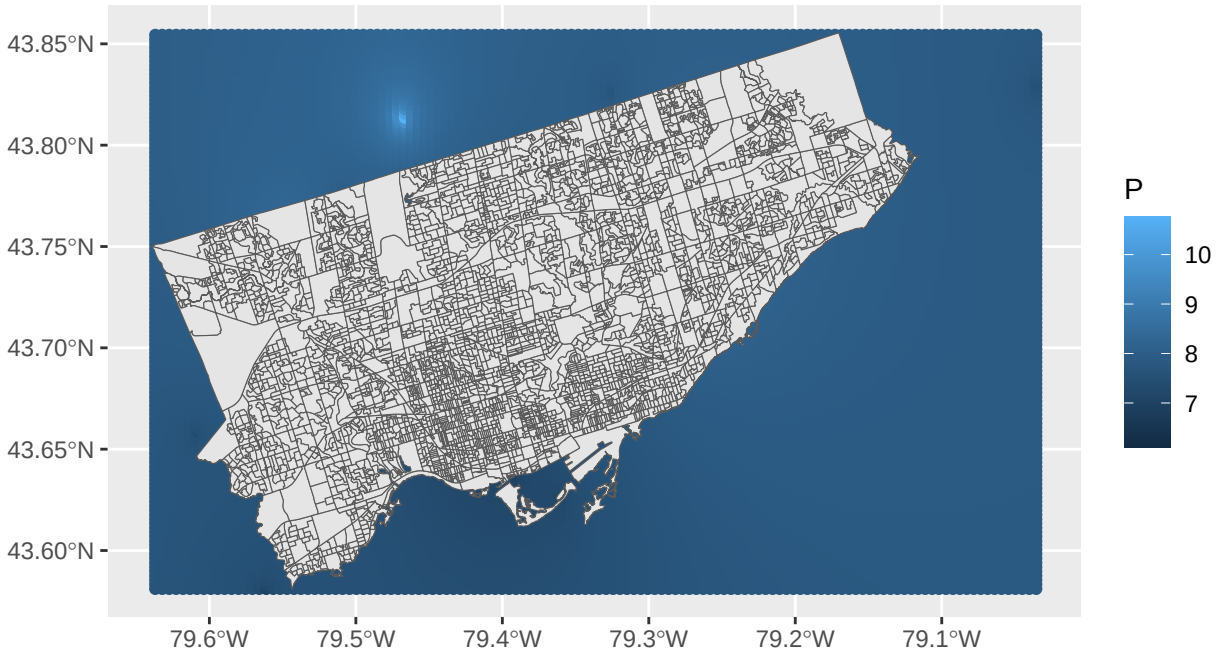


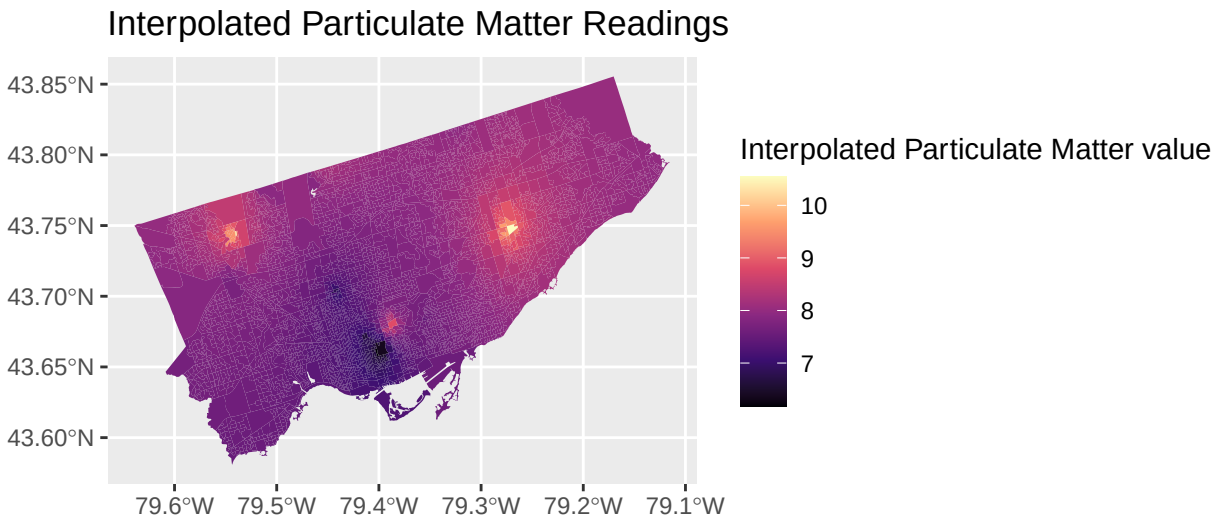
Here we use Leave One Out cross validation to determine the optimal power: https://www.statology.org/leave-one-out-cross-validation/

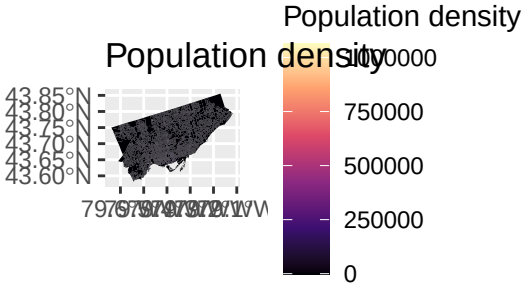
See also: https://rpubs.com/Dr_Gurpreet/interpolation_idw_R

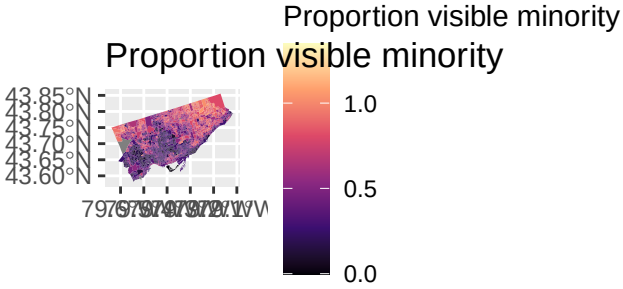
```
## [1] 0.471
```

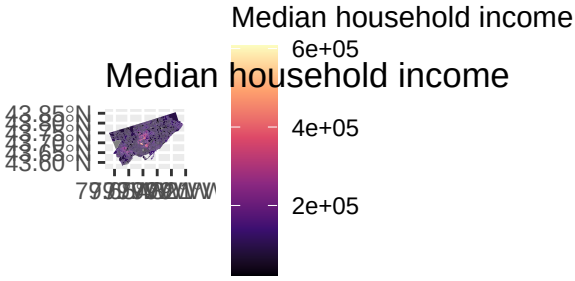

P.idw_pixels
10
9
8
7

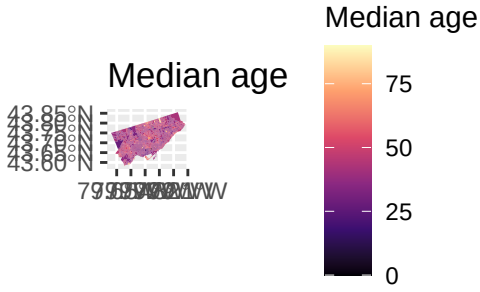

43.85°N
43.80°N
43.75°N
43.70°N
43.65°N
43.60°N
79.6°W
79.5°W
79.4°W
79.3°W
79.2°W
79.1°W
P
10
9
8
7

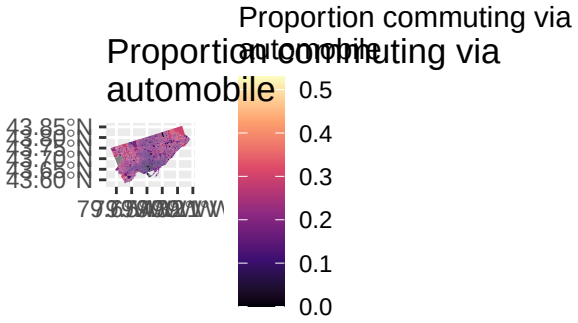

Interpolated Particulate Matter Readings
43.85°N
43.80°N
43.75°N
43.70°N
43.65°N
43.60°N
79.6°W
79.5°W
79.4°W
79.3°W
79.2°W
79.1°W
Interpolated Particulate Matter value
10
9
8
7

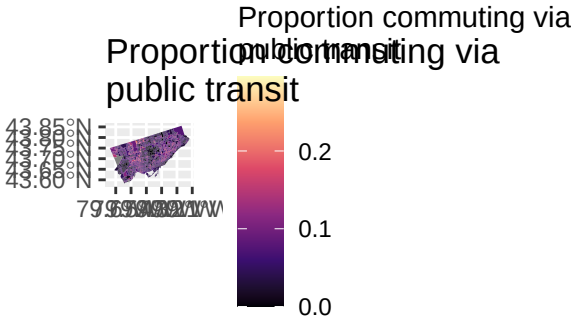

Population density
Population density
1000000
750000
500000
250000
0
43.85°N
43.80°N
43.75°N
43.70°N
43.65°N
43.60°N

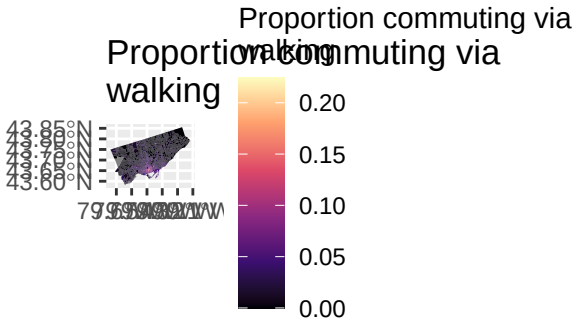

Proportion visible minority
Proportion visible minority
1.0
0.5
0.0
43.85°N
43.80°N
43.75°N
43.70°N
43.65°N
43.60°N

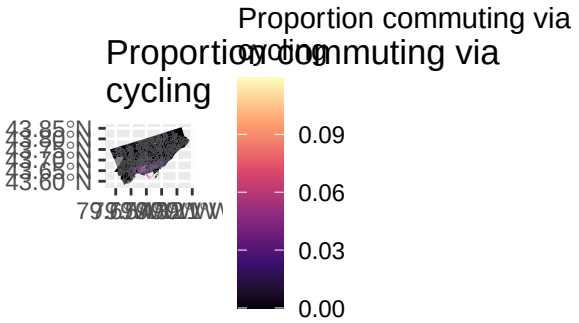

Median household income
Median household income
6e+05
4e+05
2e+05

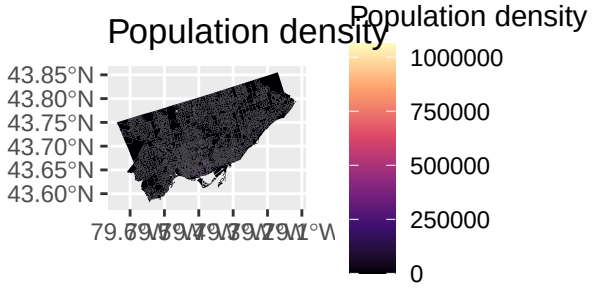

Median age
Median age
75
50
25
0

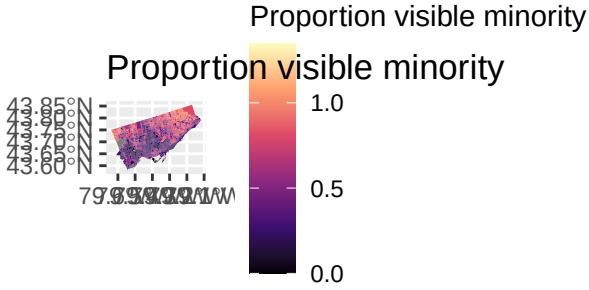

Proportion commuting via automobile
43.85°N
43.80°N
43.75°N
43.70°N
43.65°N
43.60°N
0.5
0.4
0.3
0.2
0.1
0.0

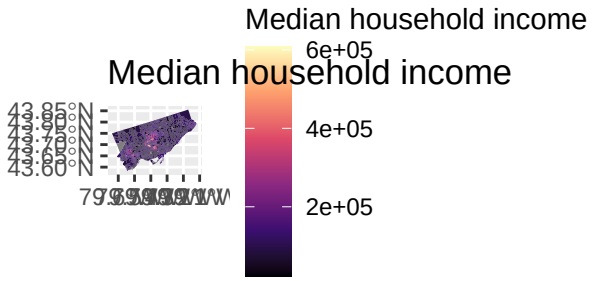

Proportion commuting via public transit
43.85°N
43.80°N
43.75°N
43.70°N
43.65°N
43.60°N
0.2
0.1
0.0

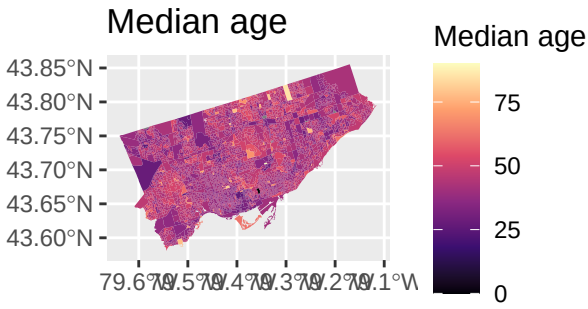

Proportion commuting via walking
43.85°N
43.80°N
43.75°N
43.70°N
43.65°N
43.60°N
0.20
0.15
0.10
0.05
0.00


Proportion commuting via cycling
43.85°N
43.80°N
43.75°N
43.70°N
43.65°N
43.60°N
0.09
0.06
0.03
0.00

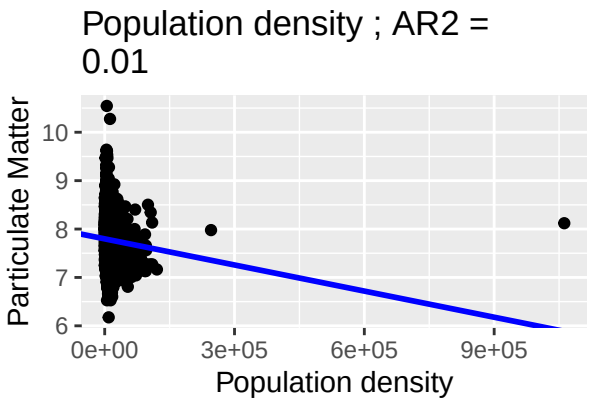


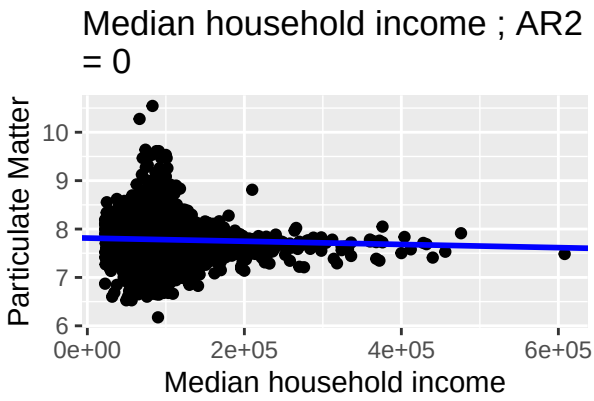


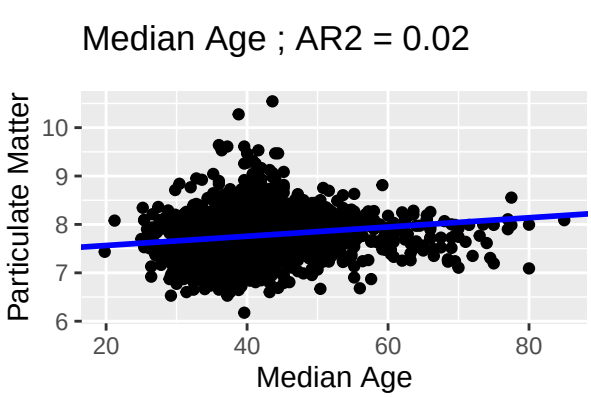


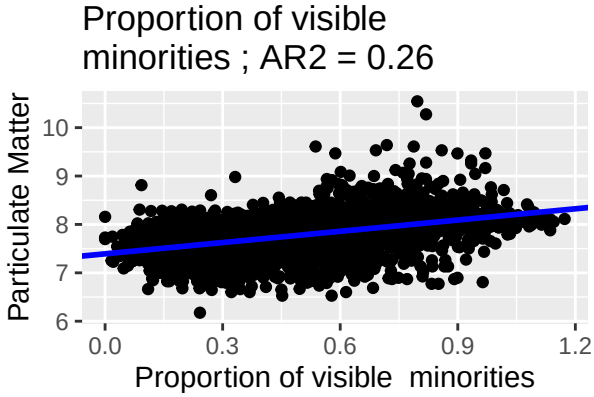


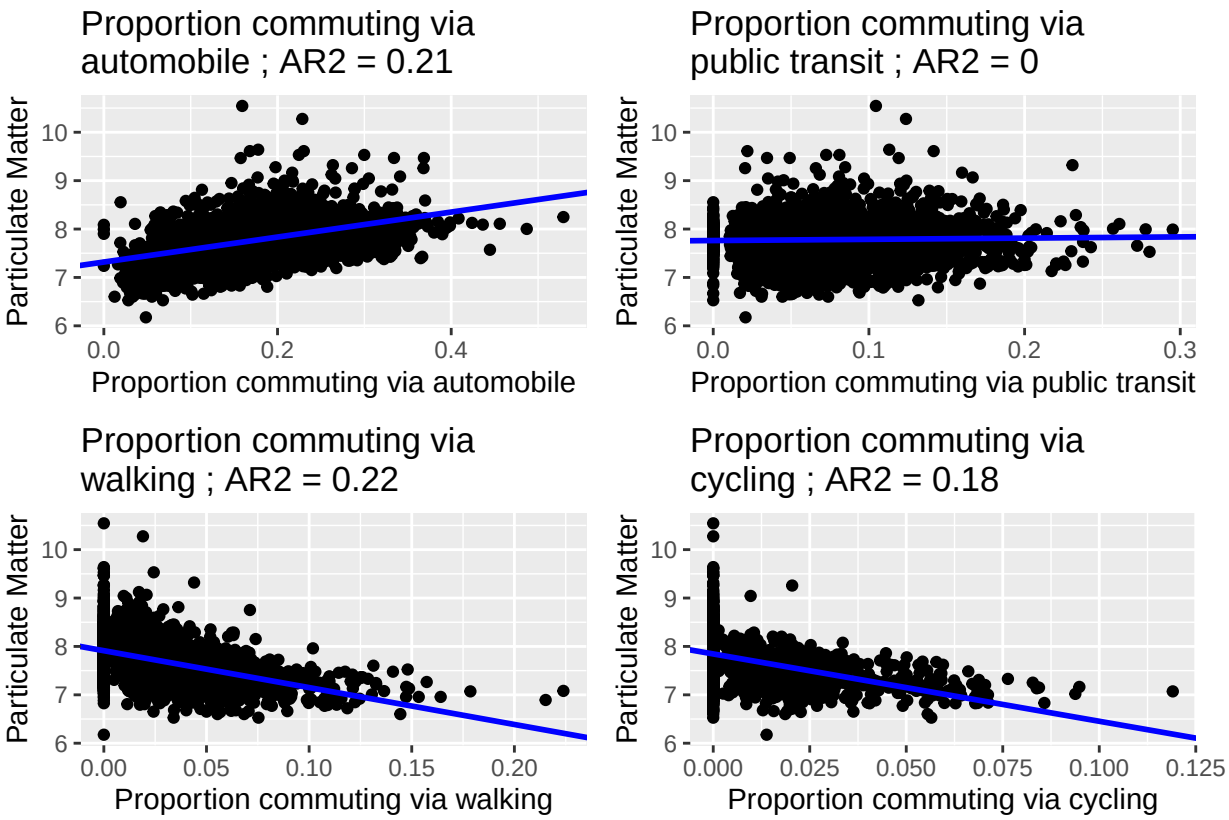

```
##      Variable                                          Moran's Statistic (I)
## [1,] "Population density"                              "0.122873369614986"
## [2,] "Median household income"                         "0.545540163845709"
## [3,] "Median age"                                      "0.305545970526739"
## [4,] "Proportion visible minority"                     "0.747715284796605"
## [5,] "Proportion commuting via automobile"             "0.556743175149464"
## [6,] "Proportion commuting via public transportation" "0.335970618239244"
## [7,] "Proportion commuting via walking"                "0.583850511148544"
## [8,] "Proportion commuting via cycling"                "0.40985144504303"
##      p-value
## [1,] "2.56077593971829e-68"
## [2,] "0"
## [3,] "1.4150620141619e-226"
## [4,] "0"
## [5,] "0"
## [6,] "3.66842309312586e-273"
## [7,] "0"
## [8,] "0"
```


Population density ; AR2 = 0.01
Particulate Matter
10
9
8
7
6
0e+00
3e+05
6e+05
9e+05
Population density


Median household income ; AR2 = 0
Particulate Matter
10
9
8
7
6
0e+00
2e+05
4e+05
6e+05
Median household income


Median Age ; AR2 = 0.02
Particulate Matter
10
9
8
7
6
20
40
60
80
Median Age


Proportion of visible minorities ; AR2 = 0.26
Particulate Matter
10
9
8
7
6
0.0
0.3
0.6
0.9
1.2
Proportion of visible minorities



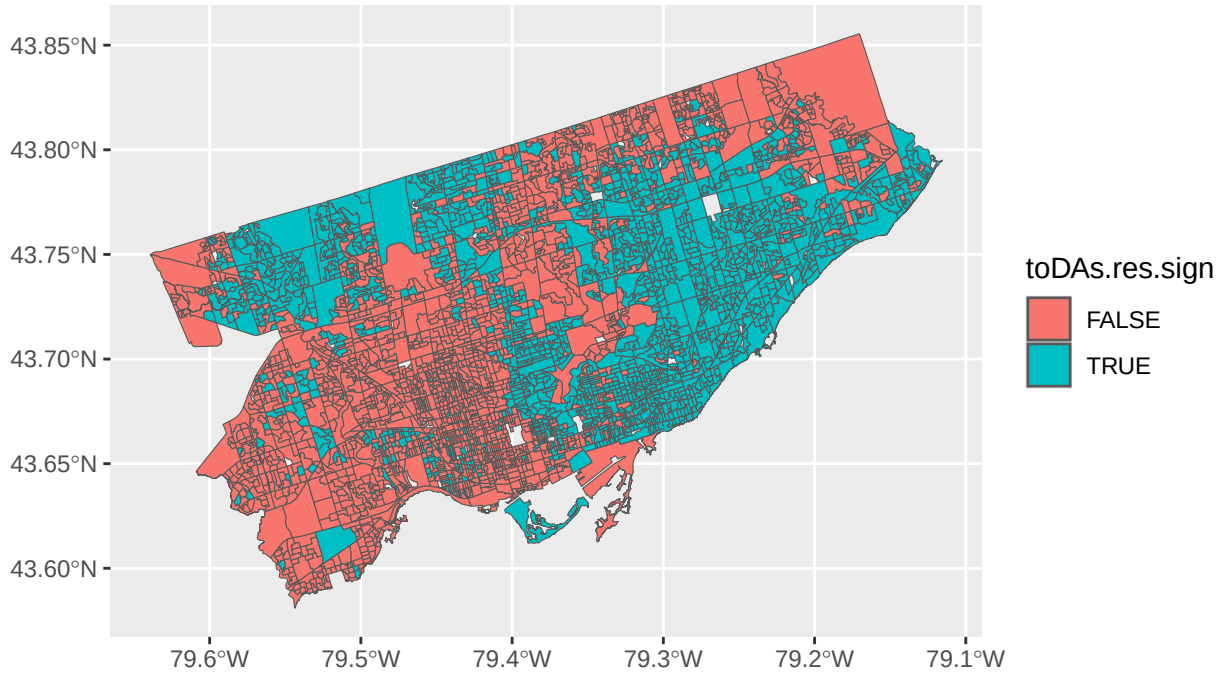
To further investigate the relative importance of each covariate, linear regression was performed using all eight covariates. It is also possible that this might result in a better fit than models involving only single covariates.

```
##
## Call:
## lm(formula = P ~ population_density + minority_proportion + median_age +
##     median_hh_income + automobile_prop + public_tr_prop + walking_prop +
##     cycling_prop, data = toDAs.NAomit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.33640 -0.17765 -0.01834  0.16016  2.50336
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          7.051e+00  4.948e-02 142.520  < 2e-16 ***
## population_density  -4.722e-07  2.383e-07  -1.981  0.04762 *
## minority_proportion  6.684e-01  2.225e-02  30.039  < 2e-16 ***
## median_age           7.913e-03  8.249e-04   9.593  < 2e-16 ***
## median_hh_income     3.541e-07  1.250e-07   2.832  0.00465 **
## automobile_prop      7.975e-01  8.085e-02   9.863  < 2e-16 ***
## public_tr_prop      -4.072e-01  1.287e-01  -3.164  0.00157 **
## walking_prop        -3.683e+00  2.314e-01 -15.914  < 2e-16 ***
## cycling_prop        -5.215e+00  4.412e-01 -11.819  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.29 on 3666 degrees of freedom
## Multiple R-squared:  0.4841, Adjusted R-squared:  0.483
## F-statistic:   430 on 8 and 3666 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = P ~ minority_proportion + median_age + automobile_prop +
##     walking_prop + cycling_prop, data = toDAs.NAomit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.32615 -0.17306 -0.01627  0.16371  2.49662
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          7.0361088  0.0415344  169.40   <2e-16 ***
## minority_proportion  0.6154436  0.0200923   30.63   <2e-16 ***
## median_age           0.0088286  0.0007926   11.14   <2e-16 ***
## automobile_prop      0.8531586  0.0800063   10.66   <2e-16 ***
## walking_prop        -3.8315499  0.2294299  -16.70   <2e-16 ***
## cycling_prop        -5.3605210  0.4406873  -12.16   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2912 on 3669 degrees of freedom
## Multiple R-squared:  0.4795, Adjusted R-squared:  0.4788
## F-statistic: 675.9 on 5 and 3669 DF,  p-value: < 2.2e-16
```

Despite using only the covariates with the lowest p-value, the Adjusted R-squared value decreases slightly and the residual standard error increases. This indicates that reducing the number of variables in the model does not in fact improve the fit. Therefore, for general prediction it might be best to use all 8 of the selected variables.



Words

Methods

Describing some methods.

Results and discussion

Here we will discuss the results of the analysis.

References