Metro Vancouver Dissemination Area Income and Pollution of PM2.5 Case Study

Geog 418: Devin Hewett V00821219

1 Introduction

Vancouver, British Columbia is home to some of Canada's richest and poorest neighbourhoods located within a very close proximity to one another. As the city continues to grow the risk of poor and harmful air quality increases[1]. A measurable indicator of air quality is $PM_{2.5}$, which refers to atmospheric particle matter (PM) that have a diameter of less than 2.5 micrometers[2]. $PM_{2.5}$ is either emitted directly into the air usually from a combustion source such as power plants, motor vehicles, wood burning, forest fires or agricultural burning, or it is formed when gas and particles interact with one another[2]. For instance, sulfur dioxide gas emitted from a combustion reaction reacts with oxygen and water vapor to form sulfuric acid as a byproduct[2]. Exposure to $PM_{2.5}$ is known to cause lung damage[1] and, according to a study published by The Journal of American Medical Association, can cause plaque deposits in the arteries which may lead to vascular inflammation and hardening of the arteries, increasing the risk of heart attack and stroke[3]. Income segregation has been measured by looking at variable factors such as median household income, number of parks and proximity to services[4]. However, there is much less research looking into the relationship between air quality and neighbourhood income. A study in The Canadian Medical Association Journal showed that in Greater Hamilton, Ontario, mean pollutant levels tended to be higher in lower income areas and both low income and high pollutant levels were associated with increased mortality[1]. Thus, given a neighbourhood with high pollution levels and low income they are more prone to mortality caused by harmful air quality illnesses[1]. $PM_{2.5}$ is measured in Standard ($\mu g/m^3$) and air quality is considered good between 0-12, moderate between 12-35.4, poor between 35.5-55.4 and unhealthy when greater than 55.5[5]. This report attempts understand which neighbourhoods in Vancouver are the most susceptible to exposure to these types of pollutants and could help inform polices in order to reduce harmful effects in these areas[4]. The first step would be to determine if income is spatially segregated in Metro Vancouver neighbourhoods by testing the characteristic spatial pattern. Then, determine if the spatial variability in pollution level explains the spatial variability in income by conducting regression analysis. The residuals of the regression could be analyzed to test the model and determine if they are different from random. Then lately, validate the sampling strategy used for collecting the $PM_{2.5}$ measurements.

The objective of this study is to determine if Metro Vancouver's neighbourhoods are spatially segregated by income. Then determine if the $PM_{2.5}$ levels explain the spatial variability in income and validate the findings by testing if the pollution samples are clustered or random across the study area.

2 Methods

2.1 Study Site and Data

The study site is the Metro Vancouver area (Figure 1), located in Southwest British Columbia mainland next to the Pacific Ocean. The site consists of very dense urban areas like Vancouver, Burnaby, and Surrey, accompanied with surrounding sub-urban areas and the Coastal Mountain range to the north. The area is known to consist of a wide range of wealth, ethnicities, industry, and development[4]. Also, due to Greater Vancouver's continued population growth, increasing industry, and the effect of local wildfires air quality is becoming of greater concern[6]. There are three main datasets used in this report. The first was a CSV file of Income data collected from the 2016 Canadian Census at different dissemination area levels. The second dataset used was a shapefile of 2016 dissemination area boundaries from the Canadian Census; the file was projected in NAD83 BC Albers and contained 7582 features, 3363 of which intersects with the study area and was used in the analysis. The final data set was a CSV file from The Canadian Urban Environment Health Research Consortium (CANUE) of $PM_{2.5}$ measurements in Standard ($\mu g/m^3$) simulated with respect to postal codes.

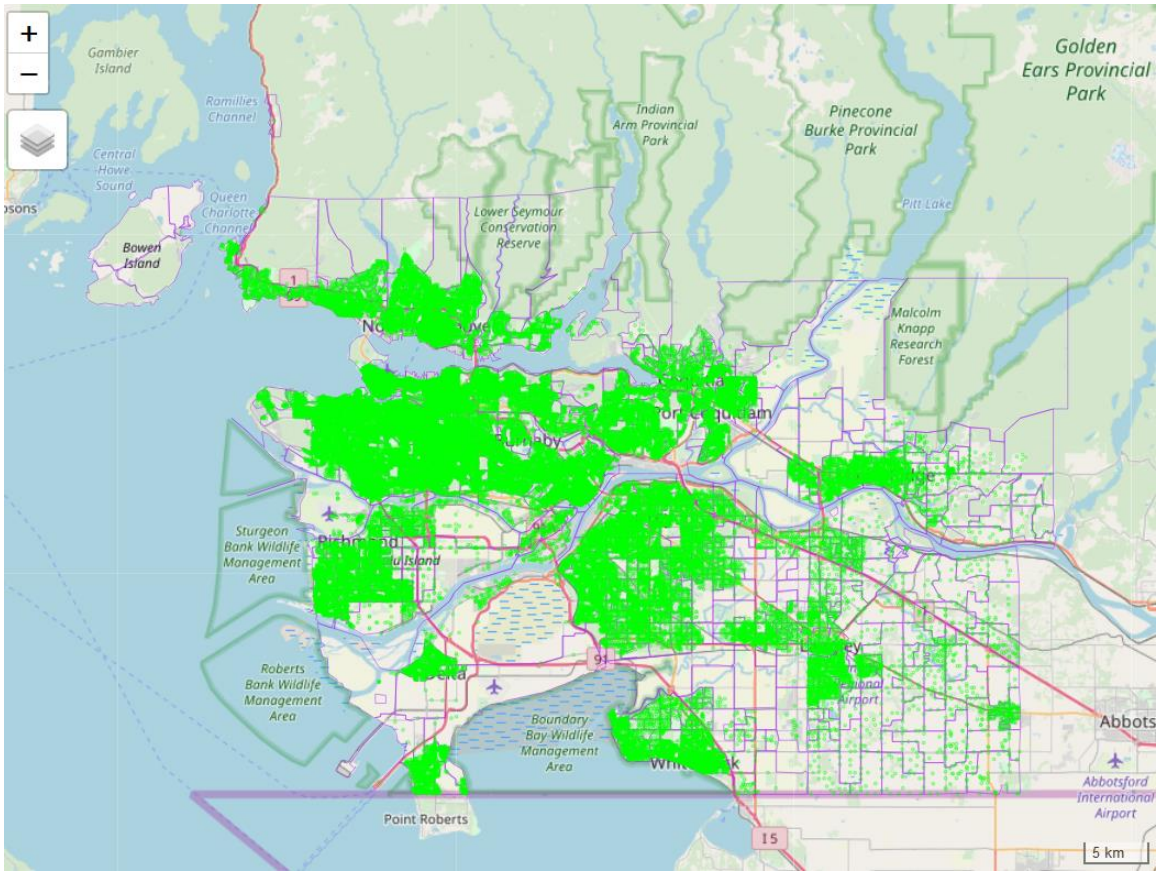


Figure 1: Metro Vancouver Study Site, with $PM_{2.5}$ sample points in green and census tracts outlined in purple.

2.2 Descriptive Statistics

To better describe the data that was used in the analysis, descriptive statistic can be performed. First, in order to visualize the income levels of different dissemination areas, the median income was used and plotted in a choropleth map for each neighbourhood. Then the mean value (Equation 1) of the $PM_{2.5}$ measurements in each neighbourhood were taken and a choropleth map was construction for visual interpretation. Median was used instead of mean for income as the distribution of the income data was quite negatively skewed, therefore the median better represented the population. Histograms of both the median income and $PM_{2.5}$ concentrations were created to better interpret the distribution of the data.

$$\bar{X}_i = \frac{\sum_{i=1}^{n} X_i}{n_i}$$

Equation 1: Calculation of the Mean.

2.3 Spatial Autocorrelation of Income

In order to compare the different incomes amongst different dissemination areas, all the neighbourhoods must be related somehow. This is where a neighbourhood weight matrix comes into play, defining the relationship between neighbourhoods. Since the number of neighbours is large, the queen's method of creating the matrix was used as it has the possibility of connecting up to eight neighbours. Next, a choropleth map of lagged means which shows how each neighbourhood differs from the mean was plotted. This is a good way of determining outliers in the data, as observations that are more different from the mean are impacting the statistic more than those closer to the mean. Next, a Global Moran's I, which determines the degree of autocorrelation between Income between the different areas of Vancouver was calculated. Global Moran's I compares each census tract with the global mean and then compares that difference with each neighbour's variation from the mean (Equation 2). A Z-Test was used if we can say with a certain confidence the data is significantly different from random (Equation 3). Then to compare, a local Moran's I test was conducted. Local Moran's I is very similar to global, with the difference being that instead of using the entire study site, each region only compares each value to the mean of its neighbours (Equation 4). Local indicators of spatial association (LISA) was mapped to understand which areas are positively or negatively autocorrelated. Then all regions income was compared in a scatter plot to spatially lagged income values to observe the general trend of autocorrelation for the entire site.

$$I = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\left(\sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j}\right)\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

Equation 2: Global Moran's I Calculation.

$$Z_I = \frac{I - E_I}{s^2}$$

Equation 3: Z-Statistic Calculation.

$$I_i = \frac{x_i - \bar{X}}{S_i^2} \sum_{j=1}^{n} w_{i,j}(x_j - \bar{X})$$

Equation 4: Local Moran's I Calculation.

2.4 Regression Analysis

Linear regression of the variables was performed in order to test if the spatial variability in $PM_{2.5}$ explained the spatial variability in income. The ordinary least squares (OLS) regression fits a linear line (Equation 5) which minimizes the squared sum of residuals. A P-value was produced to be able to reject or accept the null hypothesis that the $PM_{2.5}$ does not help explain income. The R^2 value was then calculated to determine how well the model fits the data and residuals, which is the difference of observed points with the model itself. We assume using the OLS model that the data is normally distributed, will fit a linear line, residuals are indeed errors and they are independent of each other. To test this, first a map of the residuals was plotted then Local and Global Moran's I was conducted on the residuals to see if they are random or clustered. Finally, a geographically weighted regression (GWR) was conducted. While very similar in purpose to the linear regression, this method is a bit more complex and instead takes a measure for every individual location. The GWR used a bandwidth method of AICc and an adaptive kernel type. Then both the resulting r-squared values and coefficients were mapped.

2.5 Point Pattern Analysis

The last step was to determine if the sample structure of air quality measurements were randomly distributed or not. First, Nearest Neighbour Distance (NND) point pattern analysis was performed. The study area's window was first defined from the sample point's max and min longitude and latitude of 49.00192 to 49.4767°N and 122.40864 to 123.34592°W. From that, mean nearest neighbour distance, the NND for a completely random distribution, the R value, and finally the z value are calculated to determine the distribution. The k-function for the sample dataset will be determined as the k-function, a ratio between the number of points within a distance from a random point and the density of the study area. The study area again defined by the bounds of furthest latitudes and longitudes in each direction were broken up into 100 quadrats and the frequency of each quadrat was counted. The VMR and chi-squared were both determined to understand the sampling structure of the samples across the site.

3 Results

The resulting Choropleth map showed the median income value for dissemination areas in Greater Vancouver with a minimum median yearly income of $11504 and a maximum of $79872 (Figure 2). Areas of lower median income appear to be located in dense city centres, where areas of high income are found in close proximity to the ocean and in sub-urban areas (Figure 2). The mean $PM_{2.5}$ concentration for each neighbourhood is shown by the choropleth map (Figure 3) which has a maximum mean concentration of 57 $\mu g/m^3$ and a minimum of 1. The histogram of median income (Figure 4) and mean $PM_{2.5}$ concentration (Figure 5) show the distribution of different ranges and how many neighbourhoods fall under which range.

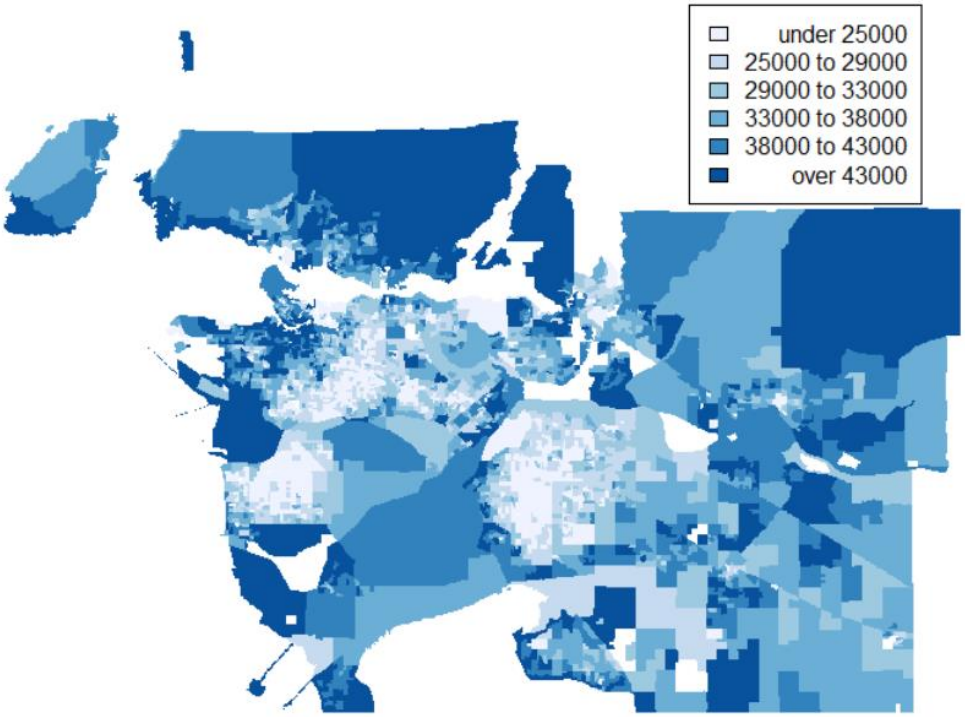


Figure 2: Choropleth Map of Median Income for Neighbourhoods in Metro Vancouver.

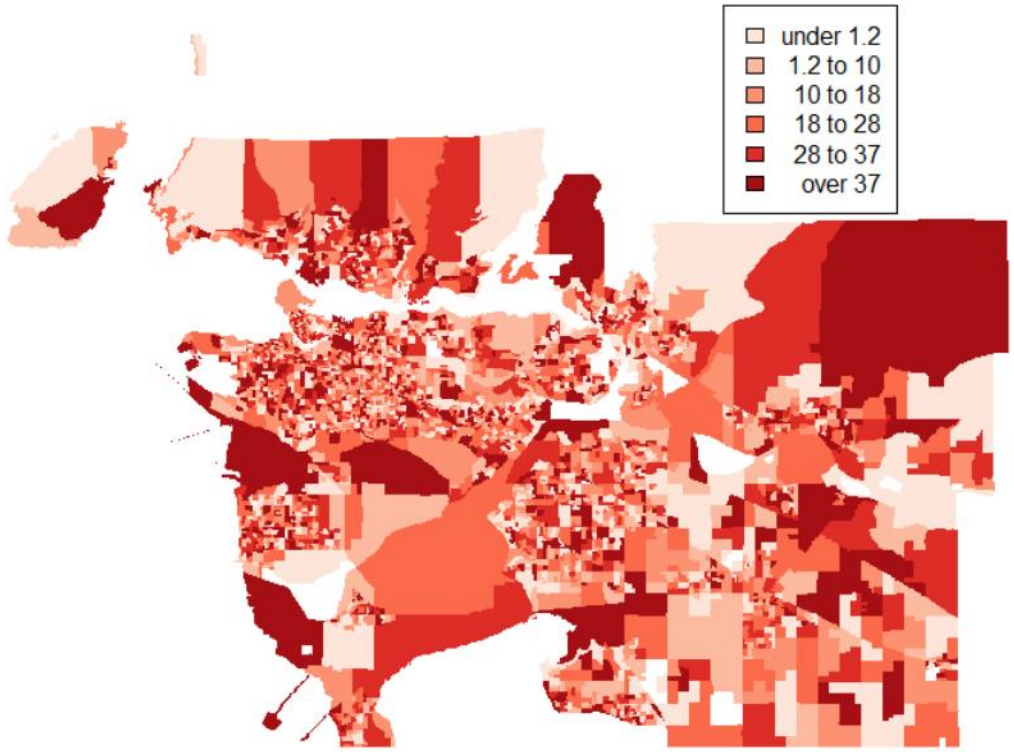
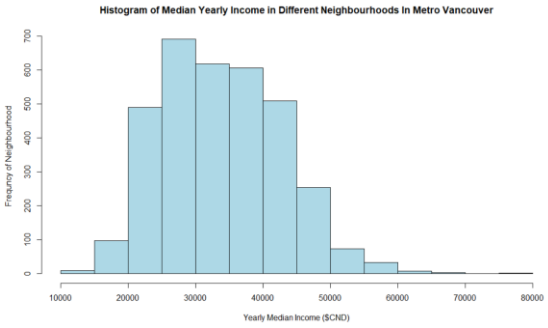


Figure 3: Choropleth Map of Mean $PM_{2.5}$ ($\mu g/m^3$) for Neighbourhoods in Metro Vancouver.



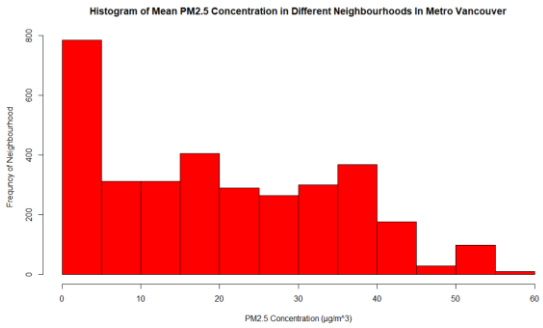


Figure 4: Median Income of Neighbourhoods. **Figure 5:** $PM_{2.5}$ for Metro Van Neighbourhoods.

The Resulting Queens method matrix which relates all the neighbours to one another (Figure 6) in order to test for spatial autocorrelation contains 3363 regions, 21062 nonzero links, and has an average link number of 6.26. The plotted lagged means for median income in the neighbourhoods (Figure 7) showed areas that differ drastically from the mean as a darker colour gradient. Areas close to city centres seem to have less variation then those that are further (Figure 7). The global Moran's I calculated a statistic standard deviate of 67.539 and a p-value of less than $2.2e^{-16}$. The Moran's I statistic was 0.6789, with an expectation of -0.000297 and a

variance of 0.000101. The potential Moran's range was between -1 and 1.051804 and the Z statistic calculated was 6715.8, which is an extremely high value and indicates income is clustered with a confidence interval of 95% or a p-value less than 0.05. The LISA values map for the resulting local Moran's I (Figure 8) shows an overwhelming number of regions are positively spatially autocorrelated with a few areas being negatively autocorrelated. The scatterplot was fitted with a trendline that showed there was a positive autocorrelation trend between income and different neighbourhoods in Metro Vancouver (Figure 9).

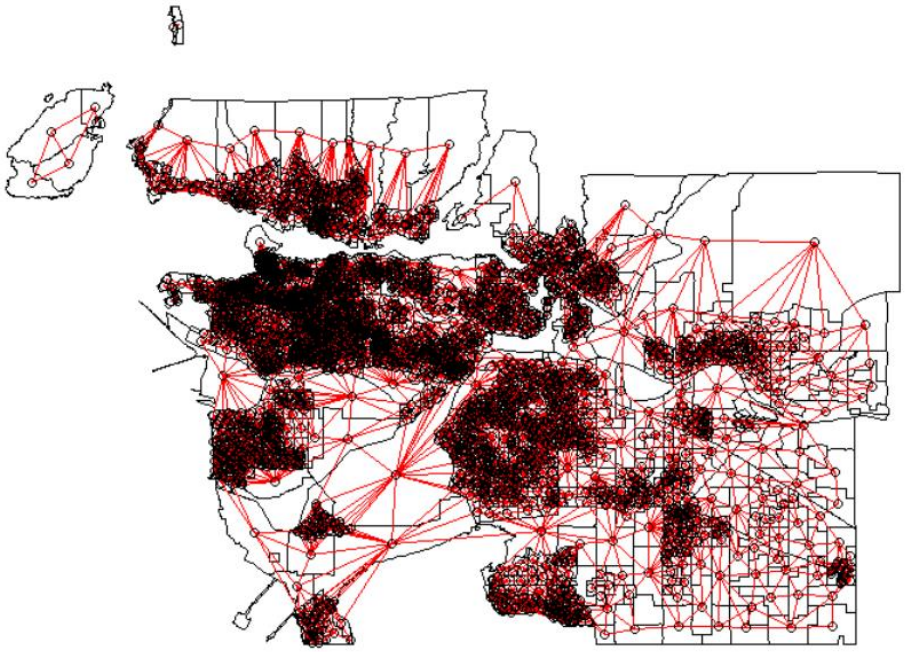


Figure 6: Queens Method Neighbourhood Matrix of Metro Vancouver.

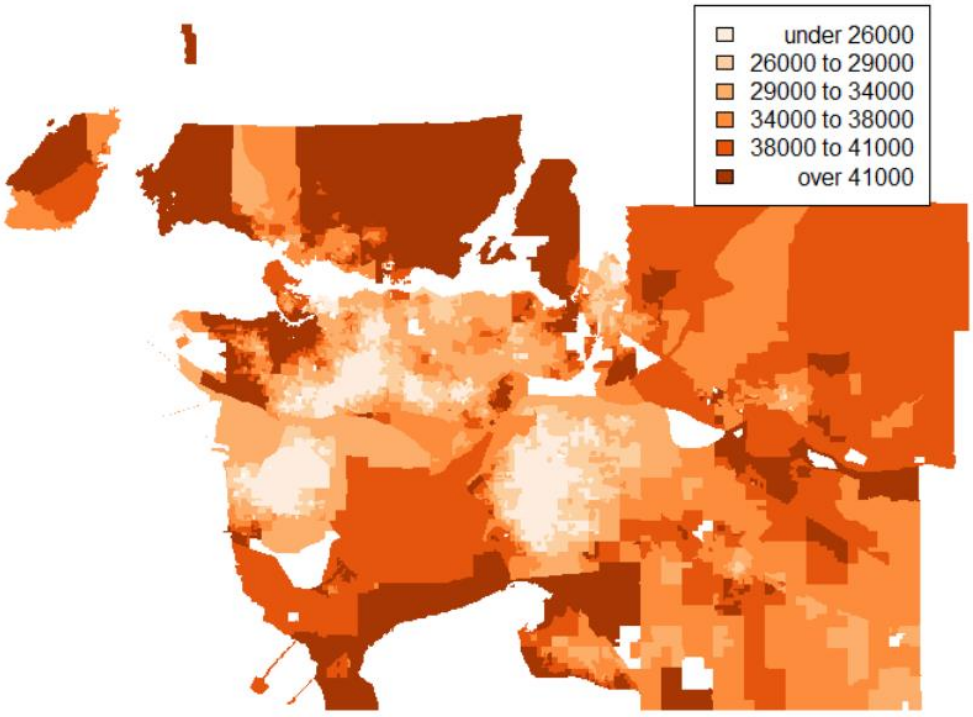


Figure 7: Lagged means of Income in Metro Vancouver Neighbourhoods.

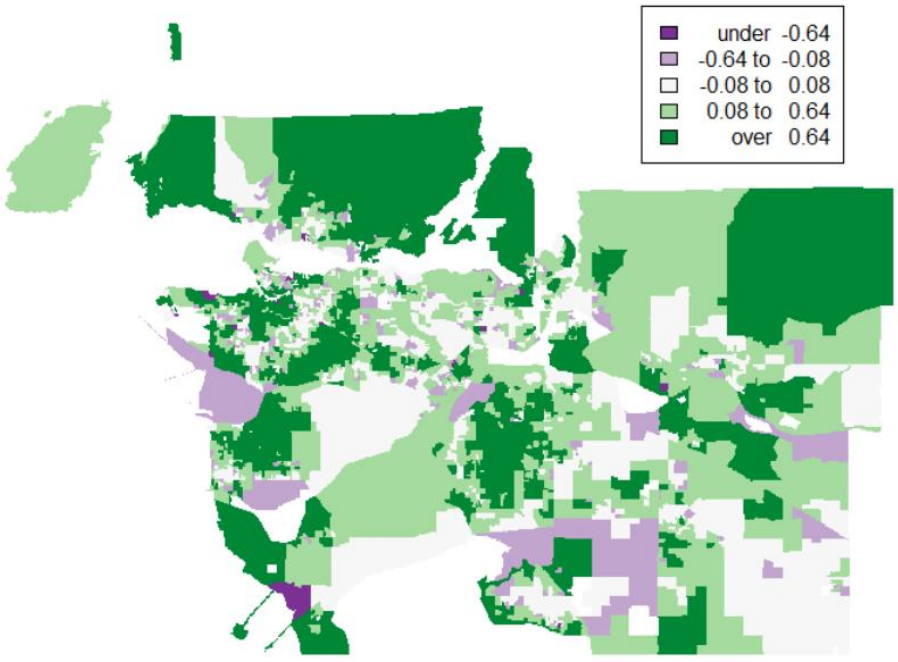


Figure 8: Map of Income Moran's I's LISA Values For Greater Vancouver Neighbourhoods.

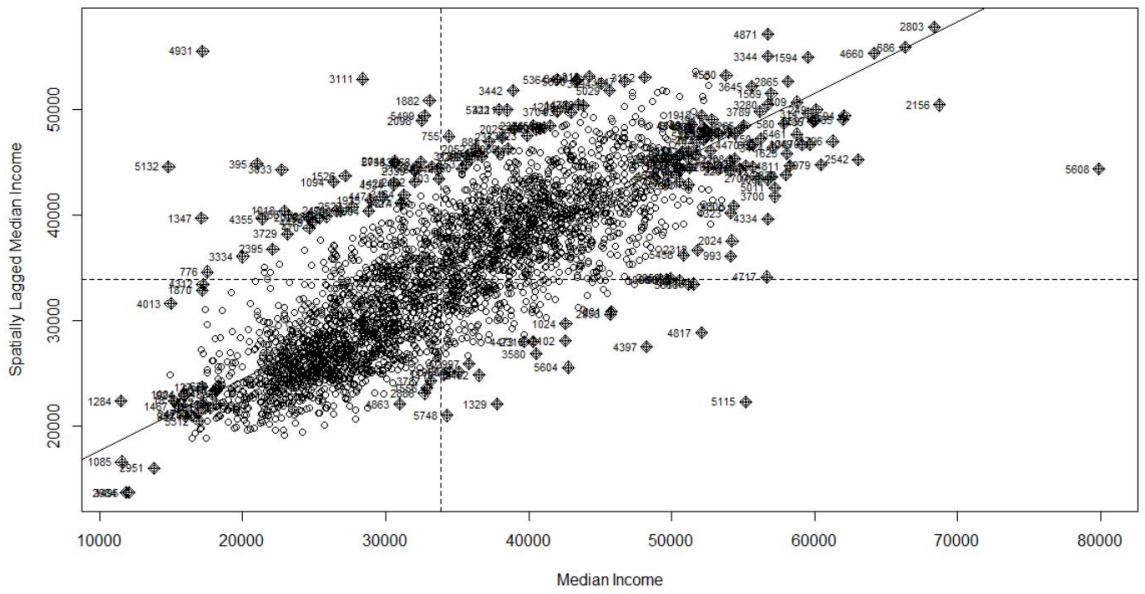


Figure 9: Scatterplot of Spatially Lagged Median Income Vs Median Income in Greater Vancouver Neighbourhoods.

The linear regression has resulting residuals with a min of -24452, a median of -761 and a max of 44407. The model has an intercept of 30603.96 with an error of 332.7 and a slope of 134.32 with an error of 12.26. The R^2 value was 0.0409 with a p value less than 2.2e-16, meaning that there is somewhat of a dependence of income on PM2.5 with a high confidence interval. The map of the residuals plotted on a choropleth map (Figure 10) was used to visualize where error is located. The global Moran's I on the residuals resulted with a Moran's I statistic of 0.66769, a z-statistic of 5437.112 and p value of less than 2.2e-16 which with high confidence the residuals are different than random and are indeed clustered. The LISA values map of plotted Local Moran's I (Figure 11) and the scatterplot (Figure 12) of residuals both show a heavy positive autocorrelation, meaning that the residuals are clustered. The GWR resulted in a quasi-global r^2 value of 0.794 and a residual sum of squares of 47124331198. The map of the r squared values (Figure 13) and coefficient values (Figure 14) show spatially where the values appear.

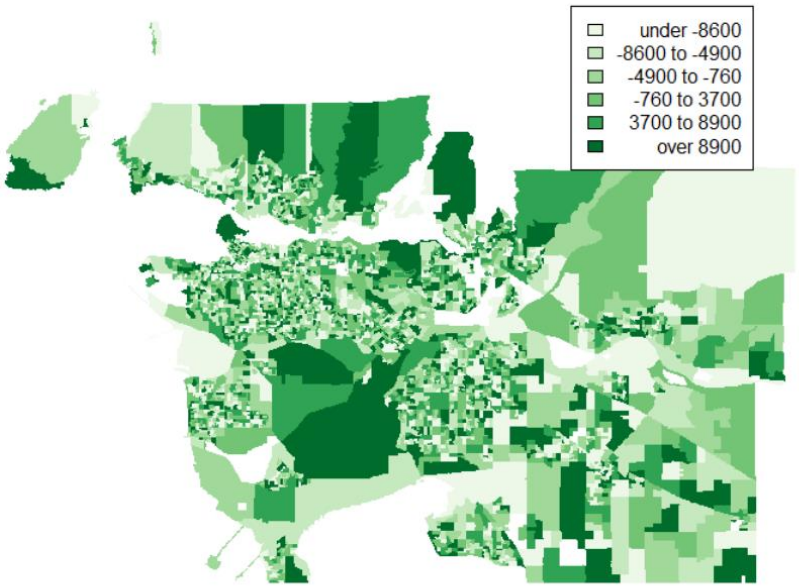


Figure 10: Residuals from Linear Regression of Metro Vancouver Neighbourhoods.

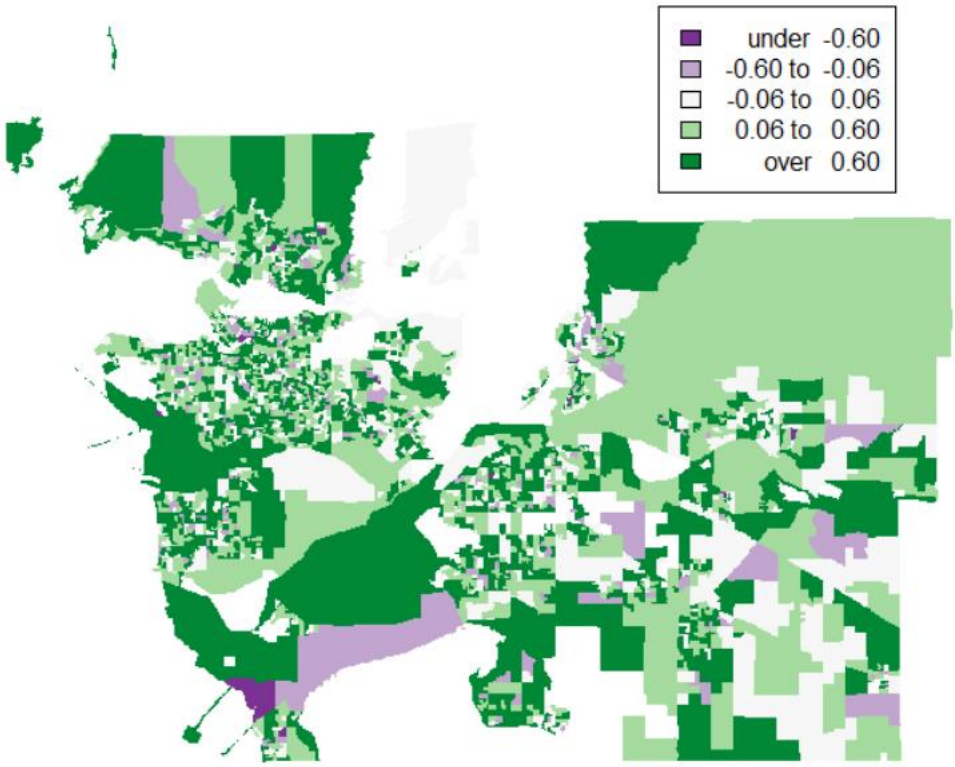


Figure 11: Map of Residuals Moran's I's LISA Values for Greater Vancouver Neighbourhoods.

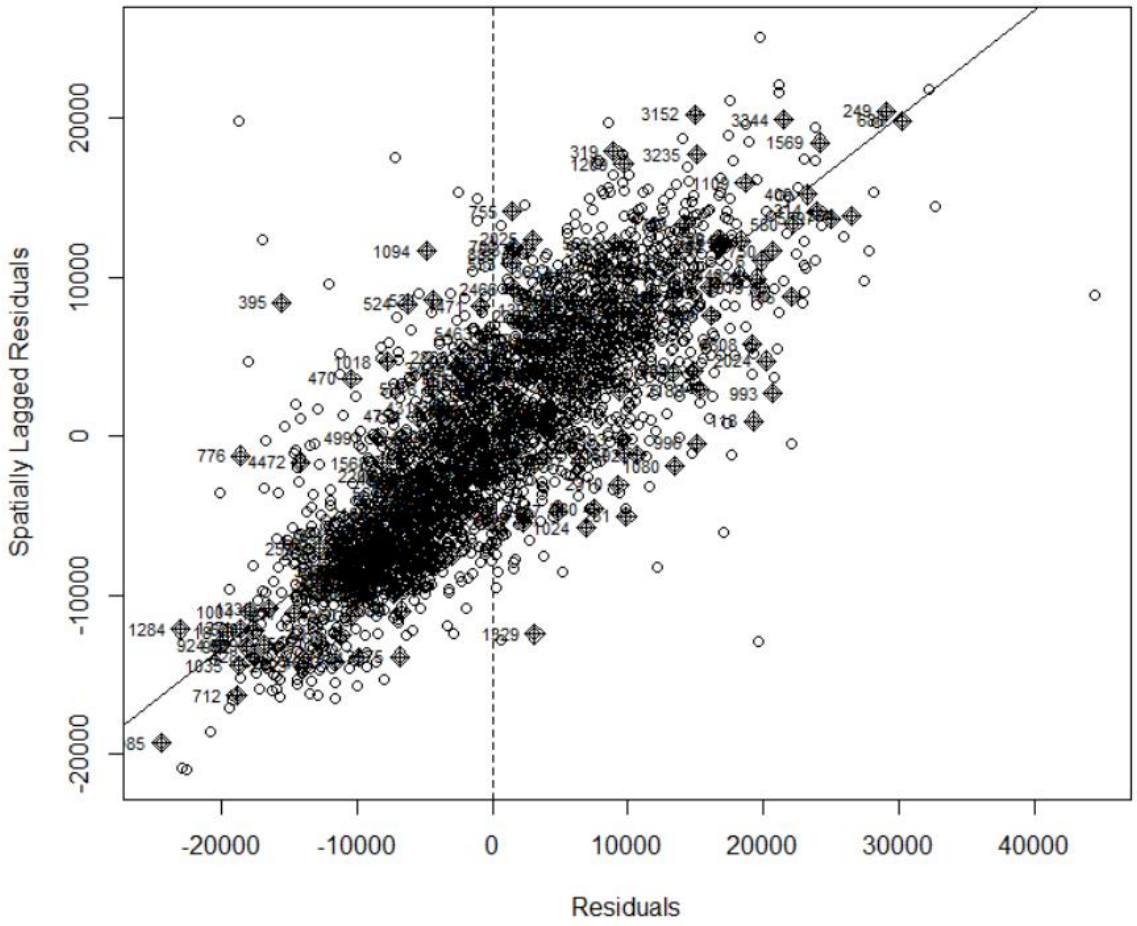


Figure 12: Scatterplot of Spatially Lagged Residuals vs Residuals in Greater Vancouver Neighbourhoods.

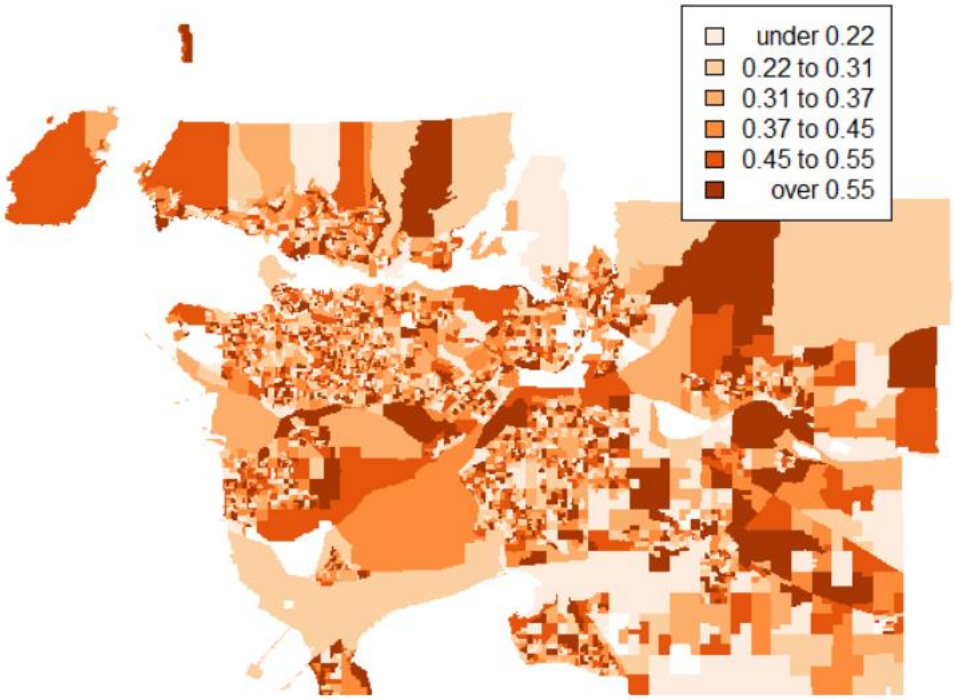


Figure 13: Choropleth map of GWR R-Squared values in Metro Vancouver.

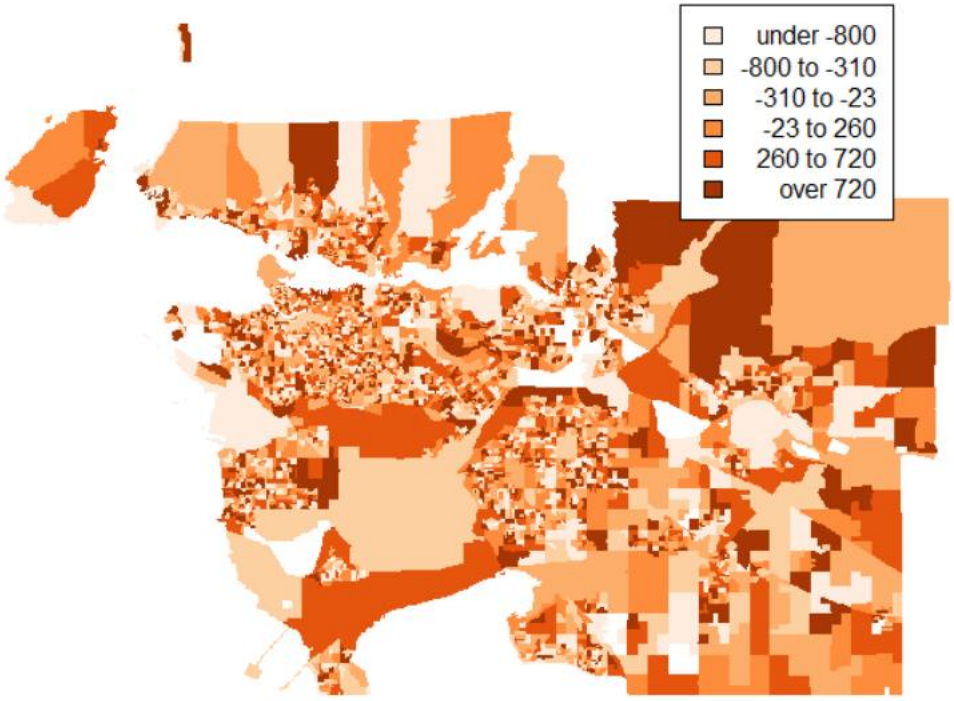


Figure 14: Choropleth map of GWR Coefficient values in Metro Vancouver.

The result of analysis of the distribution type of the air quality measurements from NND was a mean NND of 0.004105°. If the points were totally random, the NND would be 0.006293°. The normalized r value was 0.6533, which resulted of a z score of -35.20061. This means that with high confidence the sample points trend is clustered. The result of the K-function (Figure 15) shows a buffer around the red line, which represents 99% confidence. Therefore, it is significantly different than random and indeed is clustered. The result of the quadrat analysis (Figure 16) is a VMR value of 128.3 and a chi-squared value of 13691.8. With a degree of freedom of 2817, this results in a very low p value, less than 0.00001. Because of the high VMR and low p value, the sample points are again proven to be clustered.

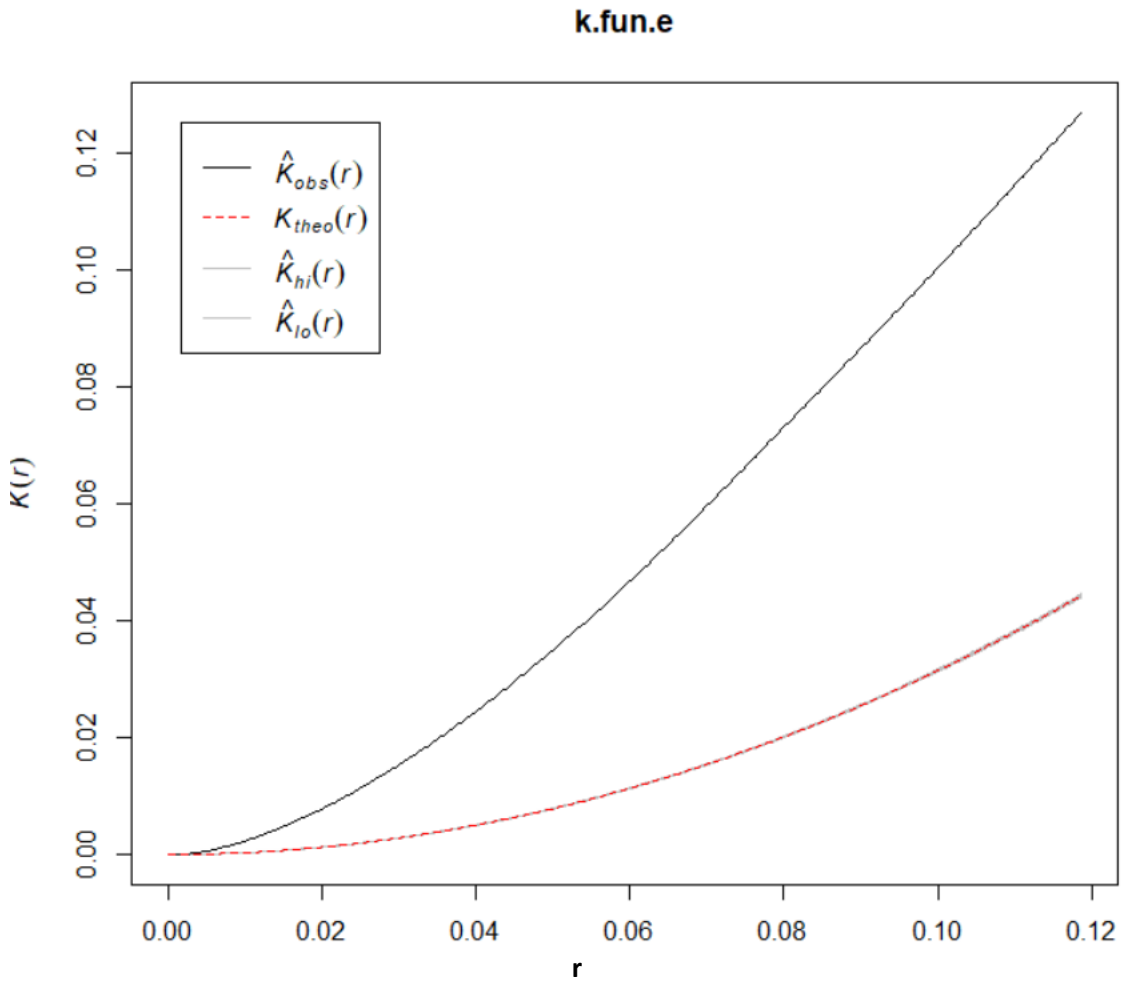


Figure 15: K-Function of air quality sample point. Black line is the data, red line is random distribution and the buffer around is a confidence of 99%.

0	1	0	0	0	0	0	0	0	0
2	0	2	0	0	0	0	0	0	0
10	29	57	30	1	0	0	0	0	0
0	9	122	58	4	1	11	2	1	0
0	118	363	162	51	19	67	2	3	1
0	29	202	188	77	70	29	69	41	5
0	47	103	5	55	132	29	35	10	3
0	42	44	1	33	93	51	53	8	3
0	0	6	12	0	10	3	53	12	18
0	0	33	7	0	64	1	7	5	4

Figure 16: Quadrat Analysis Map of Air Quality Sample Points.

4 Discussion

The report studied ground level $PM_{2.5}$ concentrations from samples site neighbourhoods in Metro Vancouver and the annual median income in the same dissemination areas to determine if the different neighbourhoods are spatially segregated by income and if income is dependent on air quality between neighbourhoods. In general, from the results described above, the income between neighbourhoods are overwhelmingly positively spatially autocorrelated. From the regression analysis there appears to be a small dependence of income on $PM_{2.5}$ concentration. The spatial analysis of the residuals produced from the regression were strongly positively autocorrelated in nature. The GWR showed that the data fit the regression model well and produced values for coefficients and r-squared values for each of the neighbourhoods. The point pattern analysis determined that the sample technique conducted for collecting $PM_{2.5}$ was significantly different than random.

Therefore, based on the analysis in this report, income is very spatially segregated amongst Greater Vancouver neighbourhoods. $PM_{2.5}$ does explain the spatial variability in income in the area. However, the residuals or model errors of the regression are extremely clustered, meaning the model does not fit the data for specific areas. The validation of the sample technique for collecting $PM_{2.5}$ values are also very clustered in nature, meaning the samples were not randomly taken; they followed a supervised sampling model and definitely were not randomly stratified.

The resulting descriptive statistic showed that the median income for neighbourhood close to the dense city centres were more likely to be low (Figure 2). The spatial autocorrelation via the queen's method model (Figure 6) resulted in a global Moran's I z-statistic of 6715.8, affirming that income was indeed segregated (Figure 8) with an extremely high positive autocorrelation (Figure 9). The linear regression compared if income was spatially dependent on $PM_{2.5}$. It resulted in an R^2 value of 0.0409 with a p value less than 2.2e-16. This means that the with a small positive R^2 that there is a small trend of dependence with a very significant confidence above 99%. Testing if the linear regression model fits the data by testing the spatial autocorrelation of the residuals (Figure 10) showed that the residuals were cluster spatially (Figure 12) meaning that specific groups of neighbourhoods in the study site did not fit the model well (Figure 11). The GWR uses local neighbours as it better fits a model regression for each neighbourhood. The neighbour's coefficients (Figure 14) and r values (Figure 13) resulted in a global r-squared value of 0.794, meaning that the model reasonably fit the data. Finally, the points where the $PM_{2.5}$ was sampled from the NND analysis produced an r-value of 0.6533 and z score of -35.2, showing that the sample was spatially clustered. The k-function showed that the points were significantly different than random (Figure 15) and the quadrat analysis with VMR of 128.3 and chi-squared of 13691.8 confirmed that conclusion (Figure 16).

The results of this case study were opposite of the Canadian Medical Association Journal study in Hamilton[1]; this discrepancy in results could be due to a number of factors. As $PM_{2.5}$ can change drastically with weather conditions and wind[2], daily measurements in the same areas could fluctuate greatly. Therefore, in order to improve this study, collection and analysis should be repeated over a longer time frame, exploring the different factors which could influence $PM_{2.5}$. Ultimately comparing and averaging data over time to gain a more informed result. Since the sample points were spatially clustered, a new strategy for collection should be used, perhaps a stratified random sampling method. Since the residuals from the regression were clustered, a look at better regression models would also improve the validity of the analysis. Overall, new policies could be put into place in areas of higher income to help reduce the pollution in those areas and prevent long term exposure[3]. As the city continues to grow, people with lower income often have less resources to help protect themselves[4]. Therefore, both low and high income households should make this a high priority and actively help reduce pollution together.

5 References

1. Finkelstein, Murray M., et al. "Relation between income, air pollution and mortality: a cohort study." *Canadian Medical Association Journal* 169.5 (2003): 397-402.
2. Ji, X., Yao, Y., & Long, X. (2018). What causes PM2.5 pollution? cross-economy empirical analysis from socioeconomic perspective. *Energy Policy, 119*, 458-472. doi:10.1016/j.enpol.2018.04.040
3. Pope III, C. A., Burnett, R. T., Thun, M. J., Calle, E. E., Krewski, D., Ito, K., & Thurston, G. D. (2002). Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *Jama, 287*(9), 1132-1141. doi:10.1001/jama.287.9.1132
4. Ley, D. (2012). *Divisions and disparities in lotus-land: Socio-spatial income polarization in greater vancouver, 1970-2005*. S.l.: Cities Centre, University of Toronto.
5. Kang, D., Mathur, R., & Rao, S. (2010). Real-time bias-adjusted O-3 and PM2.5 air quality index forecasts and their performance evaluations over the continental united states.*Atmospheric Environment, 44*(18), 2203-2212. doi:10.1016/j.atmosenv.2010.03.017
6. Rolfe, C., Klenavic, N., Canadian Electronic Library (Firm), & West Coast Environmental Law Research Foundation. (2005). *Strengthening the GVRD air quality management plan: Submissions to the greater vancouver regional district regarding the 2005 draft air quality management plan*. Vancouver, B.C: West Coast Environmental Law.