# Aggregation and analysis of $PM_{2.5}$ exposure prenatally

Jonathan Viswasam

August 17, 2013

**Abstract**

In this project, data was retrieved from the state of Massachusetts about the level of pollutant data, specifically $PM_{2.5}$ (particulate matter with a diameter of less than 2.5$\mu$m), and demographic information of pregnant mothers. The effects of $PM_{2.5}$ are well known to adults but there have not been conclusive studies detailing the effects on newborns from prenatal exposure. There are massive datasets which contain geographic district zones with $PM_{2.5}$ data for a nine year period. Using MATLAB code, the average exposure level for each person by weeks and trimesters were calculated from algorithms that mapped the person to the zone, and then to the $PM_{2.5}$ value, dealing with any missing data entries and analysis of data. While this work is part of a larger project, some conclusions from the analysis include distance to roadway having little effect on $PM_{2.5}$ exposure, and that most of trimesters are not correlated with each other except trimesters 1 and 3, which exhibited a moderate correlation.

## 1 Introduction

The work this past summer is part of a much larger project involving professors at the University of California - Irvine: Dr. Bartell and the project leader Dr. Veronica Vieira. The critical purpose of this project is to find the relationship between prenatal $PM_{2.5}$ (fine particles that come from certain industrial activities with a diameter of less than 2.5$\mu$m) exposures and outcomes such as bronchiolitis and otitis media. The relationship is understood for adults but the degree of severity is unknown for infants.

Bronchiolitis, one of the outcomes mentioned above, is a lower respiratory tract infection that is the most frequent cause of hospitalizations for children during their first year of life. It is caused more frequently by respiratory syncytial virus but is also caused by exposure to environmental tobacco smoke and indoor wood burning. Otitis media, the second outcome mentioned, is an inflammation of the middle ear which occurs predominantly in infants. Otitis media costs 3.8 billion dollars in the United States annually and contributes to stunted hearing and cognitive development. Both bronchiolitis and otitis media

have similar risk factors such as low socioeconomic status, short breast feeding durations and genetics. [2]

There are only a few studies done that address this question, and only for one specific outcome. There was a study in Southern California looking at the $PM_{2.5}$ exposure effect on bronchiolitis. The study reported that there was little to no effect that $PM_{2.5}$ affected bronchiolitis, but the study urged greater investigation and study.

For otitis media, there was a study using a birth cohort from Germany and the Netherlands that reported a slight positive association with $PM_{2.5}$ and otitis media, citing a odds ratio of 1.13 with a 95 percent confidence interval. But the study did not look at the severity of OM. [1]

There were two attributes from this project that allow the innovation to build off the previous pieces of literature. The first method was the remote sensing data. This allowed access to daily measurements of aerosol optical density (AOD). The AOD is correlated with the PM2.5 concentration given this equation:

$$AOD = \frac{PM_{2.5}Hf(RH)3Q_{ext,dry}}{4\rho * r_{eff}}$$

Where H is the height of the satellite, f(RH) is the ratio of the extinction and dry coefficients, and $r_{eff}$ is the radius of the particles. With this satellite data, they were able to measure the pollutant level for $5,359$ geographic zones at each day for nine years from $2001 - 2009$. The second level of uniqueness involved in this project is the use of the PELL (Pregnancy to Early Life Longitudinal) dataset. The PELL dataset contained demographic information of the pregnant mother such as conception, geographical zone, gestational period, and their addresses' distance from the road. [2]

# 2 My Work

As stated above, the PELL dataset was used for this project. With nine years of data, there were some problems actually visualizing the data at first. Upon importation into Microsoft Excel, an error message relayed that the row limit had been exceeded. If the first file, 2001, was exhibiting trouble, then the eight following years would also yield problems. The next logical step involved importing directly onto MATLAB's workspace through its import commands; this displayed similar problems as Excel. Where Excel's limit was $1,048,576$, the limit of the MATLAB workspace viewing window was $65,576$. Although the window size was smaller, all of the entries were able to be accessed and operated on in MATLAB so the importation problem was dealt with.

Although importation had been taken care of, there arose a new problem. There was missing data in different parts of the demographic information - conception date, geographic id zone, and the pollutant levels. Additionally, the information were in different data types, so the data had to be treated differently. Using the command textread, different data types were able to be

imported, accessed and operated on by assigning the string format to the dates with slashes, e.g. 1/1/2001, or assigning numbers with many digits floating point format to prevent overflow.

With the data imported, there was the task of mapping a pregnant mother's geographical id to the corresponding demographic file, and matching the date of conception with the date in the pollutant file. There posed another issue, the correspondence of the dates proved difficult as written in the slash format. While textread was able to import the characters which contained slashes, operating on them was unsuccessful. Therefore, the command datenum was extensively used in order to convert the mm/dd/yyyy format into a serial number with the units in days. In MATLAB, the default reference year is a hypothetical date 1/1/0000, but this was of no consequence because the dates would still be equal as long as datenum was used consistently throughout the code.

After the mapping code was created,the $PM_{2.5}$ values for each person would be added then averaged for their entire gestational period, but one problem mentioned earlier had to be dealt with first - there were still missing pollutant data. We were unable to use values associated with the missing demographic information as there is no effective method to predict the conception date or geographic zone. But for the missing $PM_{2.5}$ values, we used linear regression in order to impute the missing values. Using regression, we had averages for each of the months during the nine year period. Our code would check which month the missing value was in and then substitute it in.

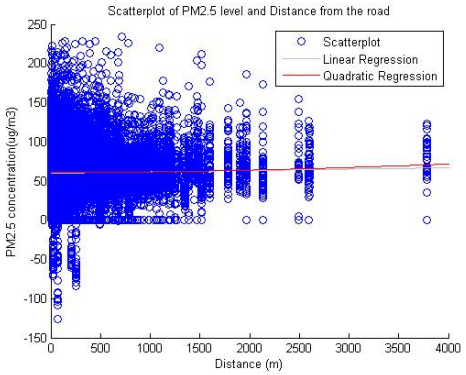
# 3 Results

While using regression, it is necessary to have multiple covariates, variables that serves as predictors of the outcome. In this case, the outcome would be the $PM_{2.5}$ exposure. The first covariate added in the regression was months as the pollutant level changes throughout the months. But the month must be treated as a categorical variable as they are up to 12 different months. Initially, the regression equation was where Y is the $PM_{2.5}$ exposure for some person i, month j, day k and alpha is the average $PM_{2.5}$ exposure for month j:

$$Y_{i,j,k} = \alpha_j$$

We then assessed a covariate in our regression equation. As stated before, we had the distance from a road of where the mother lived. We performed a linear regression initially, but then included a quadratic term in the regression equation. As shown in the figure below, there is little to no correlation between the distance from the road and $PM_{2.5}$ exposure. Therefore, distance from the roads was not considered as a covariate of $PM_{2.5}$ exposure from this point.

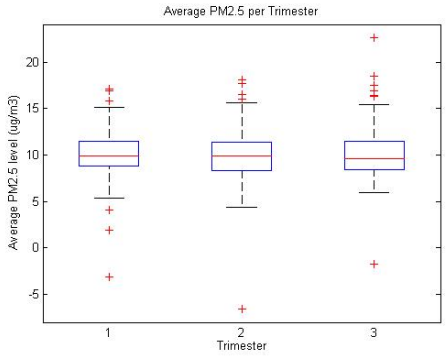
$$Y_{i,j,k} = \alpha_j + \beta_0 + \beta_1 X_1 + \beta_2 {X_2}^2 + \epsilon_i$$




With the weekly $PM_{2.5}$ averages, we converted them to the trimester averages. As seen in the boxplot, the median $PM_{2.5}$ level for the first trimester was 9.8825, the second trimester was 9.8528 third trimester was 9.5763.

After finding the average PM2.5 for each of the trimesters, we checked for the correlations between them. For each of the trimesters, we found that there was little to no correlation between each other, see the Spearman's correlation matrix and boxplot below:

$$\begin{bmatrix} 1 & 0.08217006 & 0.45435626 \\ 0.08217006 & 1.00000000 & 0.09926287 \\ 0.45435626 & 0.09926287 & 1.00000000 \end{bmatrix}$$




The matrix can be interpreted as follows: For some row i and some column j, the correlation coefficient would be between trimester i and trimester j. According to the matrix, all combinations of trimesters except 1 and 3 have little to no correlation between each other as their correlation coefficient is greater than $-0.3$ and less than 0.3. While trimesters 1 and 3 have correlation coefficients around 0.4, so they are moderately correlated.

This is a positive result for future work on the project as outcome data from the regression would not be subject to multicollinearity. Multicollinearity is a

phenomenon where two variables are *highly* correlated with each other, and as a result, makes it hard to see which variable is having an effect on the outcome Y. Since at most the trimesters were only moderately correlated, there should not be further problems pertaining to computer analysis of data such as inversion of any matrices to produce a best fit line that would stem from multicollinearity.

# 4 Conclusion/Future Work

The project for this summer consisted of the importation of data through MATLAB, then accessing the data and concatenating parts of it to have a more readily convertible form so that it could be analyzed. The specific objective was to create code to sum the $PM_{2.5}$ levels in terms of weekly and trimester averages. Before dealing with the data, there were missing conception dates, gestational ages, grid IDs, and $PM_{2.5}$ levels. The only missing data that we could remedy were the $PM_{2.5}$ levels. In order to account for the missing data, we used regression with covariates such as the months and distance from the roads. While assessing distance from the roads as a covariate, we used quadratic regression and found that the distance from the road is uncorrelated with $PM_{2.5}$ exposure.

Dealing with the trimesters, we produced a boxplot of the average trimester $PM_{2.5}$ values and found that they did not change among trimesters. Upon creating a Spearman correlation matrix, there were little to no correlations between trimester 1 to 2 and 2 to 3, but there was a moderate correlation between trimesters 1 and 3.

With the data for the exposures compiled, Dr. Vieira and her group can proceed with the advanced statistical analysis in order to determine the precise relationship of $PM_{2.5}$ exposure to bronchiolitis and OM.

# 5 Acknowledgements

I would like to thank the Mathematical BioSciences institute, the MCBU program at the University of California - Irvine, my project supervisor Dr. Veronica Vieira, and my faculty mentor Dr. Scott Bartell.

# References

[1] Brauer et. al. Traffic-related air pollution and otitis media. *Environ Health Perspect*, 2006.

[2] Vieira et.al. R01 grant application to the national institute of environmental health sciences. *NA*, 2010.