Fine Particulate Matter & Cardiovascular Death in Washington State

Team

Kathryn Cuff, Steve DeBroux, Christine Malinowski

Abstract

The relationship between acute increases of $PM_{2.5}$ (fine particulate matter) concentrations and cardiovascular deaths has been well established in previous regional studies, while the relationship between longer-term exposure to increased $PM_{2.5}$ and cardiovascular mortality has been limited to country-level data. This study provides an initial analysis of this correlation within the State of Washington, examining the EPA gathered $PM_{2.5}$ data and the Washington State Department of Health mortality data for 2007-2010 and shows a mid to high correlation with a negative Pearson product-moment correlation coefficient of -0.5104159. Regression analysis indicates an initial predictive model for the relationship though a low R^2 shows a large available margin of improvement. Utilizing 2010 census data to provide further context to this correlation, we observed deep interference of the demographic influences on cardiovascular mortality, requiring further attention to the use of census data as secondary or contextual information. These findings are further limited by 1) the lack of $PM_{2.5}$ measurements for all counties, and 2) the introduced bias of influential demographic data resulting from the limited selection of census variables tested in this small study. In order to develop a more refined assessment of the $PM_{2.5}$ and cardiovascular mortality relationship established in this work, an increase in $PM_{2.5}$ site data within Washington State is needed and a reexamination of the use of specific census-collected variables is suggested.

Introduction

The link between $PM_{2.5}$ (particulate matter smaller than 2.5 micrometers) and cardiovascular mortality has been suggested in several studies, where acute episodes of increased particulate matter (PM) are associated with increased cardiovascular deaths (Holloman, Bortnick, Morara, Strauss, & Calder, 2004; Hoek, Brunekreef, Fischer, & Van Wijnen, 2001; Ostro et al., 2008). More recently, a shift to examining the effects of long-term exposure to increased PM has suggested statistically positive associations between pollution and cardiovascular mortality (Crouse et al., 2012). This most recent work, however, is limited to the effects on a generalized, national population – in this case, Canada – and further study is necessary in order to understand the regional variation of pollution levels and its effects on the inhabitant's cardiovascular health.

Through our work, we aimed to identify regions of concentrated fine particulate matter ($PM_{2.5}$) within the State of Washington during the period spanning 2007-2010 and align it with incidences of cardiovascular mortality. This data was assessed to address the primary research question: ***Is there a potential relationship between $PM_{2.5}$ and cardiovascular mortality in the State of Washington?***

Datasets. The United States Environmental Protection Agency (EPA) captures and publicly provides data on local air quality conditions in the United States. Among the datasets available are $PM_{2.5}$ levels, taken daily (and sometimes hourly) at specific monitoring sites throughout the country[1]. The datasets are presented by year, and, for the purposes of this study, we accessed the four yearly datasets spanning 2007-2010. More specific details as to the features and characteristics of these datasets are available in Table A.2.

To examine possible correlations between $PM_{2.5}$ levels and the incidences of cardiovascular mortality in Washington State, we sought vital statistics data at the county level. After initially examining overly complex national datasets – specifically the US CDC Vital Statistics Mortality Multiple Cause-of-Death data[2] – we determined that datasets from the Washington State Department of Health, Center for Health Statistics offered public datasets at the appropriate level of detail (both in terms of geography and relevant death statistics) for our study. Specifically, we decided to utilize the Mortality Table C7 - Diseases of the Heart, Ischemic Heart Diseases, and Cerebrovascular Diseases by County of Residence[3].

Finally, to add context to any observed relationship between $PM_{2.5}$ and cardiovascular mortality, we used data from the U.S. Census Bureau, narrowed to Washington State statistics (again at the county level) that included information on specific variables such as gender, race, etc. that could then be used to provide demographic and socioeconomic context to our statistically relevant observations[4]. This census data also includes a separate metadata file containing corresponding headers or feature information for the provided yearly data.

---

[1] US EPA Air Quality System, PM2.5 - Local Conditions datasets for 2007-2010, available at http://www.epa.gov/ttn/airs/airsaqs/detaildata/downloadaqsdata.htm.
[2] US CDC Vital Statistics Mortality Multiple Cause-of-Death, available at http://www.cdc.gov/nchs/data_access/VitalStatsOnline.htm.
[3] Washington State Department of Health, Center for Health Statistics, available at http://www.doh.wa.gov/DataandStatisticalReports/VitalStatisticsData/DeathData/DeathTablesbyTopic.aspx.
[4] US Census Bureau statistics, Washington-specific data available via advanced search at http://factfinder2.census.gov/faces/nav/jsf/pages/searchresults.xhtml?refresh=t

While the specificity of some of our data sets could be narrowed beyond the county level, for the purposes of our comparisons across datasets, we chose to maintain our specificity at the county level for this analysis, allowing for future investigation at a more narrowed regionalization. Our preparation of the three data sets for this comparative analysis is discussed in further detail in the methods section below (in preliminary analysis).

Methods

Preliminary analysis. Prior to any analysis using our three datasets, we first optimized the sets to allow for easier data manipulation and munging. For the $PM_{2.5}$ data, we first created subsets of each of the four yearly datasets, using the State.Code column as our constraint where Washington State’s State.Code equals 53, according to the Federal Information Processing Standard (FIPS) utilized by the EPA. We then ran summary statistics for each yearly data set; the results of which can be seen in Table A.2.

We then created a single data set with all of the $PM_{2.5}$ data, adding a unique ID for each record in the individual datasets and a column denoting the year of the observation. The four yearly data sets were combined to include all 15,023 observations into a single data frame.

To begin to understand the variance of the $PM_{2.5}$ levels across the years, we created both scatterplots and histograms of the observed $PM_{2.5}$ values versus County.Code (again a FIPS standard value). The U.S. EPA has defined the annual health National Ambient Air Quality Standard (NAAQS) for fine particles ($PM_{2.5}$) to be 12.0 micrograms per cubic meter (µg/m3) and the 24-hour fine particle standard to 35 µg/m3 (EPA, 2013). We added an abline to the produced histograms and scatterplots to represent the 24-hour standard of 35 µg/m3 to help contextual these preliminary results (see Plot A.1. and Plot A.2.). Finally, we ran simple t-tests (Welch Two Sample t-tests, default t-tests in R) to begin to examine the variance of the data provided, an element further investigated in the next section.

We also began to examine the cardiovascular mortality data. Again, we cleaned the dataset to streamline the data relevant to this study, isolating the “Diseases of the Heart” data, adding unique IDs for each observation, adding the corresponding FIPS county code to match the county names provided, and combining the four years (2007-2010) into one data frame. We then ran summary statistics of the new dataset in isolation (Tables A.1 and A.2.) and plotted the frequency of cardiovascular deaths in Washington State for 2007-2010 (Plot A.3.) and the number of cardiovascular deaths in Washington State by county for 2007-2010 (Plot A.4.).

Variance of $PM_{2.5}$ data. After our preliminary examination of the datasets and before our examination of the potential relationship between $PM_{2.5}$ and cardiovascular death, we investigated the variance of the $PM_{2.5}$ data by running t-tests to determine if there were significant differences in the $PM_{2.5}$ levels both across the years as well as within each year of observed data. We then also ran a t-test against the EPA’s acceptable level of $PM_{2.5}$ to see if there was a statistically significant difference between the measured data and the acceptable level of air quality.

Relationship between $PM_{2.5}$ and cardiovascular death. With a general sense of the variance and summary statistics of our datasets, we then analyzed our $PM_{2.5}$ and death statistics to determine if a statistically significant relationship exists between this data. We chose to first create a scatterplot matrix to have a visual summary of the potential relationship (Plot A.5.). We then did a basic Pearson product-moment correlation test and followed this analysis with a more substantial regression analysis to determine the significance of $PM_{2.5}$ levels influence on cardio deaths. The outcomes of this work are discussed in the result section below.

Context of observed relationship. As discussed in the introduction, it is not enough to determine if a relationship exists between our two main datasets: $PM_{2.5}$ levels and WA State cardio mortality data. We need to add context to this assessment to better understand the potentially impacting factors influencing this correlation. To do this, we utilized the 2010 census data for the State of Washington that provided county-level data on 371 demographic factors. We specifically chose to limit our assessment to five variable categories, given the time and scope of this project. The variables we chose to investigate are: age (65+), gender, race, owner/renter status, and part of a household or living alone. See R code A.4. for actual corresponding feature names (column headers). We then ran additional regressions in order to improve our regression model, identifying the selected variables that positively influenced the model’s fit.

Results

Variance of $PM_{2.5}$ data. In running the t-tests for the observed $PM_{2.5}$ data, we have gained a clearer picture of the distributive differences within the data.

When comparing the observations within a given year against the average observation of that year, we see in Table 1 that the p values for each of our four years are well above the widely accepted 0.05 threshold. Therefore, we cannot reject the null hypothesis nor can we state with any statistical certainty that the observed values are produced by chance.

Table 1. Measured $PM_{2.5}$ within individual years

	t-value	p-value
2007	0.0025	0.998
2008	$2x10^{-4}$	0.9998
2009	0.0024	0.9981
2010	-0.0086	0.9932

However, looking at the measured $PM_{2.5}$ values for individual years against the full set of observations of all years, we see that there is a statistically significant difference where our p values are less than 0.05 (Table 2), allowing us to reject the null hypothesis and suggest that the data is not random.

Table 2. Measured $PM_{2.5}$ for individual years against PM2.5 for all years

	t-value	p-value
2007	8.6649	$2.2x10^{-16}$
2008	7.8376	$7.53x10^{-15}$
2009	8.811	$2.2x10^{-16}$
2010	-11.5804	$2.2x10^{-16}$
All years	38041.72	$2.2x10^{-16}$

In Table 3, we see that comparing our measured $PM_{2.5}$ data for individual years against the average of all years allows us to claim a statistically significant difference for 2007, 2008 and 2009 data, but in 2010, the p-value is well above 0.05 and we are, once again, unable to reject the null hypothesis.

Table 3. Measured $PM_{2.5}$ for individual years against average of all years

	t-value	p-value
2007	3.4371	0.008861
2008	3.1617	0.01336
2009	6.6852	0.001132
2010	1.0792	0.3219

Lastly, our comparison of the measured $PM_{2.5}$ data reveals statistically relevant results, implying a significant difference between the measured results and the known EPA-acceptable limits. That said, further study is required to add context to this statement to discern whether the data is within the perceived EPA limits or exceeds it.

Table 4. Measured $PM_{2.5}$ for all years against acceptable EPA value

	t-value	p-value
All years	-534.5873	$2.2x10^{-16}$

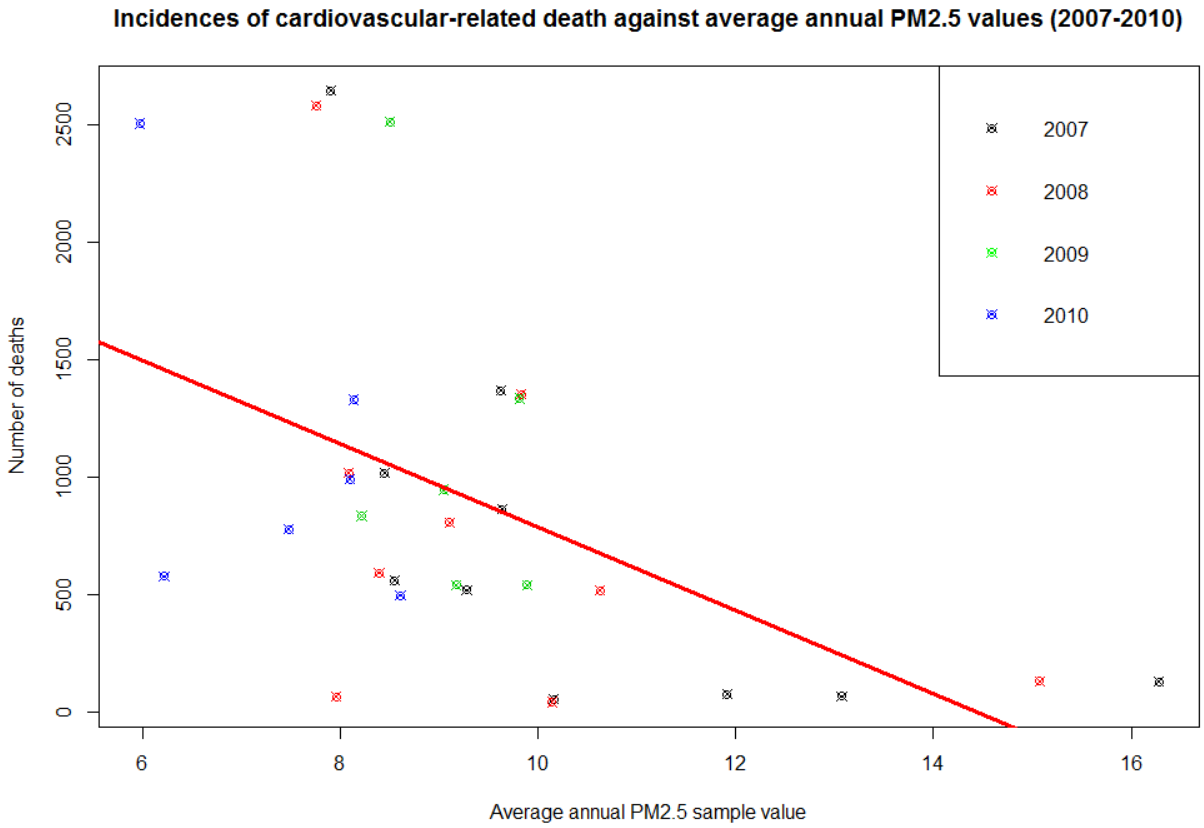
Relationship between $PM_{2.5}$ and cardiovascular death. Based on the Pearson's product-moment correlation, we see a significant negative correlation with the Pearson's product-moment correlation coefficient of -0.5104159, a value within the acceptable range of high-correlation results. See Results A.1. for the full data result of this Pearson's product-moment correlation test.

The initial linear regression model (see Plot 1 below) of the relationship between cardiovascular deaths and $PM_{2.5}$ levels produces a model of

$$\text{Cardio deaths} = 2560.75 - 177.33(PM_{2.5}) + \text{error}$$

However, this regression, with its R^2 value of 0.2605 (Results A.2.) is a mediocre model for the observed correlation.

Plot 1. Incidences of cardiovascular-related death against average annual PM2.5 values for 2007-1010



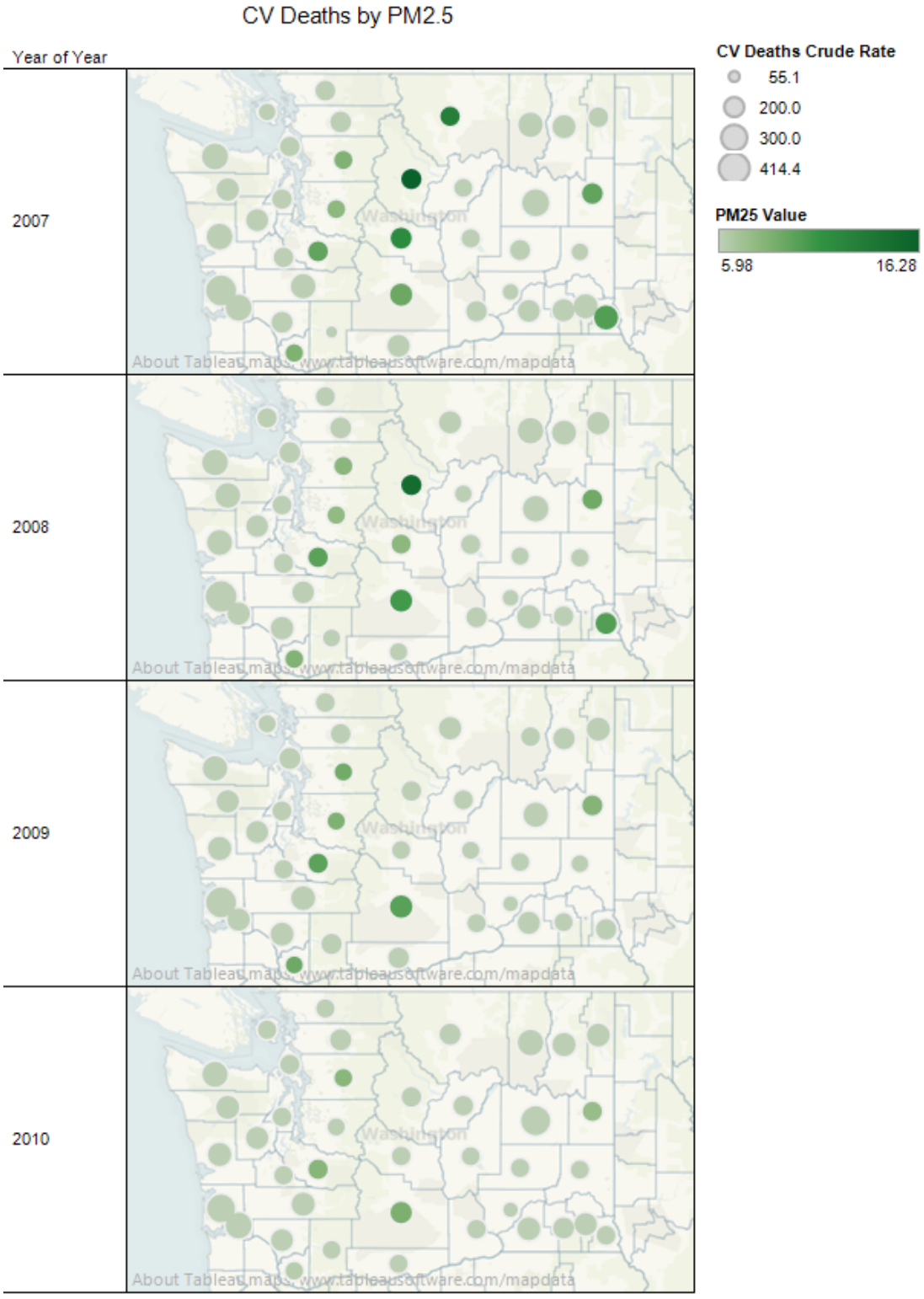
Context of observed relationship. Because of the already strong correlation between individual demographic factors such as race and age to the increase in cardiovascular deaths, the use of this data as secondary data to the $PM_{2.5}$ and cardiovascular mortality relationship proved problematic. We ran several regression analyses adding in the census data in an attempt to improve the initial linear regression model (see Table 5 below).

Proposed model of cardio deaths	Intercept std error	R^2	Corres. data
cardio deaths = $\alpha + \beta*(PM_{2.5}) + error$	534.55	0.2605	Plot 1
cardio deaths = $\alpha + \beta*(PM_{2.5}) + \gamma*(age\ 65+) + error$	344.2	0.9932	Results A.3.
cardio deaths = $\alpha + \beta*(PM_{2.5}) + \gamma*(female) + error$	393.7	0.9906	Results A.4.
cardio deaths = $\alpha + \beta*(PM_{2.5}) + \gamma*(part\ of\ household) + error$	420.4	0.9894	Results A.5.
cardio deaths = $\alpha + \beta*(PM_{2.5}) + \gamma*(living\ alone) + error$	621.1	0.977	Results A.6.
cardio deaths = $\alpha + \beta*(PM_{2.5}) + \gamma*(male) + error$	440.4	0.9882	Results A.7.
cardio deaths = $\alpha + \beta*(PM_{2.5}) + \gamma*(American\ Indian) + error$	867.56	0.9235	Results A.8.
cardio deaths = $\alpha + \beta*(PM_{2.5}) + \gamma*(Asian) + error$	1007	0.934	Results A.9.
cardio deaths = $\alpha + \beta*(PM_{2.5}) + \gamma*(African\ American) + error$	538.0	0.9786	Results A.10.
cardio deaths = $\alpha + \beta*(PM_{2.5}) + \gamma*(Other\ Race) + error$	2039.99	0.554	Results A.11.
cardio deaths = $\alpha + \beta*(PM_{2.5}) + \gamma*(Pacific\ Islander) + error$	1255.78	0.8822	Results A.12.
cardio deaths = $\alpha + \beta*(PM_{2.5}) + \gamma*(Caucasian) + error$	572.4	0.9818	Results A.13.
cardio deaths = $\alpha + \beta*(PM_{2.5}) + \gamma*(One\ Race) + error$	458.9	0.9874	Results A.14.
cardio deaths = $\alpha + \beta*(PM_{2.5}) + \gamma*(Owner\ of\ Home) + error$	497.9	0.9852	Results A.15.
cardio deaths = $\alpha + \beta*(PM_{2.5}) + \gamma*(Renter) + error$	487.7	0.9857	Results A.16.
cardio deaths = $\alpha + \beta*(PM_{2.5}) + \gamma*(Two+\ Races) + error$	473.06	0.9836	Results A.17.

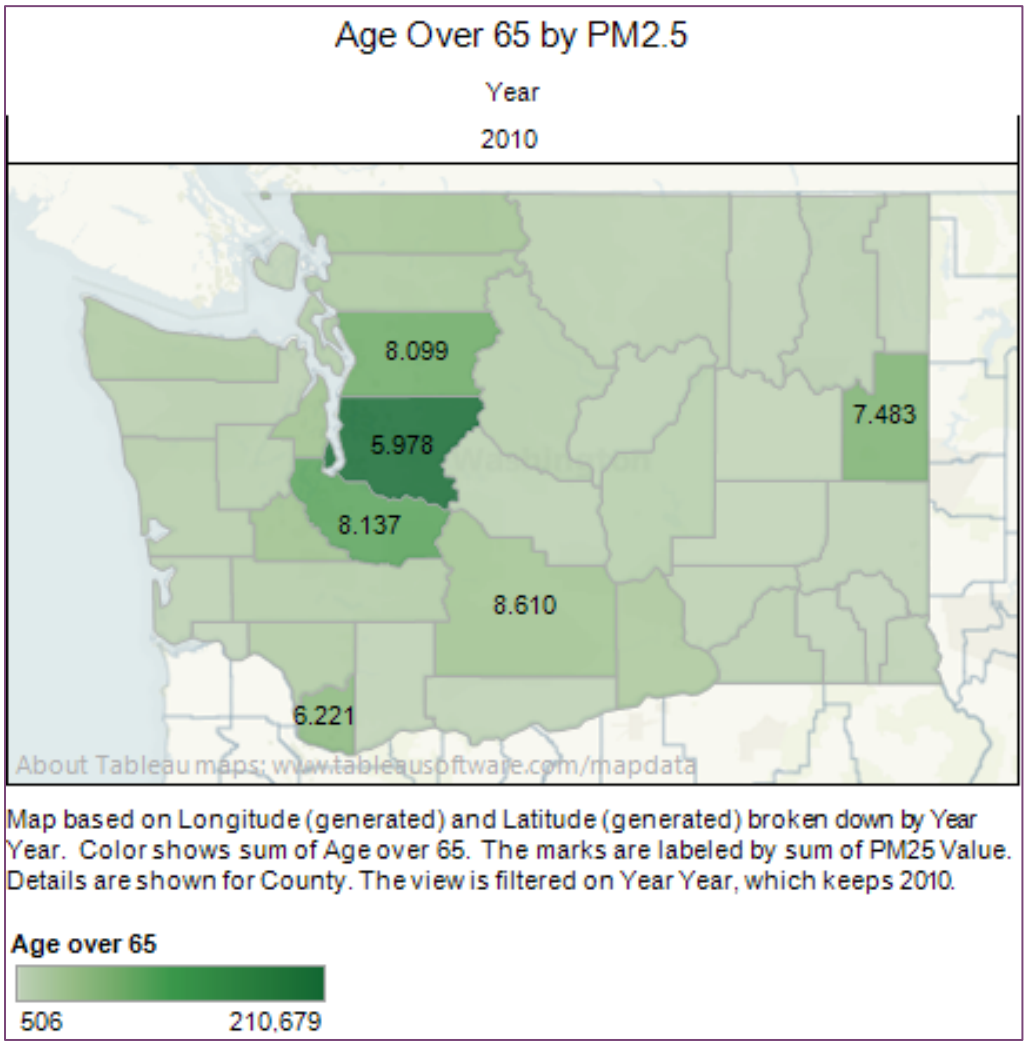
From the various regression models we can see where individual races, for example have a higher correlation to cardiovascular death. Particular to the improved model for the correlation between $PM_{2.5}$ and cardiovascular death, this approach is inadequate to provide influencing variables independent of their own relationship to cardiovascular death (outside of the environmental factors desired here).

Preliminary Visualization. Despite the preliminary nature of this study, we also wanted to explore potential visualizations relevant to both the scientific and lay audiences that could provide summary information as to the observed relationship between cardiovascular deaths and $PM_{2.5}$ concentrations in the State of Washington. Plot 2 below serves as such a preliminary visualization. Further work towards the incorporation of census data as major influences is desired as the treatment of this data is further contemplated. Visualizations such as Plot 3 could also provide further contextual visualizations where census data against $PM_{2.5}$, while not directly compared, could be an interesting point of interest.

Plot 2. Preliminary Visualization of Cardio Death and PM2.5 values



Plot 3. Preliminary Visualization of PM2.5 values and 65+ census data



Discussion

Within the scope of our primary research question, we have shown that there is a correlation between cardiovascular deaths and $PM_{2.5}$ concentrations in the State of Washington. The degree of the influence of $PM_{2.5}$ concentrations on cardiovascular death requires further refinement as our indicated predictive model could be better constructed with the right combination of secondary variables.

The inclusion of census data, initially perceived as an appropriate addition, added undesired complexity to our model where the selected census variables overshadowed the influence of $PM_{2.5}$ on cardiovascular deaths. Race and elderly population, for example, proved to be major influencing factors by themselves on cardiovascular deaths, skewing any models including $PM_{2.5}$ concentrations.

It is also important to note several caveats regarding our datasets and the above stated correlation conclusions. First, while the initial datasets were quite large, our level of geographic specificity decreased the number of relevant $PM_{2.5}$ observations and showed a lack of data for entire counties within Washington State, limiting our ability to create a full picture of the $PM_{2.5}$ concentrations of the state.

With the addition of the contextual census data, we also limited the number of variables (selecting 5 major demographic qualities out of the more than 250 provided). This introduced bias was necessary in order to remain in the scope of our project timeline, but it is important to note that the inclusion of additional census data (perhaps less independently influential) may provide a better model for the variables influencing the observed $PM_{2.5}$ and cardiovascular mortality relationship.

Conclusions

In this project, we sought to provide a framework by which we can begin to formulate valid assertions regarding the relationship between $PM_{2.5}$ (particulate matter smaller than 2.5 micrometers) and cardiovascular mortality. We were able to show in this study that there is a statistically significant relationship between these two observations. Initial regression analysis provided a preliminary regression model though further work on the predictive model with additional relevant variables is needed to improve its overall fit.

Through the course of this work, it was determined that some census data may prove problematic as secondary data in the $PM_{2.5}$/cardiovascular mortality relationship. Additional studies with more rounded $PM_{2.5}$ data covering all counties is necessary and the expansion or further investigation of appropriate linked demagraphic data is suggested before more granular conclusion can be made regarding this relationship. This study, however, provides not only a proof of concept for a preliminary assessment of the $PM_{2.5}$ and cardiovascular mortality relationship on the regional level; it also suggests a strong relationship within the State of Washington that warrants further investigation.

References

Crouse, D. L., Peters, P. A., van, D. A., Goldberg, M. S., Villeneuve, P. J., Brion, O., Khan, S., ... Burnett, R. T. (2012). Risk of nonaccidental and cardiovascular mortality in relation to long-term exposure to low concentrations of fine particulate matter: a Canadian national-level cohort study. *Environmental Health Perspectives, 120*, 5, 708-14.

EPA. (13 January 2013). National Ambient Air Quality Standards for Particulate Matter; Final Rule. *Federal Register, 87,* 10 3086. Retrieved from http://www.epa.gov/pm/actions.html

Hoek, G., Brunekreef, B., Fischer, P., & van, W. J. (2001). The Association between Air Pollution and Heart Failure, Arrhythmia, Embolism, Thrombosis, and Other Cardiovascular Causes of Death in a Time Series Study. *Epidemiology, 12,* 3, 355-357.

Holloman, C. H., Bortnick, S. M., Morara, M., Strauss, W. J., & Calder, C. A. (2004). A Bayesian hierarchical approach for relating PM(2.5) exposure to cardiovascular mortality in North Carolina. *Environmental Health Perspectives, 112*, 13, 1282-8.

Ostro, B. D., Feng, W. Y., Broadwin, R., Malig, B. J., Green, R. S., & Lipsett, M. J. (2008). The impact of components of fine particulate matter on cardiovascular mortality in susceptible subpopulations. *Occupational and Environmental Medicine, 65*, 11, 750-6.

Appendices

Table A.1. Unique observations of each dataset
Table A.2. Observational Details (information/features/characteristics) of Source Datasets

Plot A.1. Histograms, PM2.5 Sample.Value v. County.Code (each year and combined)
Plot A.2. Scatterplots, PM2.5 Sample.Value v. County.Code (each year and combined)
Plot A.3. Histogram, Frequency of Cardiovascular Deaths in Washington State
Plot A.4. Number of Cardiovascular Deaths in Washington State by county (2007-2010)
Plot A.5. Scatterplot Matrix, Incidences of cardiovascular death vs. $PM_{2.5}$ by county and year

Results A.1. Pearson’s product-moment correlation test
Results A.2. Linear Regression, cardio deaths vs. PM2.5 concentration
Results A.3. Linear Regression, with 65+ age
Results A.4. Linear Regression, with female
Results A.5. Linear Regression, with part of household
Results A.6. Linear Regression, with living alone
Results A.7. Linear Regression, with male
Results A.8. Linear Regression, with American Indian
Results A.9. Linear Regression, with Asian
Results A.10. Linear Regression, with African American
Results A.11. Linear Regression, with ‘Other’ race
Results A.12. Linear Regression, with Pacific Islander
Results A.13. Linear Regression, with Caucasian
Results A.14. Linear Regression, with one race
Results A.15. Linear Regression, with homeowner status
Results A.16. Linear Regression, with renter status
Results A.17. Linear Regression, with 2+ races

R code A.1. Preliminary analysis
R code A.2. Variance of $PM_{2.5}$ across years
R code A.3. Relationship between $PM_{2.5}$ and cardiovascular death
R code A.4. Context of observed relationship (census data)