BREN SCHOOL OF ENVIRONMENTAL SCIENCE & MANAGEMENT
University of California - Santa Barbara

Diabetes and Air Quality in California

Degree Objective: Master of Environmental Science and Management
Prepared by: Vanessa Guenther, Yirui Zhang, Angie Bouche, Hope Cupples

Client: Sansum Diabetes Research Institute

Faculty Advisor:
Ashley Larsen

PhD Advisor:
Patrick Hunnicutt

April 2020

Bren School of Environmental Science & Management
UNIVERSITY OF CALIFORNIA, SANTA BARBARA

JAMES S. BOWER FOUNDATION
Re-envisioning.

SANSUM DIABETES RESEARCH INSTITUTE

Abstract

A growing body of research has linked air pollution with a myriad of chronic and acute health conditions. However, the relationship between air pollution and one widespread and increasingly common condition, type 2 diabetes, has yet to be rigorously tested. This project aims to fill this crucial gap by assessing relationships between particulate matter 2.5 ($PM_{2.5}$) and diabetes prevalence in California, USA, using a cross sectional and panel data approach. The two model types assess the years 2014 through 2017 to understand the possible relationship between diabetes and $PM_{2.5}$ in the state.

Cross sectional linear models for the years 2014 through 2016 show a positive association between $PM_{2.5}$ and diabetes prevalence (0.06 increase in diabetes prevalence with 1 ug/m^3 increase of $PM_{2.5}$). Results from our fixed effects analysis are qualitatively similar (0.04 increase in diabetes prevalence with 1 ug/m^3 increase of $PM_{2.5}$). The 2017 cross-sectional model and the fixed effects model with all years (2014-2017) is near null and not significant. We explore possible explanations for the 2017 results relating to changes in socioeconomic conditions and the possibility of non-linear relationships between $PM_{2.5}$ and diabetes. The relationship between $PM_{2.5}$ and diabetes is complex and could vary depending on functional form and timescale of the interaction.

Executive Summary

Public health initiatives have long focused on health behaviors and lifestyle factors that contribute to the incidence of non-communicable diseases, but now environmental conditions are also being examined. In particular, air pollution has been associated with a range of negative health impacts including heart disease, stroke, chronic obstructive pulmonary disease, lung disease, and lower respiratory infections in children. Air pollution also contributed to 2.7 million deaths worldwide in 2012 (Kelly and Fussell, 2015). Entities including the United Nations and World Health Organization have now placed air pollution as one of their risk factors for non-communicable diseases (Linou et al., 2018).

One non-communicable and preventable disease that has been increasing in prevalence since the end of the 20th century is type 2 diabetes. Within the US, one in ten people are diagnosed with diabetes, which cost the United States $327 billion in 2017 alone (ADA, 2018). Type 2 diabetes, which represents 95% of all cases, is known to be influenced by a range of factors including environmental conditions, socioeconomic status, and health behaviors (WHO, 2016). Research on environmental conditions has focused on air pollution, specifically particulate matter 2.5 ($PM_{2.5}$). One longitudinal cohort study on $PM_{2.5}$'s effects on diabetes prevalence projected that ambient $PM_{2.5}$ contributed to 3.2 million cases of type 2 diabetes globally (Bowe et al., 2018).

Air pollution has proven to be a more widespread issue in California compared to other states. The 2019 American Lung Association "State of the Air" report compiled data from the U.S. Environmental Protection Agency to rank the cities with the highest levels of air pollution. The report found that California continues to dominate this list, containing six of the 10 most polluted cities in the country (American Lung Association, 2017). Our project will target a gap in literature by assessing the relationship between $PM_{2.5}$, a major air pollutant, and type 2 diabetes, an increasingly prevalent disease, in a state that has high rates of both. Understanding if and how environmental factors are associated with diabetes can allow healthcare providers in California to more effectively implement prevention techniques that go beyond diet, exercise, and prescriptions.

In this project we combine publicly available datasets that include $PM_{2.5}$ concentrations from the California Air Resource Board, diabetes prevalence from the Centers of Disease Control, and sociodemographic variables from the Census Bureau's American Community Survey for the years 2014 through 2017. The $PM_{2.5}$ data was recorded as daily or hourly measurements from air quality monitors throughout the state, while diabetes prevalence and socioeconomic variables were recorded by census tract. Because of this, we first needed to wrangle the $PM_{2.5}$ data by averaging observations each year and interpolating these concentrations across the state. This allowed us to assign a $PM_{2.5}$ concentration at each census tract. After wrangling all datasets, we retained observations for 5,084 census tracts across California.

We used both cross sectional and fixed effects models to analyze data from 2014 through 2017. The cross sectional models assess the relationship between diabetes and $PM_{2.5}$,

along with various sociodemographic variables, to compare these relationships in each year of study. As a more statistically rigorous test, we also run a fixed effects model that uses panel data of these same variables to control for time-invariant factors that are not accounted for in the cross sectional models.

Among the cross sectional models, we see a positive and significant association between diabetes prevalence and $PM_{2.5}$ concentration when sociodemographic variables are included for the years 2014, 2015, and 2016 (~0.06, $p<0.001$). In the year 2017 there is no significant association. Likewise, when we run the fixed effects model across the years 2014-2016 we find a significantly positive association between diabetes prevalence and $PM_{2.5}$. However, when the fixed effects model encompasses data from 2014-2017 there is no association.

We hypothesize that these conflicting results could be attributed to a range of factors. In 2017, major wildfires were located close to census tracts incorporated in our analysis. However, since $PM_{2.5}$ values are an annual average concentration, the effects of fires should be minimized in our $PM_{2.5}$ value. Socioeconomic conditions like access to healthcare, unemployment rate and poverty rate, across California improved steadily across the years of study. Finally, in the literature the overall time scale at which air pollution and health conditions are associated is unknown. It is possible that the chronic effects of high $PM_{2.5}$ concentrations on health would not be observed over the four years of study.

In this relatively new body of research where public health meets environmental science, there is room for new research to build on our analysis. Our results show a possible association between $PM_{2.5}$ and diabetes in California that can be further explored to assess the consistency of trends and to better understand the timescale of this interaction.

Table of Contents

I. Introduction

Diabetes is an increasingly widespread disease with negative health and economic impacts. Within the US, one in ten people are diagnosed with the disease, which cost the United States $327 billion in 2017 alone (ADA, 2018). Diabetes occurs when the pancreas does not produce enough insulin (type 1), or when the body cannot effectively use the insulin it produces (type 2). Diabetes of either type can lead to blindness, amputations, strokes, heart attacks, and other serious health events (WHO, 2016). Type 2 diabetes, which represents 95% of all cases, is known to be influenced by a range of factors including socioeconomic status and health behaviors (ADA, 2015). Specific risk factors include smoking, obesity, physical inactivity, high blood pressure, high cholesterol and high blood glucose (Division of Diabetes Translation, 2017).

A growing body of research is examining the links between environmental factors, including pollution exposure, and diabetes rates. According to the literature, pollutants of concern include air pollution (particulate matter, nitrogen dioxide, etc.), drinking water pollution (accumulative arsenic exposure, inorganic arsenic drinking water), proximity to hazardous waste (persistent organic pollutants) and pesticide exposure (insecticides, herbicides, fungicides, rodenticides, and molluscicides) (Huang et al., 2011; Navas-Acien et al., 2008; Kouznetsova, M. et al., 2007; Juntarawijit et al., 2018). Of these, air pollution from very fine particulate matter shows the most strongly supported association with diabetes (Navas-Acien et al., 2008; Steinmaus et al., 2009; Saldana et al., 2007; Kouznetsova, I. et al., 2007; Sergeev and Carpenter, 2005).

Fine particulate matter is made up of a mixture of organic chemicals, dust, soot and metals (AirNow, 2017). The particulate matter particles of the greatest health and regulatory concern are those with a diameter of 2.5 micrometers ($PM_{2.5}$), which is less than the thickness of a human hair (Rodriguez and Zeise, 2017). The small size allows these particles to be inhaled, deposited in the lungs, and passed into the bloodstream (Canadian Centre for Occupational Health, 2019). They can also transport other toxic chemicals into the bloodstream that are harmful to human health (CARB, 2015). $PM_{2.5}$ is released into the atmosphere from a range of anthropogenic sources. These sources include cars and trucks, factories, and burning wood (EPA, 2019). Natural sources of $PM_{2.5}$ include dust from the wind erosion of natural surfaces, sea salt, wildland fires, and primary biological aerosol particles (EPA, 2019).

$PM_{2.5}$ has previously been linked to a range of negative health impacts ranging from Alzheimer's and dementia to heart attacks (Jung et al., 2015, Rajagopalan et al., 2018). For this reason, $PM_{2.5}$ is a U.S. EPA criteria air pollutant (EPA, 2017) and is regulated on both the state and federal level. However, based on the available scientific evidence, air quality analyses and risk assessments, the current primary annual standard for $PM_{2.5}$ of 12 $\mu g/m^3$ may not be adequate to avoid severe health impacts (EPA, 2019).

The specific mechanism in which air pollution might interact with type 2 diabetes is still not fully understood, especially at the molecular and cellular level. Broadly speaking, $PM_{2.5}$ particulates act as foreign bodies in the bloodstream and trigger an inflammatory response.

Some studies suggest $PM_{2.5}$ may stimulate oxidative and inflammatory responses in the lungs that affect the function of other organs (Xing et al, 2016), while others suggest particulates may be translocated to central nervous system receptors (Dimakakou et al., 2018).

The timescale of health outcomes between $PM_{2.5}$ exposure and diabetes are still being explored. A report released by the EPA demonstrated that adverse health effects were observed with lags ranging from one or two days to several months for different health outcomes (EPA, 2019). Seasonality further plays a role, as researchers from the University of Windsor also reported that cool and dry weather increased the adverse effects of $PM_{2.5}$ on human health, whereas warm and humid weather decreased the effect (Miller, et al, 2018). We did not find any literature on the specific time lags of $PM_{2.5}$ and diabetes prevalence.

Similarly, the functional form between $PM_{2.5}$ and diabetes is not well understood. Even short-term (acute) exposure to PM is known to cause exacerbations of diabetes leading to hospitalizations and death (Andersen et al., 2012). Some studies looking at $PM_{2.5}$ and disease outcomes strongly suggest that health effects have no threshold within the studied range of ambient concentrations and can occur at levels close to $PM_{2.5}$ background concentrations. A meta-analysis of seven studies on $PM_{2.5}$ explored this linear relationship by showing that with every 10 $\mu g/m^3$ increase in $PM_{2.5}$ concentration, diabetes risk increased by 25% with chronic long-term exposure (He et al., 2017)

In California, 55% percent of all adults have diabetes, prediabetes, or undiagnosed diabetes, costing the state more than $27 billion annually, with $19 billion of that spent on direct medical care for diabetes (Babey et al., 2016). California is also home to six of the 10 most polluted cities in the country which are home to around 20% of the state's population (American Lung Association, 2019). The state further struggles with an unequal distribution of air pollution based on one's socioeconomic or demographic status (Table A2, Boyd-Barret, 2019). Diabetes prevalence throughout California is also not distributed evenly. Type 2 diabetes prevalence among Mexican-origin Latino adults (18%) is nearly double than that among non-Latino whites (9.6%) (CDC 2016). Latinos of any race have a higher diabetes prevalence rate (11.8%) than non-Hispanic whites (8.1%) across California (Health Rankings, 2019). High pollution, along with high diabetes prevalence and equity concerns make California a compelling state to explore the possible association of these two variables.

Currently, there are no state-wide California studies exploring this association even though the state struggles with air pollution and diabetes. Our literature review revealed that investigations into this possible association have been sparse. Existing research has been conducted in developed nations in North America and Europe but has not yet been explored in California. California has struggled with high levels of air pollution since the second half of the 20th century, and as a result has a comprehensive network of air monitoring stations, particularly in urban areas compared to other states, making it an ideal location to study. Five studies investigating $PM_{2.5}$ and diabetes in various locations do show a positive association (Table A1), however these studies used different methods of calculating $PM_{2.5}$ exposure than our methodology. Notably, in previous studies $PM_{2.5}$ data is interpolated at the zip code level and typically averaged over a multi-year period. Given our access to census tract-level diabetes

data in the state and the robust air monitoring network in California, we are able to explore this association at a more granular level than some of the studies in the literature (Table A1).

Within California, rural areas have lower rates of diabetes prevalence than the national average, while in suburban and urban areas diabetes prevalence is higher than the national average (2% and 0.6% higher respectively) (Health Rankings, 2018). Urban dwellers in California appear to have the highest risk of diabetes, making them an important population to study. As such, understanding the relationship between urban air pollution and diabetes prevalence is of fundamental concern both socially and economically. Our analysis therefore focuses on the most populated areas of the state and seeks to fill this gap by leveraging detailed, publicly available air quality and diabetes data.

Specifically, we address the following questions:

1. What are the yearly average $PM_{2.5}$ concentrations at the census-tract level in California?
2. What is the relationship between these $PM_{2.5}$ concentrations and diabetes prevalence in California?
3. Is there heterogeneity in results based on thresholds to exposure levels and ethnicity?

II. Methods

We used both a cross sectional and panel data approach to assess relationships between $PM_{2.5}$ and diabetes prevalence in California, USA. Daily and hourly $PM_{2.5}$ measurements were leveraged from the California Air Resource Board (CARB) and CalEnviroScreen (CES) in combination with diabetes prevalence data from the Centers for Disease Control for approximately 5,000 census tracts across California. Sociodemographic variables from the Census Bureau's American Community Survey (ACS) were also incorporated into the models. This analysis provides an exploration of the possible relationship between diabetes prevalence and $PM_{2.5}$ at the census-tract level, accounting for demographic and socioeconomic factors statewide.

Data

Diabetes data for the years 2014-2017 were collected from the Centers for Disease Control and Prevention (CDC)'s 500 Cities database. These datasets were published in the 2016-2019 releases of 500 Cities, respectively. The CDC's 500 Cities database is an initiative to provide health-related data for the most populated 500 cities in the United States. Diabetes rates are given as model-based estimates of crude prevalence of diagnosed diabetes among adults, which are calculated based on the Behavior Risk Factor Surveillance System (BRFSS). This database includes diabetes prevalence within a census tract in populated cities throughout California. Diabetes prevalence represents a percent of the adult population in a census tract diagnosed with diabetes. On average, the population of a census tract in the 500 cities database in California was 4,269 people.

Socio-demographic data is obtained from the American Community Survey (ACS). ACS is an ongoing survey conducted by the U.S.Census Bureau that produces a dataset each year

with demographic, economic, and social data based on 35 million households' responses. In this study, we selected the unemployment rate, educational attainment (percent of people who do not have a high school degree), and poverty rate (percentage of people with income less than the federal poverty level) for further analysis.

$PM_{2.5}$ data is collected from all monitoring stations within the CARB monitoring network. There are around 180 air monitoring stations placed throughout the state (Figure 1). This dataset contained daily $PM_{2.5}$ observations from two types of monitors, Beta Attenuation Method Monitors (BAM) and Federal Reference Method Monitors (FRM). BAM monitors measure air quality continuously. Hourly measurements from these monitors are averaged over a 24-hour period to provide daily observations. FRM monitors contain a filter that collects $PM_{2.5}$, which is then manually taken out of the monitor and weighed. Daily $PM_{2.5}$ observations from both types of monitors are included in datasets for years 2000-2018.

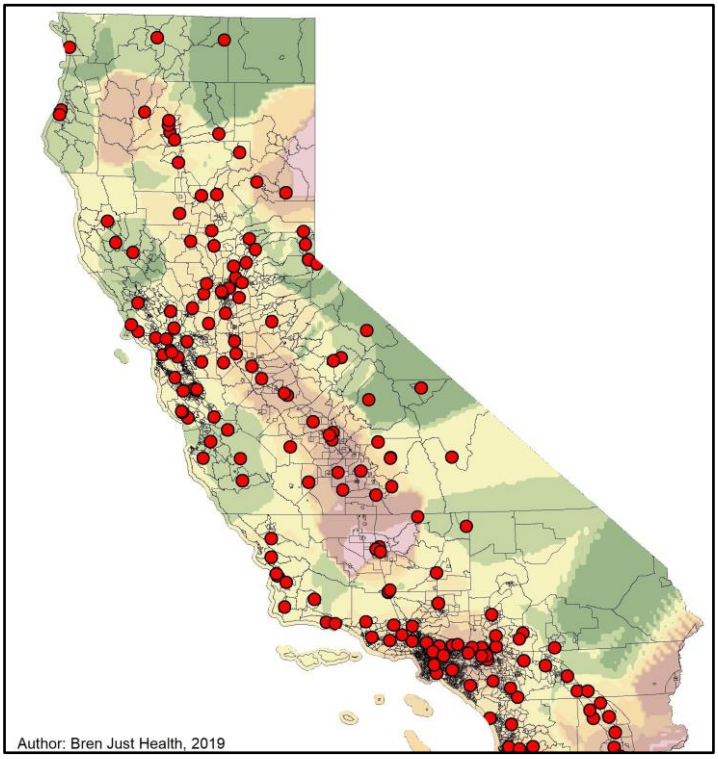


Figure 1. $PM_{2.5}$ Air Monitoring Stations in California. Of the ~190 monitors across California, 59 federal reference monitors measure daily observations, while 131 beta attenuation and speciation monitors measure $PM_{2.5}$ continuously.

These daily observations were averaged over each year of interest. Because $PM_{2.5}$ has pronounced seasonality, FRM monitoring stations that record daily observations operate on a variable sampling schedule. During the winter, these monitors record observations once every three days, and during the remainder of the year once every six days. To minimize seasonal bias, we averaged daily observations by quarter and then by year. This followed the protocol used by CARB to generate $PM_{2.5}$ maps for CES, a California state tool released by the Office of Health Hazard Assessment (Tran et al., 2008).

The data we have on diabetes are limited to the years 2014-2017. For this reason, we began our analysis of $PM_{2.5}$ values for those same years. We created a continuous $PM_{2.5}$ surface between air quality monitoring stations over California using fixed-radius ordinary kriging using ArcGIS 10.7.1 (Figure A1). This method was used and verified for $PM_{2.5}$ interpolation in a range of other studies (Rivera-Gonzalez et. al 2015) (Wu and Hung, 2016). The search radius parameter was set at 50km (Tran et al., 2008). If a monitoring station was not found within 50km, the value from the next nearest monitoring station was used.

The assigned $PM_{2.5}$ value was the average value for the entire census tract (Figure 2). A simplified version of the model in ArcGIS is included in Appendix 4.

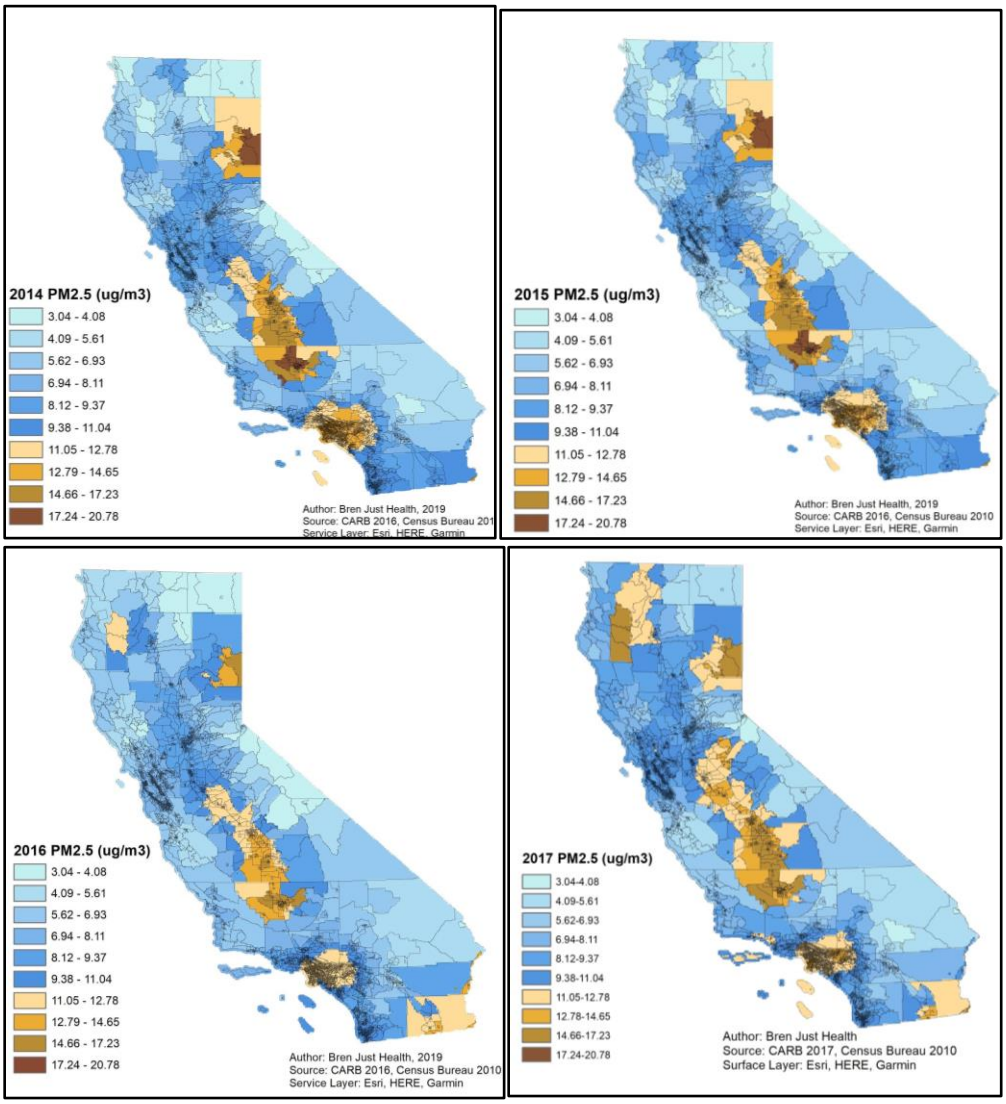


Figure 2. 2014 -2017 $PM_{2.5}$ in California. The blue values identify census tracts with $PM_{2.5}$ concentrations below the National Air Quality Standards (NAAQS) established by the EPA. The brown values identify the $PM_{2.5}$ concentration above the NAAQS established by the EPA. Hotspots are located in Plumas County, the Central Valley and Los Angeles.

The results from the above methodology were compared to an existing dataset from CES to ensure that our model accurately interpolated $PM_{2.5}$ values across census tracts. The CES data contained $PM_{2.5}$ values at the census tract level averaged over three years and thus were not compatible with our cross sectional and fixed effects approaches (Figure A3). However, they provided a way to validate our interpolation methodology. Following the above methodology, we created a comparable dataset for 2012-2014 from the raw CARB data (Figure A4). Although data from more monitoring stations were included in our interpolation from raw CARB data, 90% of our data points were still less than 13% different than CES data, with a maximum absolute difference of 5.4 $\mu g/m^3$ (Figures A5 and A6).

Areas on the outskirts of cities are typically where census tracts have a higher percent difference between our data and CES. In densely populated areas, there are many air quality monitors placed, so the addition of several extra monitors does not affect the results of kriging. On the outskirts of cities where there are fewer monitors overall, additional monitors will have a greater impact on kriging results and therefore lead to a discrepancy between our data and CES data. In areas like Bakersfield, CA, which lies inland from Los Angeles, all monitoring stations in our data and CES data were the same and the percent difference between values was less than 10% (Figures A7 and A8).

Cross Sectional Models

We created a series of cross-sectional models to investigate relationships between annual $PM_{2.5}$ concentration and diabetes prevalence. Cross sectional models contain variables that are all associated with the same single period in time. We chose specific sociodemographic variables (educational attainment, poverty rate, unemployment rate, and race/ethnicity) from a larger suite of sociodemographic variables to reduce collinearity (Figure A9).

Each cross section was analyzed using a multivariate linear model with diabetes as a linear function of $PM_{2.5}$ concentration and various combinations of sociodemographic variables. To do so, we used contemporaneous measures of diabetes, $PM_{2.5}$ and demographic characteristics. In addition, we applied the same model to examine the CES data as a model robustness test. In this test we paired $PM_{2.5}$ concentrations from CES 2.0 (2009-2011) with diabetes and sociodemographic variables from 2014. We also paired $PM_{2.5}$ concentrations from CES 3.0 (2012-2014) with diabetes and socioeconomic variables from 2016. An example of the multivariate linear model equation is:

$$Diabetes\ Prevalence_i = \beta_1\ PM2.5_i + ... + e_i$$

Fixed Effects Models

Besides using cross sections to assess relationships between $PM_{2.5}$ and diabetes prevalence, we also take a panel data approach to identify the relationships in a time series, using annual data from 2014-2017. A fixed effects model controls for variables that are unique to a census tract and are time invariant. For the fixed effects model, there is one observation for $PM_{2.5}$, diabetes prevalence, and each socioeconomic variable for each census tract each year. Based on the results from the cross-sectional analysis, we ran the fixed effects model on all

years (2014-2017) and a subset of years (2014-2016). Prior to running our fixed effects analysis, we assessed the variation in the $PM_{2.5}$ and diabetes datasets. This was done by calculating the standard deviation of linear model residuals. We confirmed that there is variation in both the $PM_{2.5}$ dataset from CES and diabetes prevalence dataset from the CDC that is not purely a function of census tract and time and allowed us to move forward with our fixed effects model.

To explore the impact of socioeconomic indicators, we ran the fixed effects model twice, once on a model that did not include sociodemographic variables and once on a model that did include these. Both linear models incorporated cluster robust standard errors to allow for heteroscedasticity and spatial autocorrelation of the errors (Vogelsang, 2012). The basic model analyzed the changes in $PM_{2.5}$ and diabetes alone that occurred in a census tract over three or four years of study, where subscripts *I* is a given census tract and *t* is time. We tested models including the same sociodemographic variables, as in the cross sectional models.

$$Diabetes\ Prevalence_{it} = \beta_1 PM_{2.5\ it} + Census\ Tract_i + Year_t + e_{it}$$

Threshold Analysis

$PM_{2.5}$ concentrations used in the models thus far were recorded as annual average concentrations in $\mu g/m^3$. These values were then incorporated in linear models. The literature is unclear on the functional form of the relationship between $PM_{2.5}$ and health conditions, therefore, we also wanted to explore the possibility of non-linear relationships into our analysis. In this introductory look at non-linear relationships, we calculate the proportion of observations that exceed the EPA's National Ambient Air Quality Standards (NAAQs) of 12 $\mu g/m^3$at a given location.

First, we selected monitoring locations with hourly monitors that recorded at least 300 days of observations for that year. This translates to 106 monitor locations across the state. We calculated the proportion of observations at each station that are above 12 $\mu g/m^3$ and incorporated that value in our ordinary kriging model in GIS; the same way we used kriging with the average $PM_{2.5}$ concentrations. Kriging assigns one $PM_{2.5}$ proportion per census tract. We then repeated our methodology of cross sectional and fixed effects models using this proportion as the $PM_{2.5}$ value. We included the same sociodemographic variables that were used in our original fixed effects analysis.

Latino Subgroup Analysis

To assess if there is a different relationship between $PM_{2.5}$ and diabetes in areas that are more or less populated by Latinos, we conducted a subgroup analysis, splitting our sample into high and low Latino-populated census tracts before running the fixed effects model. The median percentage of the population identifying as Latino across all census tracts with diabetes data was 33%. We created a binary operator where "1" was assigned to any census tract that was equal to or over 33% Latino of any race in 2014, and "0" was assigned to any census tract that was less than 33% Latino of any race in 2014. The binary assigned to a census tract in 2014 remained for all years of study, 2014-2017, regardless of whether the census tract had changes

in demographics. We repeated this process using bins of under 50% Latino and over 50% Latino corresponding with binary operators of “0” and “1” respectively. We removed all continuous racial demographic variables and retained only the Latino binary variable in the model to reduce collinearity between variables. We then ran the fixed effects models as described above, using both average concentrations and threshold proportions for $PM_{2.5}$ values, on the high and low Latino binary subgroups.

Leave-One-Out Analysis

To better understand our fixed-effects results, we explored how each county affects the aggregate coefficient. To do this we grouped the data by county and ran the fixed effects models leaving out one county each time. The model equation was the same as above sections but had different sample sizes depending on which county was left out of the model. We ran this analysis twice, first using the average concentration as the $PM_{2.5}$ value and second using the threshold proportions as the $PM_{2.5}$ value. By analyzing and comparing the leave-one-out coefficients with the original all-county coefficient, we can explore which counties that have the largest impacts on the original model coefficient.

III. Results

Cross Sectional and Fixed Effect Model Results

In the series of cross sectional models from 2014 to 2016, we see similar positive relationships between $PM_{2.5}$ and diabetes at each year controlling for sociodemographic variables. The coefficient β_1 is approximately 0.048 (+/- 0.011, p <0.001) in 2014, 0.071 (+/- 0.013, p <0.001) in 2015, and 0.059 (+/- 0.015, p <0.001) in 2016 when socioeconomic variables are included in the cross section (Figure 3 and Figure 4). A coefficient of 0.059 represents a 0.059 percentage point increase in diabetes crude prevalence when $PM_{2.5}$ concentration increases by one unit (ug/m^3). For 2017 we find a different pattern entirely. Here, the cross sectional coefficient is much smaller than the previous years (β_1=0.015, +/- 0.016, p <0.001) and it is not significantly positive (Figure 4). To verify the integrity of our $PM_{2.5}$ dataset, we compared model coefficients between cross sections using CES $PM_{2.5}$ values and saw similar coefficients regardless of data source. Additional results are provided in the Supplemental Information (SI).

Fixed effects models are a more statistically rigorous method to explore the relationship between diabetes prevalence and $PM_{2.5}$ values using panel data. With panel data we can control for time-invariant factors that are unobserved or unmeasured, resolving omitted variable bias that could be incorporated in the cross sectional models. The coefficient association between $PM_{2.5}$ and diabetes prevalence resulting from the fixed effects model across 2014-2016 is approximately 0.034 (+/-0.006, p <0.001) when sociodemographic variables are included and 0.037 (+/-0.007, p <0.001) when they are excluded (Figure 3). This represents a 0.04 percentage point increase in diabetes crude prevalence when $PM_{2.5}$ concentration increases by one unit (ug/m^3). When the 2017 data is incorporated in the fixed effects model, meaning the panel of data is now from 2014-2017, the coefficient becomes <0.001 (+/- 0.006, P > 0.05)

(Figure 4).

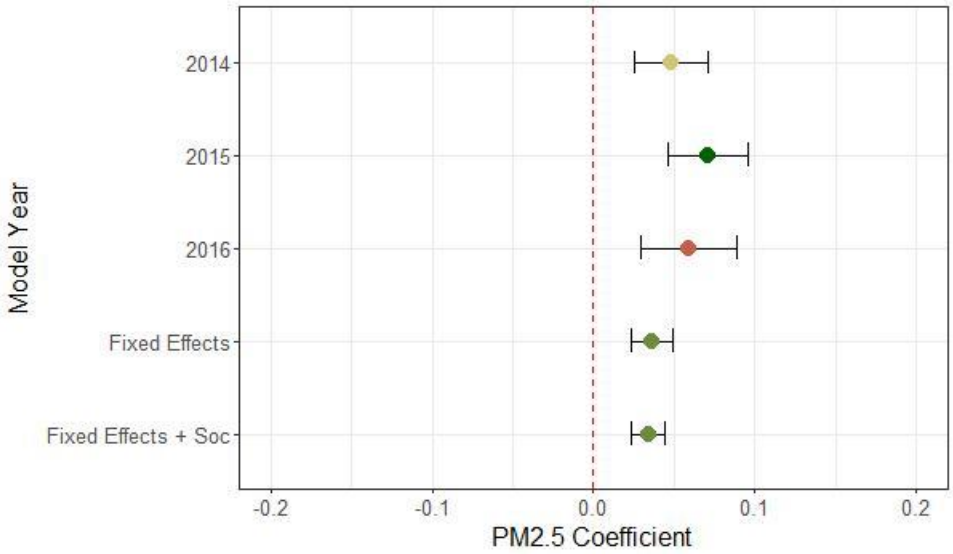


Figure 3. Cross Sectional and Fixed Effects Results 2014 - 2016. The coefficient association between $PM_{2.5}$ and diabetes prevalence shows a small but significant positive association between $PM_{2.5}$ concentration and diabetes prevalence in both the cross sectional and fixed effect models.

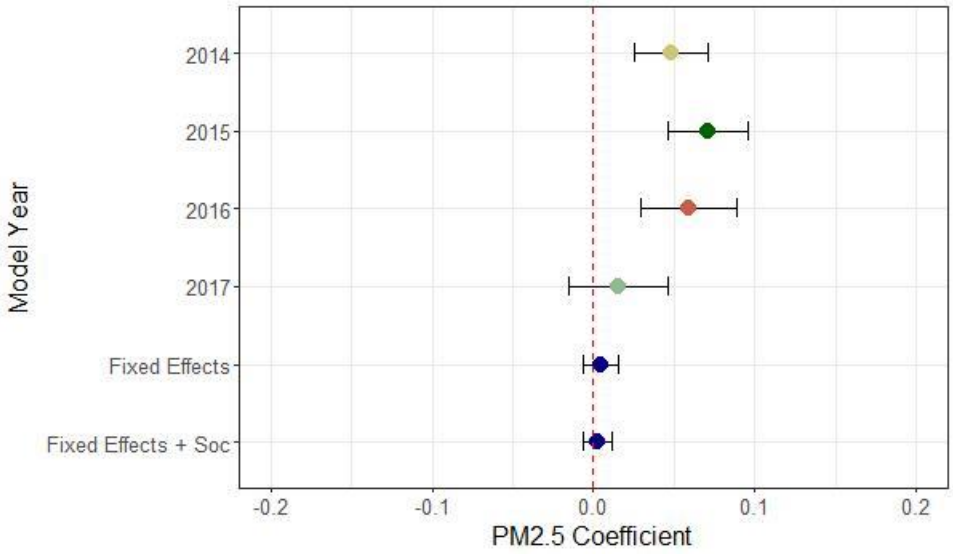


Figure 4. Cross Section and Fixed Effects Results 2014 - 2017. The coefficient association between $PM_{2.5}$ and diabetes prevalence is much smaller in 2017 than previous years. Incorporating a panel of data from 2014-2017 in the fixed effects model also results in no association between $PM_{2.5}$ and diabetes prevalence.

Table 1. Fixed Effects Model Output. In the fixed effects models using a panel of data from 2014-2016, there is a small but positive association between $PM_{2.5}$ and diabetes prevalence that is significant. In the fixed effects model incorporating 2017 data into the panel, there is no significant association.

Panel	Model Type	Coefficient	Standard Error	P Value	CI Low	CI High

2014-2016	Socio	0.034	0.006	<0.001	0.023	0.044
2014-2016	$PM_{2.5}$ Only	0.037	0.007	<0.001	0.024	0.050
2014-2017	Socio	0.003	0.005	0.593	-0.006	0.011
2014-2017	$PM_{2.5}$ Only	0.004	0.006	0.428	-0.006	0.015

Threshold Analysis

This nonlinear method is a basic exploration that needs to be refined to draw more dependable results. However, we see a positive and significant association between diabetes prevalence and proportion of days in exceedance for each of the years 2014-2017 (Figure 5). Our fixed effects model incorporating a panel of data from 2014-2017 with socioeconomic variables shows a positive association (β_1= 0.552 +/- 0.0495, p < 0.001). Our fixed effects model incorporating a panel of data from 2014-2016 with socioeconomic variables also shows a positive association (β_1= 0.622 +/- 0.0508, p < 0.001). Additional model results can be found in the SI. This means that a ten percentage point increase in the proportion of days in exceedance is associated with a 0.062 percentage point increase in diabetes prevalence.

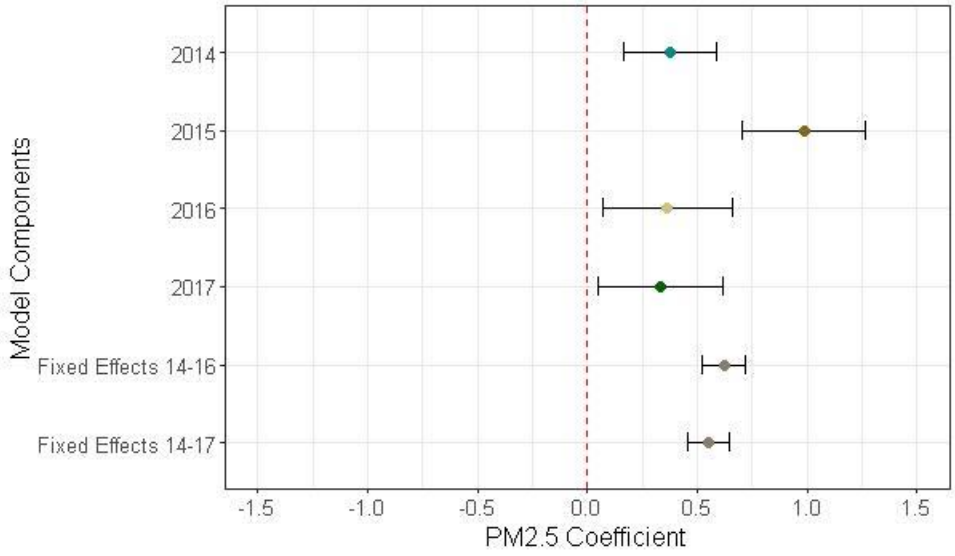


Figure 5. Cross Sectional and Fixed Effects Results Using Thresholds. The coefficient association between $PM_{2.5}$ and diabetes prevalence shows a small but significant positive association in both the cross sectional and fixed effect models. All models include socioeconomic variables.

Latino Subgroup Analysis

We explored interactions between $PM_{2.5}$ and Latino populations to assess if the relationship between air pollution and diabetes prevalence is more or less strong among areas that are heavily populated by Latinos. We first classified census tracts as either high percentage Latino or low percentage Latino. General summary statistics show that diabetes prevalence, $PM_{2.5}$ concentration and sociodemographic variables are larger in high Latino census tracts

(Table 2 & Table 3).

Table 2. Yearly Average Values for Census Tracts Above the Median Percent Latino. Yearly average values for census tracts divided into high and low percent Latino. Census tracts >33% Latino were classified as high Latino census tracts.

	2014		2015		2016		2017	
	Low	High	Low	High	Low	High	Low	High
Diabetes	8.63	11.53	8.22	11.19	8.49	11.15	8.4	11.06
$PM_{2.5}$	10.58	11.90	9.59	10.91	9.48	10.63	10.63	11.53
Unemployment	5.96	8.50	5.33	7.62	4.69	6.69	4.13	5.83
Education	8.84	31.71	8.68	31.17	8.57	30.70	8.38	29.88
Poverty	11.47	23.41	11.36	23.20	11.10	22.48	10.74	21.12

Table 3. Yearly Average Values for Census Tracts with a Majority Percent Latino. Yearly average values for census tracts divided into high and low percent Latino. Census tracts with a majority Latino population (>50%) were classified as high Latino census tracts.

	2014		2015		2016		2017	
	Low	High	Low	High	Low	High	Low	High
Diabetes	9.07	12.15	8.68	11.81	8.92	11.66	8.83	11.57
PM2.5	10.78	12.17	9.81	11.16	9.66	10.86	10.78	11.70
Unemployment	6.48	8.78	5.80	7.87	5.10	6.89	4.48	6.01
Education	11.61	38.07	11.41	37.42	11.26	36.82	11.00	35.83
Poverty	13.26	26.03	13.16	25.74	12.80	24.97	12.30	23.36

Then we ran the fixed effects model on each of these subgroups. When we use average $PM_{2.5}$ concentration in the model, there was no significant association ($p > 0.05$) between $PM_{2.5}$ and diabetes prevalence among any subgroup (Figure 6). When we ran this analysis incorporating the proportion of days in exceedance as the $PM_{2.5}$ value, the coefficient association between $PM_{2.5}$ and diabetes prevalence is larger among the high Latino subgroups (Figure 7).

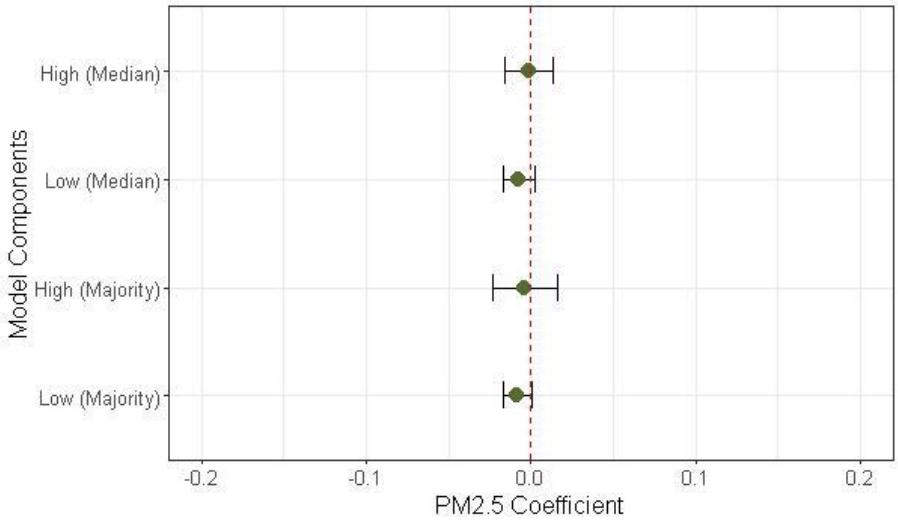


Figure 6. Latino Subgroup Fixed Effects Results 2014 - 2017. The coefficient association between average concentration of $PM_{2.5}$ and diabetes prevalence was null for high and low Latino census tracts, using both the majority and median values as the cutoff.

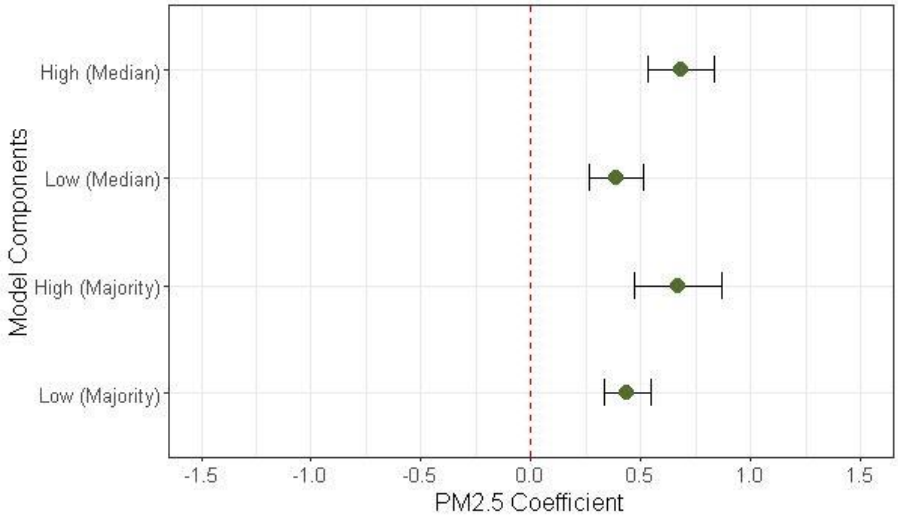


Figure 7. Latino Subgroup Fixed Effects Results 2014 - 2017 Using Thresholds. The coefficient association between proportion of exceedances of $PM_{2.5}$ and diabetes prevalence was more positive among high Latino census tracts.

Leave-One-Out Analysis

Our leave-one-out analysis explored the robustness of our fixed-effects results. We ran a series of fixed effects models for the 2014-2016 and 2014-2017 time frames, leaving out a specific county at a time. In all cases, this analysis highlighted the weight that Los Angeles County has on our result. When we use average concentration as the $PM_{2.5}$ value, in the 2014-2016 dataset our significant positive association became negative when Los Angeles County was removed from the model (β_1= -0.036, +/- 0.0058, p <0.001) (Figure 8). In the 2014-2017 time frame, the coefficient association is consistently close to zero, while leaving Los Angeles County out again shifts the coefficient to be negative (β_1= -0.093, +/- 0.0043, p <0.001) (Figure

9). It is evident that our results are strongly influenced by Los Angeles County. Additionally, the negative and significant effect in the remaining observations suggests a potential omitted variable bias or model misspecification that was not addressed by our fixed effects analysis.

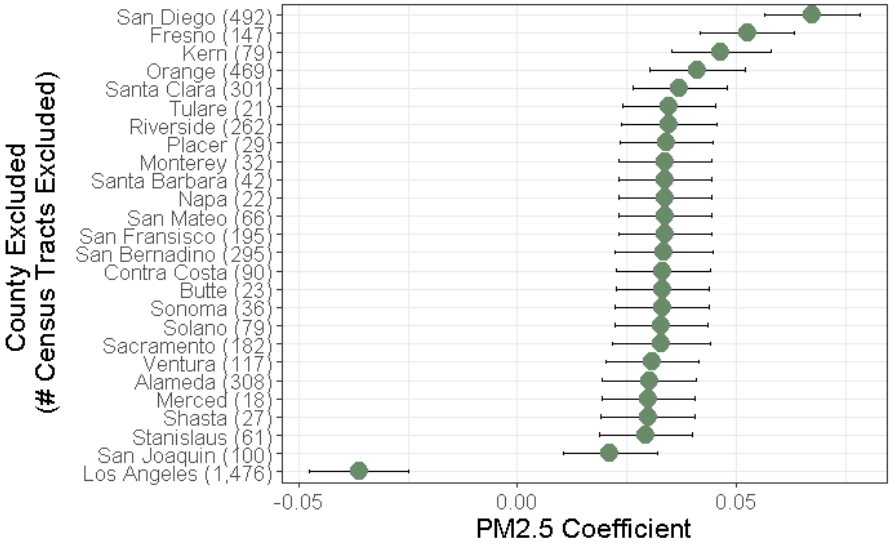


Figure 8. Leave-one-out Fixed Effects Results 2014 - 2016. Removing Los Angeles County shifts the coefficient to be negative while removing other counties maintains a positive coefficient.

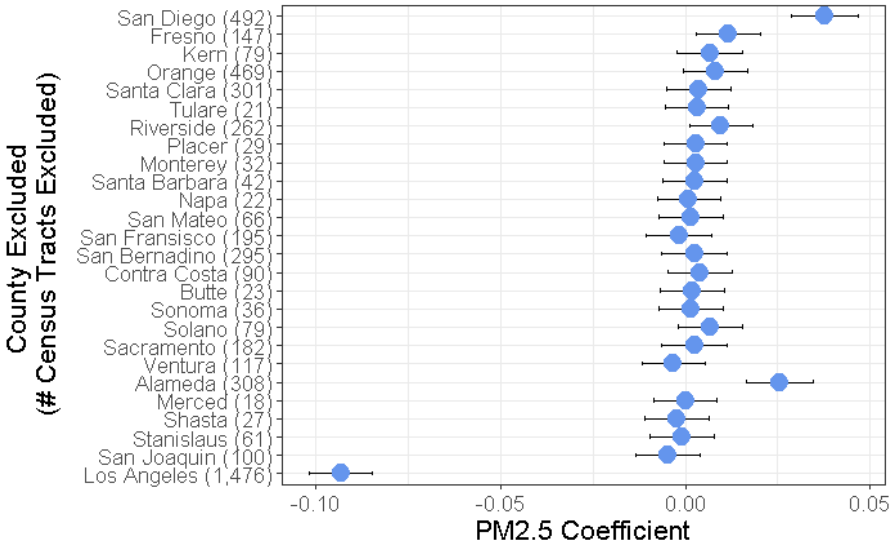


Figure 9. Leave-One-Out Fixed Effects Results 2014 - 2017. The coefficient maintains a null relationship even when counties are removed. This indicates the insignificant result in models incorporating 2017 data is not due to one particular county.

When we use proportion of exceedances as the $PM_{2.5}$ value, we still see the significant weight Los Angeles County has in the model. In the 2014-2016 dataset the coefficient association is consistently positive but becomes negative when Los Angeles County was removed from the model (β_1= -0.293, +/- 0.0679, p <0.001)(Figure 10). In the 2014-2017 time

frame, leaving Los Angeles County out again shifts the coefficient to be negative (β_1= -0.341, +/- 0.059, p <0.001)(Figure 11).

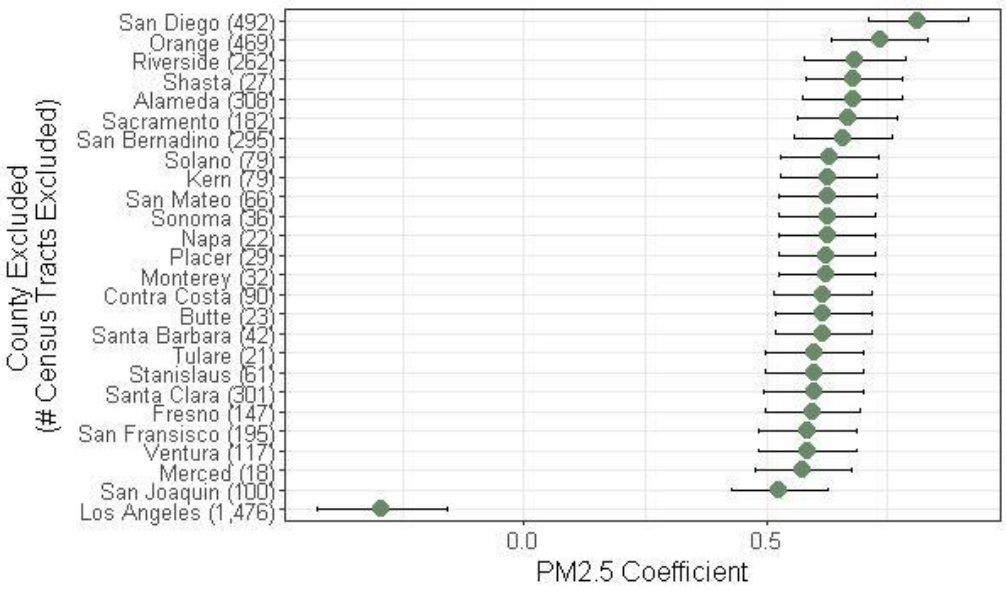


Figure 10. Leave-one-out Fixed Effects Results 2014 - 2016 Using Thresholds. Removing Los Angeles County shifts the coefficient to be negative while removing other counties maintains a positive coefficient.

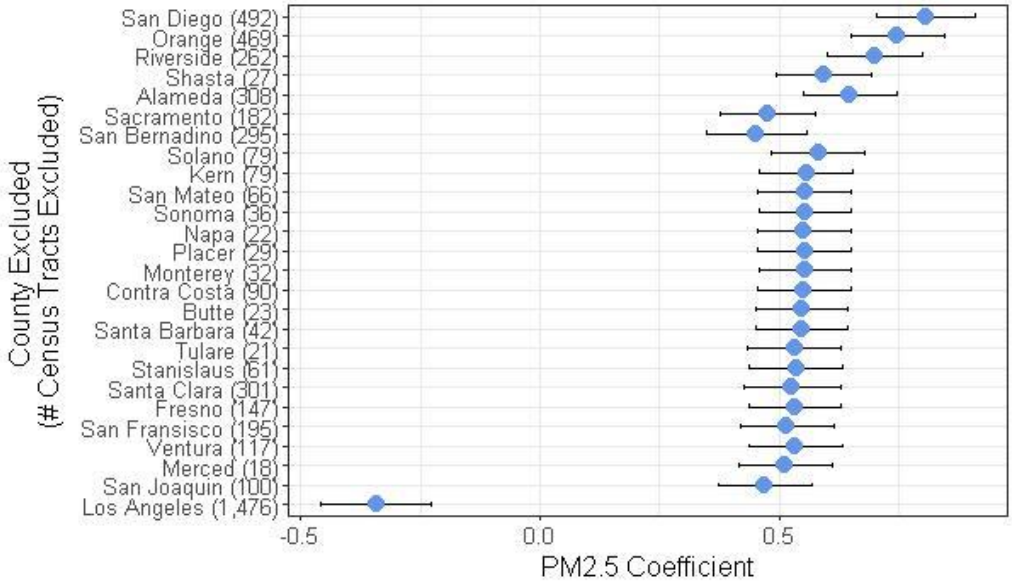


Figure 11. Leave-One-Out Fixed Effects Results 2014 - 2017 Using Thresholds. Removing Los Angeles County shifts the coefficient to be negative while removing other counties maintains a positive coefficient.

IV. Discussion

Our project aimed to assess the relationship between $PM_{2.5}$ and diabetes prevalence in California, USA, using a cross sectional and panel data approach. The two model types examined the years 2014 through 2017 to understand the possible association between diabetes and $PM_{2.5}$ in urban areas across the state.

Type 2 diabetes has been increasing in prevalence since the end of the 20th century. California is an especially interesting area to study diabetes and environmental factors because the state has high diabetes prevalence, high air pollution concentrations and an unequal distribution of pollution exposure. Understanding how and if environmental factors, such as air pollution, are associated with diabetes would allow healthcare providers in California to more effectively implement prevention techniques that go beyond diet, exercise, and prescriptions.

Our analysis returned several results. Cross sectional linear models for the years 2014-2016 show a positive association between average concentration of $PM_{2.5}$ and diabetes prevalence. Results from our fixed effects analysis are quantitatively similar. The 2017 cross sectional and 2014-2017 fixed effects model's results show no significant association.

The cross sectional model explored the relationship between $PM_{2.5}$ and diabetes by comparing across locations. The results from our cross sectional models in 2014, 2015, and 2016 indicate a positive relationship between average concentration of $PM_{2.5}$ and diabetes prevalence, with a 1 ug/m^3 increase in $PM_{2.5}$ increasing diabetes prevalence by approximately 0.06 percentage points. There are around 3 million people in California with diabetes, so a 0.06 percentage point increase in prevalence translates to around 1,800 additional cases (Health Rankings, 2019). The average annual medical expenditures of an individual with diabetes that can be attributed to the disease are $8,000 (Division of Diabetes Translation, 2017). Therefore, 1,800 additional cases of diabetes translates to an additional $14.5 million in healthcare costs statewide. However, comparing across locations using a cross sectional model is prone to omitted variable bias. To account for this bias, our analysis moved forward with a fixed effects model that incorporated a panel of data across multiple years.

Our initial fixed effects model used data from 2014-2016 and assessed the changes in average $PM_{2.5}$ concentration and diabetes that occurred within a given census tract. The fixed effects model returned coefficients of approximately 0.04. This means from every 1 ug/m^3 increase in $PM_{2.5}$, there is a 0.04 percentage point increase in diabetes prevalence. A 0.04 percentage point increase in prevalence translates to around 1,200 additional cases of diabetes (Health Rankings, 2019). This corresponds with $9.6 million in additional healthcare costs (Division of Diabetes Translation, 2017).

Interestingly, when we incorporated 2017 datasets, we saw a change in our results. Our cross sectional coefficient for the year 2017 shifted much closer to zero, with a value of 0.015. When this value was incorporated into our fixed effects model, the model results showed no significant association between $PM_{2.5}$ and diabetes prevalence. We hypothesize that several factors unique to 2017 could be responsible for this change.

In 2014-2016 the average statewide diabetes prevalence and $PM_{2.5}$ concentration values were decreasing on similar scales, while in 2017 diabetes prevalence kept decreasing and $PM_{2.5}$ increased, possibly leading to our result showing no association between the two variables. With two of the largest and deadliest wildfires on record, 2017 was an unusual year regarding $PM_{2.5}$ in California. Fires led to extreme spikes in $PM_{2.5}$ levels during the fall months of 2017 around Ventura, Los Angeles and Sonoma County (Census Bureau, 2017). However, we averaged these observations across the entire year which removed $PM_{2.5}$ spikes from the dataset and performed a sensitivity analysis that did not suggest the anomalies in 2017 were localized to any particular county, contradicting the idea that wildfires are responsible for the model results. It is still possible that those exposed to $PM_{2.5}$ in the fall wildfires did not have adequate time before the end of the year to report their diabetes cases to their doctor. More recent diabetes data is needed to assess the impacts of these wildfires.

Socioeconomic conditions also improved significantly from 2016 to 2017. Across all census tracts there was nearly a 1% decrease in the percent of people living in poverty as well as nearly a 1% decrease in unemployment. While we control for several sociodemographic conditions, we may not have captured all relevant drivers of diabetes that may be correlated with $PM_{2.5}$. These conditions would affect results from the cross sectional model in 2017, and fixed effects models. With a fixed effects model, we assume that unobserved characteristics are time-invariant at a given census tract and would thus drop out of the model. Unobserved socioeconomic conditions within a census tract that change significantly by 2017 would not drop out of the model and would affect results. Among lower income groups in particular, we see improvements in economic conditions by 2017.

In California, the Medicaid program expanded in 2014, causing the percentage of people without insurance to decline. In 2013, 17.2% of the state was uninsured, while only 7.2% were uninsured in 2017 (ACS, 2019). It is possible that this policy change and increased accessibility of healthcare, particularly among those with lower incomes, would decrease diabetes prevalence at specific census tracts in a dramatic way by the year 2017 that would cause conflicting model results when 2017 data is incorporated (Appendix 7). It is also possible that increased access to healthcare would increase access to screenings where people could be diagnosed with diabetes, leading to an increase in diabetes prevalence. Finally, it is important to note that Medicaid only covers documented individuals, thus excluding many Latinos.

There is also inconclusive literature on the timescale between air pollution and health outcomes, specifically $PM_{2.5}$. A report released by the EPA demonstrated that adverse health effects from PM were observed with lags ranging from days to years for different health outcomes (EPA, 2019). Due to the shorter time period of our study and a lack of information on the timescale of impacts, we might not be capturing an increase in diabetes cases from the 2017 wildfires within the same year. To further explore timescales, it would be useful to include a range of different time lags in future studies. Similarly, misspecification of the functional form of this relationship could be influential.

The literature is ambiguous with regard to the existence or magnitude of threshold values of exposure. In the cross sectional and fixed effects models, $PM_{2.5}$ is given as an annual

average concentration for each census tract, but it is also possible that what matters is exceedances of a particular threshold. Studies related to the functional form of $PM_{2.5}$ and health effects are not conclusive. Some research finds that chronic exposure to $PM_{2.5}$ at any level can lead to health impacts (Andersen, et al., 2015). Other studies show a non-linear response suggesting that the overall acute effects consist of two discrete patterns: a short-term response (2 to 15 days) where mortality risks decrease to near null values after the air-pollution event; or an intermediate timescale pattern (16 to 55 days) where mortality risk climbs to positive levels weeks after the event (Valari, et al). We sought to explore the possibility of thresholds as peaks of elevated $PM_{2.5}$ using the number of daily observations at a monitoring location that exceed NAAQs. We found that in all cross sectional and fixed effects models from 2014-2017 that $PM_{2.5}$ and diabetes prevalence had a significantly positive association.

Diabetes is discriminatory, disproportionately impacting Latinos in crude prevalence and mortality (Division of Diabetes Translation, 2017, Golden et al., 2012). We explored interactions between $PM_{2.5}$ and Latino populations to assess if the association between $PM_{2.5}$ changes with demographics. We found in our subgroup analysis that when we incorporate $PM_{2.5}$ as a proportion of days in exceedance of the NAAQs, that the coefficient association is larger among census tracts with large Latino populations. However, this subgroup analysis is a preliminary look at possible relationships incorporating race/ethnicity. This should be further explored in additional models using interaction effects or different functional forms.

Our leave-one-out analysis provided interesting insights. Our 2014-2017 average concentration analysis showed that even when leaving out certain counties, the resulting coefficients were still insignificant. This indicates that the insignificant relationship is not due to one particular county. In all analyses, leaving Los Angeles County out of the fixed effects models caused the coefficient to shift to a significantly negative association. This result may have been influenced by the high proportion of census tracts belonging to Los Angeles County in our dataset. However, this interesting result highlights the need for further exploration into the association in Los Angeles County, and why it is driving our coefficient outcome statewide. Misspecification of the relevant time of exposure could explain why the results are so strongly influenced by Los Angeles County and why counterintuitive negative coefficients are observed when Los Angeles County is removed from the analysis.

Our results showed significant positive associations between diabetes prevalence and $PM_{2.5}$ concentration in most models, and no association under others. We hypothesize that these conflicting trends could be due to changes in $PM_{2.5}$ distribution or socioeconomic conditions in 2017, or due to time lags that affect health outcomes. There are many different directions that additional studies could take in developing more robust models that incorporate $PM_{2.5}$ and diabetes.

V. Conclusion

Diabetes is a global epidemic. In the U.S., nearly 1 in 10 Americans (9.4% of the population or 30.3 million people) live with diabetes (95% of which is type 2 diabetes), with 1.5 million more diagnosed every year (ADA, 2018). Diabetes is the seventh leading cause of death in the U.S. In 2017, diagnosed diabetes cost the U.S. $327 billion (ADA, 2018). In our analysis we assessed the relationship between $PM_{2.5}$ and diabetes prevalence in California using a cross sectional and panel data approach.

Our results suggest a possible positive relationship between exposure to $PM_{2.5}$ and diabetes prevalence in California. In 2014-2016 we see a significant positive association between average concentration of $PM_{2.5}$ and diabetes prevalence while in models incorporating 2017 data we see no relationship. Further, changes in socioeconomics from 2017 could have an influence. There is also an unknown timescale of the interaction between diabetes and $PM_{2.5}$ and because of the short time period of our study, it is possible that we are not capturing the effects of $PM_{2.5}$ on diabetes. Although these results are preliminary, when we use the proportion of observations in exceedance of NAAQs as the $PM_{2.5}$ measurement we do see consistently positive associations.

There is plenty of room for new and more robust research in this area. Future studies could explore thresholds and the associations between spikes of $PM_{2.5}$ and diabetes prevalence in a more powerful way. Research could also focus on the association in Los Angeles County. Additional years of diabetes prevalence data will be released under 500 Cities and will allow future studies to capture different time lags and functional forms. In California, diabetes and air pollution have continuously been at the forefront of health and environmental initiatives. Understanding the possible relationship between diabetes and environmental factors could have important implications for prevention and treatment initiatives in the future.

VI. Acknowledgments

We thank the James S. Bower foundation for their funding to this project. We would also like to acknowledge our advisors at the Bren School of Environmental Science & Management, Ashley Larsen, Patrick Hunnicutt, Olivier Deschenes, and Kyle Meng, and our clients at Sansum Diabetes Research Institute, Namino Glantz and David Kerr, and at Groundswell Technologies, Mark Kram.

VII. Appendix

Appendix 1

$PM_{2.5}$ and Diabetes Prevalence

Table A1. Literature Review Exploring Relationships Between *$PM_{2.5}$* and Diabetes Prevalence. Rows in green found a positive association between *$PM_{2.5}$* and diabetes prevalence. Rows in yellow show studies where results were partly consistent with a link between long-term exposure to air pollution and the risk of diabetes. Red highlighted studies display no evidence of an association between the two variables.

Title, Author, Date	Location of Study	Sample Size
Brook RD, Cakmak S, Turner MC, et al. Long-term fine particulate matter exposure and mortality from diabetes in Canada. Diabetes Care 2013; 36: 3313–3320.	Canada	2,145, 400 participants
Chen H, Burnett RT, Kwong JC, et al. Risk of incident diabetes in relation to long-term exposure to fine particulate matter in Ontario, Canada. Environ Health Perspect 2013; 121: 804–810.	Ontario, Canada	6,310 incident cases of diabetes over 484,644 total person-years of follow-up
Coogan PF, White LF, Jerrett M, et al. Air pollution and incidence of hypertension and diabetes mellitus in black women living in Los Angeles. Circulation 2012; 125: 767–772.	Los Angeles, Ca	3,992 participants - all African American
To T, Zhu J, Villeneuve PJ, et al. Chronic disease prevalence in women and air pollution-A 30-year longitudinal cohort study. Environ Int 2015; 80: 26–32.	Ontario, Canada	29,549 participants - women only
Weinmayr G, Hennig F, Fuks K, et al. Long-term exposure to fine particulate matter and incidence of type 2 diabetes mellitus in a cohort study: effects of total and traffic-specific air pollution. Environ Health 2015; 19: 53.	Heinz Nixdorf, Germany	3,607 participants
Park SK, Adar SD, O'Neill MS, et al. Long-term exposure to air pollution and type 2 diabetes mellitus in a multiethnic cohort. Am J Epidemiol 2015; 181: 327–336.	Chicago, New York, and St. Paul, USA	5,839 participants
Hu H, Ha S, Henderson BH, et al. Association of Atmospheric Particulate Matter and Ozone with Gestational Diabetes Mellitus. Environ Health Perspect 2015; 123: 853–859.	Florida, USA	410,267 women - gestational diabetes
Robledo CA, Mendola P, Yeung E, et al. Preconception and early pregnancy air pollution exposures and risk of gestational diabetes mellitus. Environ Res 2015; 137: 316–322.	Springfield, Massachusetts; Los Angeles, California; Newark, DE; Washington, DC; Indianapolis, Indiana; Salt Lake City, Utah; Brooklyn, New York; Cleveland, Ohio; Akron, Ohio	219,952 women

Fleisch AF, Kloog I, Luttmann-Gibson H, et al. Air pollution exposure and gestational diabetes mellitus among pregnant women in Massachusetts: a cohort study. Environ Health 2016; 24: 40.	Massachusetts, USA	gestational diabetes - 159,373 women
Fleisch AF, Gold DR, Rifas-Shiman SL, et al. Air pollution exposure and abnormal glucose tolerance during pregnancy: the project Viva cohort. Environ Health Perspect 2014; 122: 378–383.	Boston, USA	2,093 women
Puett RC, Hart JE, Schwartz J, et al. Are particulate matter exposures associated with risk of type 2 diabetes? Environ Health Perspect 2011; 119: 384–389.	Northeastern and Midwestern, USA	74,412 participants

Appendix 2

Diabetes by Race in California

Table A2. Diabetes Prevalence by Race, California vs. National Averages. Racial/ethnic groups do not include Hispanic/Latinos, except for Hispanics/Latinos of any race (Health Rankings, 2019).

Racial Group	Diabetes Prevalence California	Diabetes Prevalence United States
Overall	10.4%	10.9%
White/Caucasian	8.1%	10.7%
Black/African American	14.8%	14.9%
Asian	10.6%	9.2%
Native American/Alaska Native	24%	11.7%
Hispanic/Latino of Any Race	11.8%	11.3%

Appendix 3

Kriging Surface in GIS

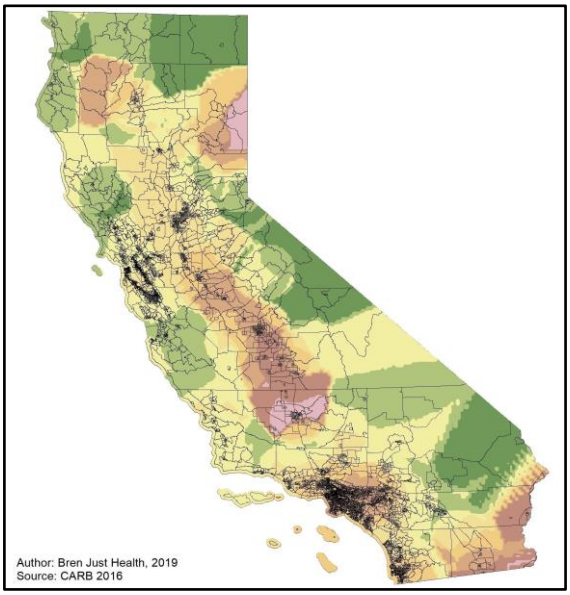
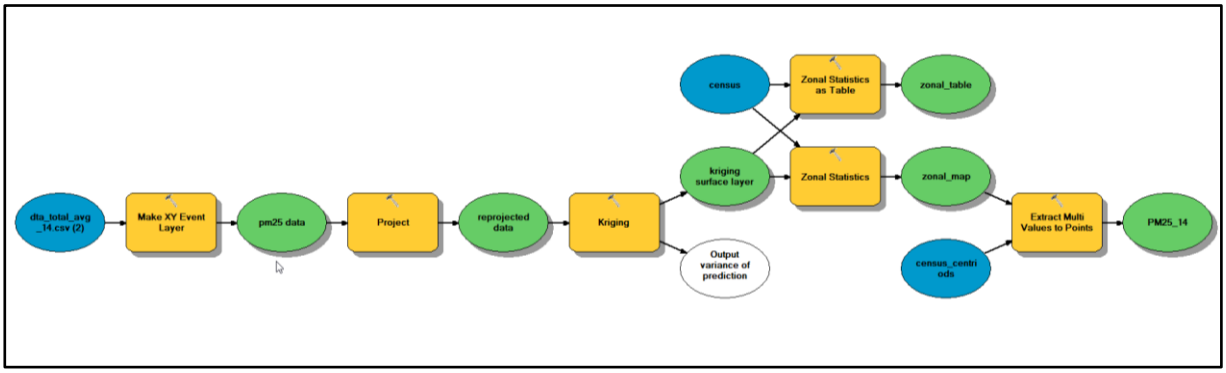


Figure A1. Example of Continuous Kriging Surface Over California. Once this surface was created, values were averaged over each census tract and the average was attached to the census tract centroid.

Appendix 4



GIS Model

Figure A2. ArcGIS Ordinary Kriging Spatial Interpolation Model of $PM_{2.5}$ data. Monitoring station data was provided with latitude and longitude coordinates. These stations were mapped, then reprojected into NAD 1983 Teale Albers (m). Ordinary kriging was used with a search parameter of a 50 km radius. If no monitoring station was found within 50 km, the next nearest station was used.

Appendix 5

Model Verification Figures and Results

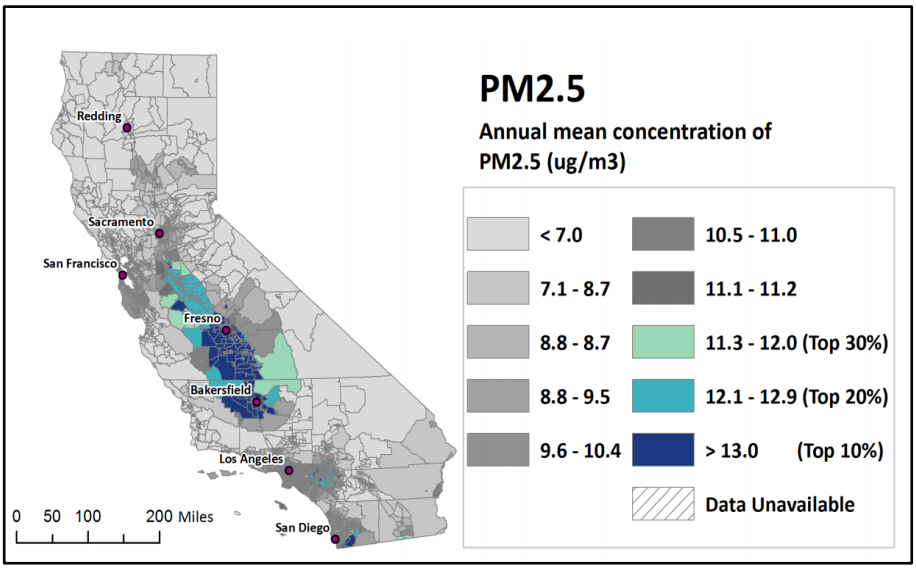


Figure A3. CES 3.0 $PM_{2.5}$ Values for 2012-2014. Concentration of $PM_{2.5}$ in each census tract calculated as an average from 2012-2014. We used this dataset for comparison of our kriging methodology. Source: CES 3.0

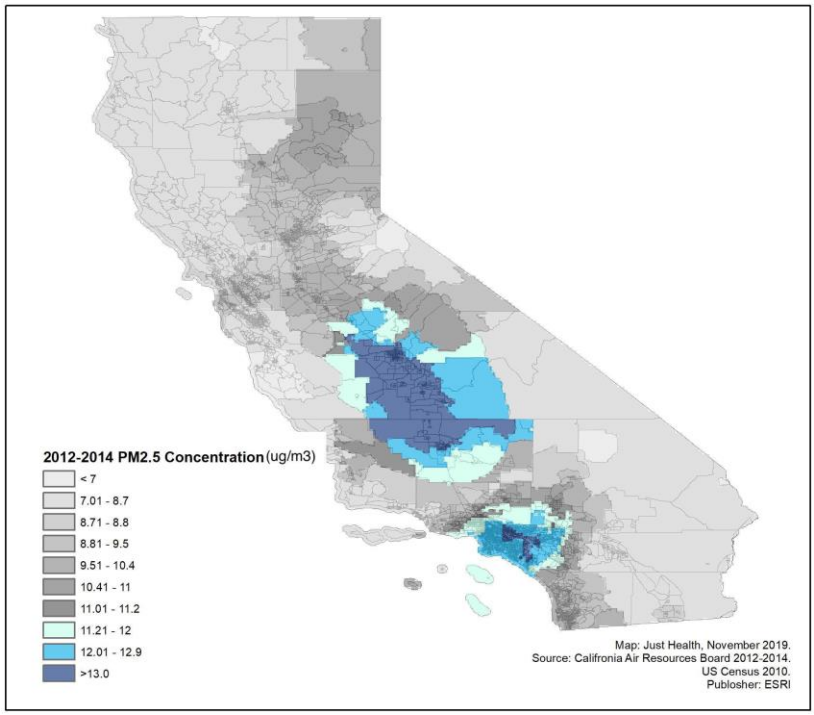


Figure A4. Calculated $PM_{2.5}$ Values for 2012-2014. We averaged our $PM_{2.5}$ dataset across 2012-2014 and used kriging in GIS to calculate one average $PM_{2.5}$ concentration per census tract.

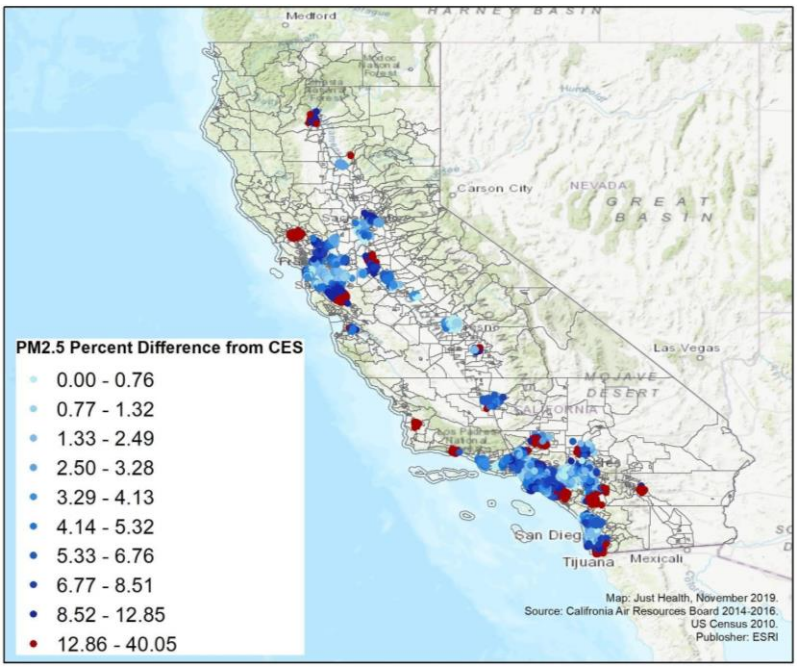


Figure A5. Percent Difference Between Dataset and CES $PM_{2.5}$. Points represent the centroid of census tracts where we have both $PM_{2.5}$ data and diabetes data. Red values represent a difference in concentrations given by CES and our results above 12.86%.

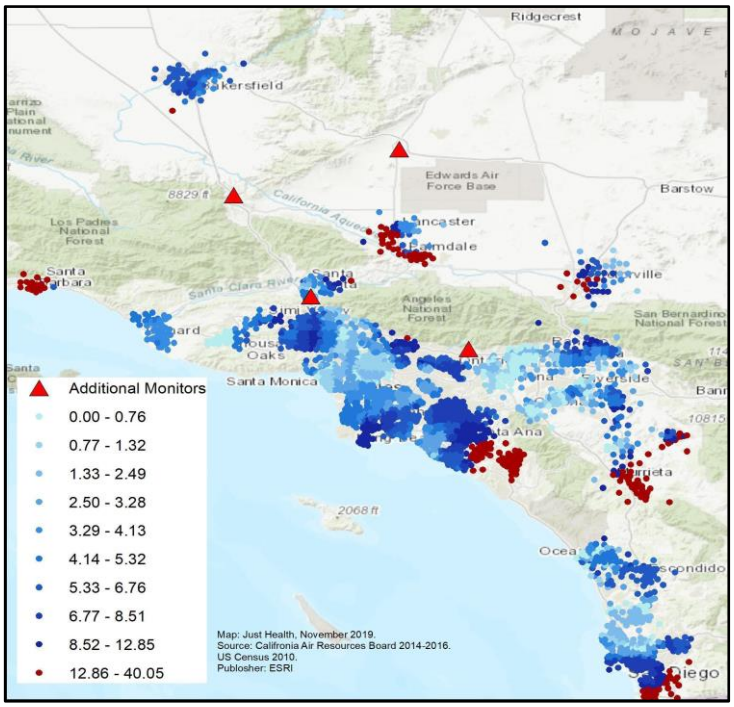


Figure A6. Case Study LA area: Percent Difference Between Dataset and CES $PM_{2.5}$ Values. Points represent the centroid of census tracts where we have both $PM_{2.5}$ data and diabetes data within the Los Angeles area.

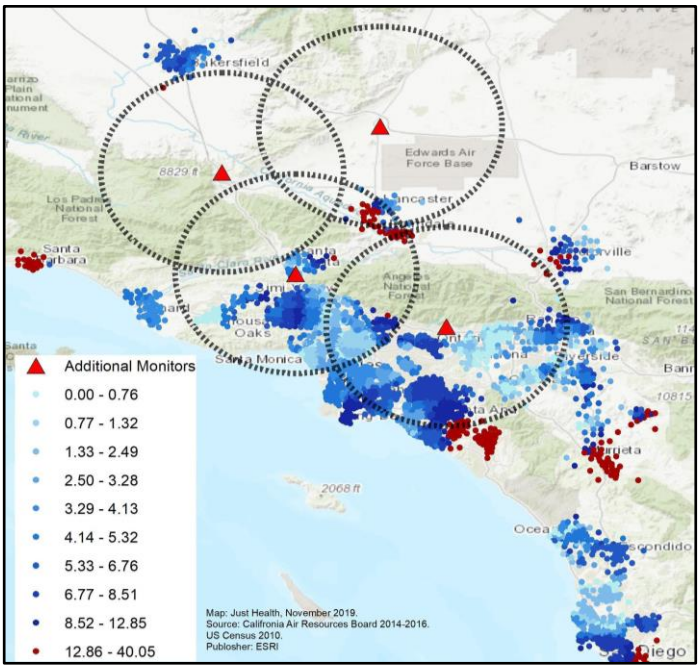


Figure A7. Case Study LA area: Percent Difference Between Dataset and CES $PM_{2.5}$ Values. 50km buffers are drawn around our dataset's additional air monitoring stations.

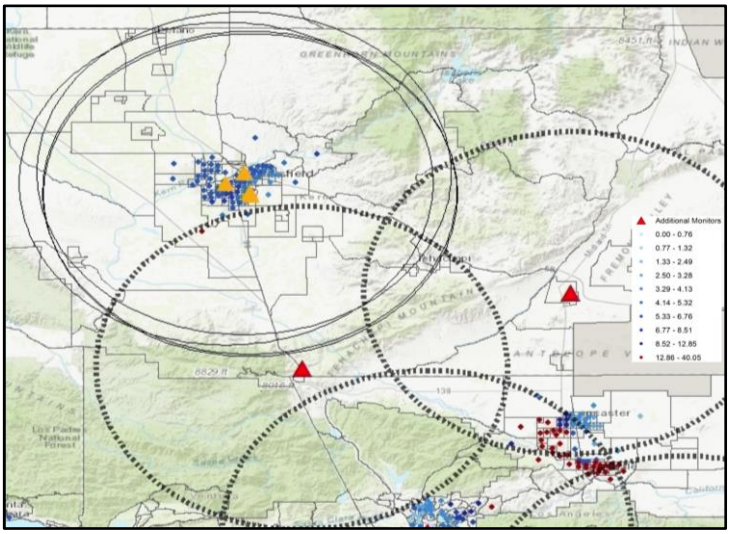


Figure A8. Case Study LA area: Percent difference between dataset and CES $PM_{2.5}$ Values. Solid lines represent a 50 km radius drawn around common (both our dataset and CES) air monitoring stations. Dotted lines represent a 50 km radius of our additional air monitoring stations.

Appendix 6

Collinearity among Socioeconomic Variables

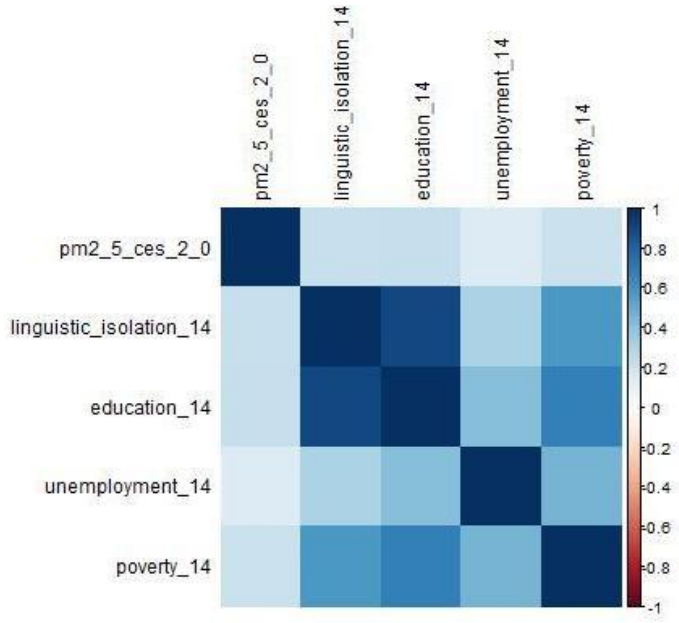


Figure A9. Collinearity among Socioeconomic Variables 2014. In 2014, there is the highest positive collinearity between education and linguistic isolation. There is not much of a collinear relationship with $PM_{2.5}$ and the sociodemographic variables. These trends are the same in the 2015, 2016, and 2017 datasets.

Appendix 7

Changes in Socioeconomic Variables 2014-2017

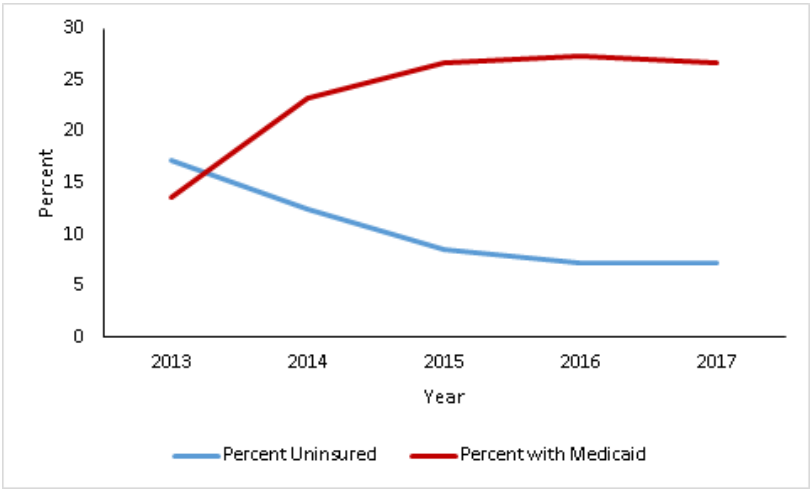


Figure A10. Trends in Health Insurance Coverage. In 2014, California expanded their Medicaid program under the Affordable Care Act. The percent of Californians with Medicaid increased dramatically after this policy change, corresponding with a decrease in the percent of Californians with no insurance.

Supplemental Information

Diabetes

Diabetes is a threat to human health around the world and its prevalence is increasing. Globally, an estimated 422 million adults were living with diabetes in 2014, compared to 108 million in 1980 (WHO, 2016). In 36 years, global diabetes prevalence almost doubled from 4.7% (1980) to 8.5% (2014) (WHO, 2016). Diabetes arises when the pancreas does not produce enough insulin, or when the body cannot effectively use the insulin it produces (WHO, 2016). This defect causes blood glucose (sugar) levels to rise higher than normal.

In 2015, there were 250,000 diabetes-related deaths recorded and on average the disease decreases life expectancy by 8.5 years (Division of Diabetes Translation, 2017). Diabetes and its complications bring about substantial economic loss to people with diabetes, their families, health systems and national economies through direct medical costs and loss of work and wages (WHO, 2016). While very serious, type 2 diabetes can be prevented and successfully managed. Current treatments include environmental adjustments, lifestyle changes, oral medications, and insulin injections. Behavioral and social factors that influence diabetes prevalence are well-known (Figure S1).

Environmental conditions are becoming an increasingly studied potential risk factor. To better visualize the interactions between risk factors, Figure S1 outlines the complexities of the environmental and behavioral actions that lead to an increased risk of diabetes (Dendup et al., 2018).

Diabetes in California

Compared to the rest of the U.S. which has an overall diabetes rate of 10.9%, California shows slightly lower levels of diabetes at 10.4% (Health Rankings, 2019). However, over the last ten years, diabetes prevalence increased by 35% throughout the state (Health Rankings, 2018). Education seems to be associated with diabetes prevalence. The group of adults with the highest diabetes rate in California (18.4%) has an education level of “less than high school”. This educational attainment group makes up 17.5% of California’s population (Health Rankings, 2019). It is also significantly more common in adults living below 100% of the federal poverty level (FPL) than in those with incomes at or above 300% FPL (7.8% vs. 4.5% respectively) (Diamant et al., 2003).

Trends also differ by race/ethnicity. Nationally, Hispanics have higher rates of end-stage renal disease caused by diabetes, and they are 40% more likely to die from diabetes than non-Hispanic whites (Office of Minority Health, 2016). Furthermore, Latinos are more highly affected than non-Hispanic whites within each age group (Diamant et al., 2003).

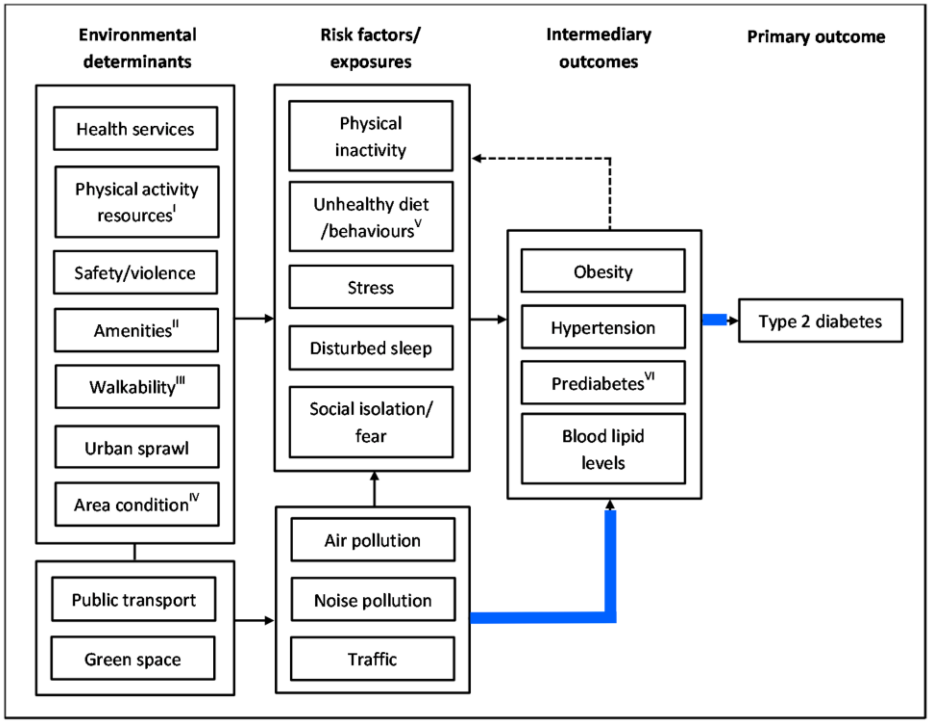


Figure S1. Schematic Diagram of Possible Pathways that Influence Diabetes Prevalence. Literature has explored a variety of relationships between environmental, health, and demographic indicators associated with diabetes. We will explore the pathway highlighted in blue by focusing on $PM_{2.5}$ pollution. Source: Dendup et al. 2018.

Proposed Mechanisms of $PM_{2.5}$ *Effects on Diabetes*

The specific mechanism in which air pollution interacts with type 2 diabetes is still not fully understood, especially at the molecular and cellular level. The timeline of impacts is also not understood in the literature. $PM_{2.5}$ may stimulate oxidative and inflammatory responses in the lungs that affect the function of other organs, or particulates may be translocated to central nervous system receptors (Dimakakou et al., 2018). One study suggests that at a cellular level, $PM_{2.5}$ contributes to insulin resistance and type 2 diabetes through disrupting the CC-chemokine receptor 2 pathway which regulates visceral adipose inflammation and by triggering the "unfolding protein response" within a cell's endoplasmic reticulum (Feng et al., 2016).

$PM_{2.5}$ *State and Federal Air Quality Standards*

Currently, annual National Ambient Air Quality Standards (NAAQS) for $PM_{2.5}$ across the U.S.are 12 ($\mu g/m^3$). Policy documents are now taking note of the effects that low levels of $PM_{2.5}$ could have on health. The EPA recently released a Policy Assessment of the NAAQS for PM in September 2019. This draft states that based on the available scientific evidence, air quality analyses and risk assessments, the current primary annual standard for $PM_{2.5}$ may not be adequate (EPA, 2019).

$PM_{2.5}$ particles are released into the atmosphere from a range of anthropogenic sources. These sources include cars and trucks, factories, and burning wood (USEPA, 2019). Natural

sources of PM include dust from the wind erosion of natural surfaces, sea salt, wildland fires, and primary biological aerosol particles (Figure S2).

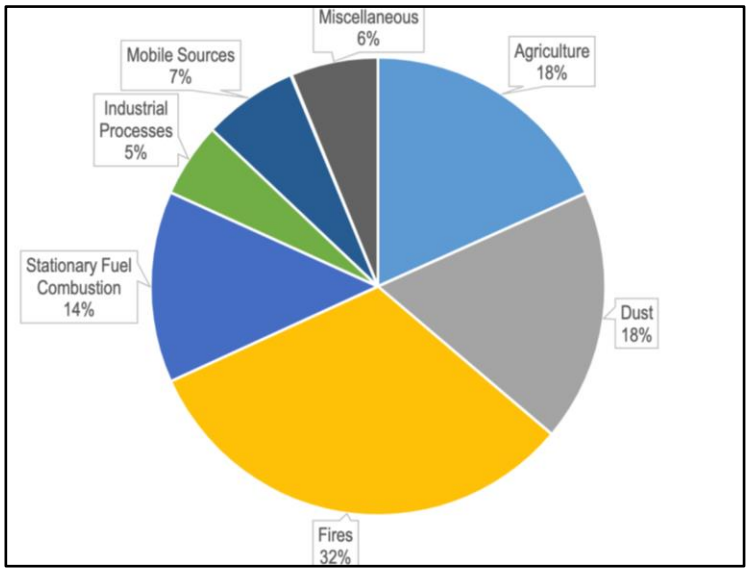


Figure S2. Percent Contribution of $PM_{2.5}$ Emissions by Source. Significant emissions of $PM_{2.5}$ come from both anthropogenic and natural sources. Source: 2014 National Emissions Inventory

$PM_{2.5}$ and Other Health Impacts

In addition to diabetes, particulate matter exposure has been associated with a range of other health impacts. Particulate matter exposure in the workplace has been associated with neurodegenerative diseases such as Alzheimer's and dementia (Jung et al., 2015). Short term increases of $PM_{2.5}$ raise the incidence of acute cardiovascular events like heart attacks by 1-3% within a few days; with long term exposure, this number increases to 10% and the development of chronic cardiovascular diseases like hypertension increase (Rajagopalan et al., 2018).

In 2019 the EPA released a report that analyzed health outcomes where evidence supports either a causal, likely to be causal or a suggestive relationship with $PM_{2.5}$. Between the years of 2009 and 2018, none of the listed causal relationships between $PM_{2.5}$ and a given health outcome have been downgraded, while cancer, nervous system effects, and metabolic effects have seen a strengthening in causality determinations (Table S1).

Table S1. Key Causality Determinations for $PM_{2.5}$ in 2009 and 2018. Exposures to $PM_{2.5}$ have been associated with a range of health outcomes. In 2018, metabolic effects (which includes diabetes) are now considered to have a suggestive relationship with $PM_{2.5}$ (EPA, 2019).

Health Outcome	Exposure Duration	2009	2018
Mortality	Long-term	Causal	Causal
	Short-term	Causal	Causal
Cardiovascular Effects	Long-term	Causal	Causal
	Short-term	Causal	Causal

Respiratory Effects	Long-term	Likely to be causal	Likely to be causal
	Short-term	Likely to be causal	Likely to be causal
Cancer	Long-term	Suggestive of, but not sufficient to infer	Likely to be causal
Nervous System Effects	Long-term	---	Likely to be causal
	Short-term	Inadequate	Suggestive of, but not sufficient to infer
Metabolic Effects	Long-term	---	Suggestive of, but not sufficient to infer
	Short-term	---	Suggestive of, but not sufficient to infer
Reproduction and Fertility	Long-term	Suggestive of, but not sufficient to infer	Suggestive of, but not sufficient to infer
	Short-term	Suggestive of, but not sufficient to infer	Suggestive of, but not sufficient to infer

Assessment of Residuals

We used a fixed effects panel regression as a more statistically rigorous method than cross sectional models to explore the relationship between diabetes prevalence and $PM_{2.5}$ values. In order to run a fixed effects model, there must be variation in the $PM_{2.5}$ and diabetes prevalence datasets that are not purely a function of census tract and time. A linear model was used to compare CES 2.0/3.0 and CDC data from 2014/2016 incorporating time, then census tract and time. Only census tracts with diabetes data are included in the model. The format of these linear models is reported below.

First, we assess whether there is variation in the $PM_{2.5}$ and diabetes datasets that is not purely a function of time. We would expect this to be true because certain areas of California have consistently higher concentrations of $PM_{2.5}$ than others. This step is not important for model verification but serves as a quantitative check on our understanding of $PM_{2.5}$ distribution across the state. We used the following linear models showing $PM_{2.5}$ or diabetes prevalence as a function of time:

$$PM_{2.5\,t} = \beta_1\ CES\ Version_{\ t}$$

$$Diabetes\ Prevalence_{\ t} = \beta_1\ Year_{\ t}$$

Next, we assess whether there is variation in the $PM_{2.5}$ and diabetes datasets that is not purely a function of time and location. This needs to be true in order for the fixed effects model to return accurate results. The fixed effects model will analyze variation within each census tract across the years in which we have panel data. The variables that do not change considerably

within a census tract across the years of study will be dropped. We used the following linear model showing $PM_{2.5}$ or diabetes prevalence as a function of location and time:

$$PM_{2.5\,it} = \beta_1\, CES\ Version_{\,t} + \beta_2\, Census\ Tract_{\,i}$$

$$Diabetes\ Prevalence_{\,it} = \beta_1\, Year_{\,t} + \beta_2\, Census\ Tract_{\,i}$$

Calculating the standard deviation of linear model residuals indicates that there is variation in both the $PM_{2.5}$ dataset from CES and diabetes prevalence dataset from CDC that is not purely a function of census tract and time. The standard deviation of residuals when the location is controlled for is much less than the standard deviation of residuals when only time is considered in the model. This is expected because, during the same year, different locations in the state have very different $PM_{2.5}$ concentrations and diabetes prevalence. These findings indicate the fixed effects model is likely to be reasonably statistically precise (Table S2).

Table S2. Standard Deviation of Residuals in $PM_{2.5}$ and Diabetes Data. There is significant variation in $PM_{2.5}$ concentration and diabetes prevalence that exists within a census tract between years, which justifies our choice in using a fixed effects model.

	$PM_{2.5}$	Diabetes
SD	2.29	2.83
SD of Residuals (year)	2.28	2.82
SD of Residuals (census tract + year)	0.523	0.395

Comparison of sociodemographic Variables in Cross-Sectional Model

Cross sectional models allow us to explore the relationship between $PM_{2.5}$, diabetes and combinations of socioeconomic variables with minimal data wrangling. We examine how coefficient $\beta 1$ changes when different sociodemographic variables are included. However, since the cross sectional models examine data from only one year at a time, the results are more prone to omitted variable bias. For each cross section, two types of linear models are shown in the coefficient plot:

1. $PM_{2.5}$ - this model assessed diabetes prevalence at a census tract as a function of $PM_{2.5}$ concentration at that census tract.
2. $PM_{2.5}$ + socioeconomic - this model assessed diabetes prevalence at a census tract as a function of $PM_{2.5}$ concentration and the following sociodemographic variables:
    - Educational attainment
    - Poverty rate
    - Unemployment rate
    - Race (African American, Native American, and Latino of any race)

We see this coefficient is largest (β_1=~0.25) when no sociodemographic variables are incorporated into the model but remains positive when they are (Figure S3). A coefficient of 0.25

represents a 0.25 percentage point increase in diabetes crude prevalence when $PM_{2.5}$ concentration increases by one unit (ug/m^3). This value is in line with and on the same order of magnitude of other associations between $PM_{2.5}$ and diabetes prevalence reported in the literature (Pearson et al., 2010). The coefficient is reduced to approximately 0.06 when sociodemographic variables are included in the cross section (Figure S3). A coefficient of 0.06 represents a 0.06 percentage point increase in diabetes crude prevalence when $PM_{2.5}$ concentration increases by one unit (ug/m^3). When we incorporate socioeconomic variables into the cross sectional model, we see the coefficient decrease due to a positive association between socioeconomic indicators like poverty rate and unemployment rate with $PM_{2.5}$ that is not accounted for in the basic cross section. Among all cross sections except for 2017 with socioeconomic variables, the $PM_{2.5}$ coefficient indicates a positive association between $PM_{2.5}$ concentration and diabetes prevalence even when standard error, represented as the 95% confidence interval, is incorporated.

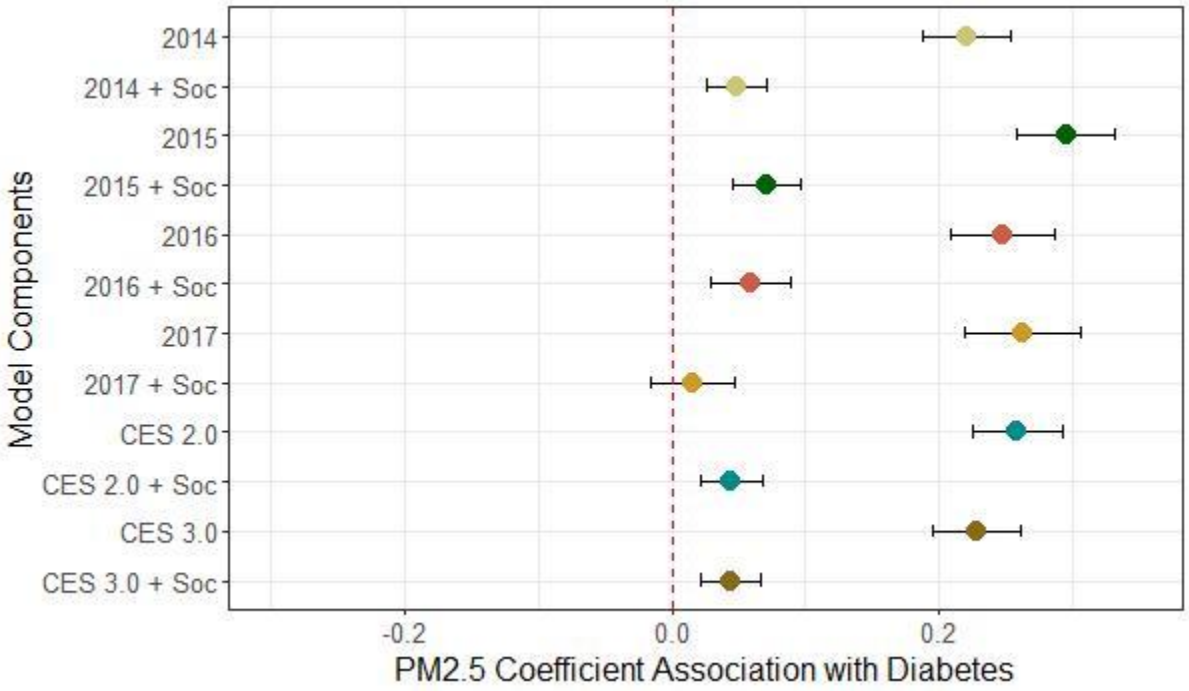


Figure S3. Coefficient Plot Comparison Between Cross Sections with and without Sociodemographic Variables. When no sociodemographic variables are included in the cross sectional model, coefficient associations are more positive than when sociodemographic variables are incorporated into the model.

This nonlinear method is a basic exploration that needs to be refined to draw more dependable results. Using ordinary kriging as a spatial interpolation method on annual proportions is not the ideal methodology, however due to time constraints other methods were not feasible. It would be more appropriate to use kriging on each day of the year and then combine these daily values to find a single annual value that could be incorporated into the models. Thresholds can be a useful way to look at non-linear relationships between diabetes prevalence and $PM_{2.5}$, however, conclusive results require using a more robust methodology.

VIII. References

AirNow. “Particle Pollution (PM),” 2017. Online. Internet. 15 Dec. 2019. Available: https://cfpub.epa.gov/airnow/index.cfm?action=aqibasics.particle [Accessed: 27 April 2019].

American Diabetes Association 2015. Facts About Type 2 [Online]. Available at: http://www.diabetes.org/diabetes-basics/type-2/facts-about-type-2.html [Accessed: 27 April 2019].

American Diabetes Association 2018. Facts About Type 2 [Online]. Available at: https://www.diabetes.org/resources/statistics/statistics-about-diabetes [Accessed: 18 Dec 2019].

American Lung Association. 2017. State of the Air, Most Polluted Cities. Retrieved March 10, 2020, from https://www.lung.org/our-initiatives/healthy-air/sota/city-rankings/most-polluted-cities.html

American Community Survey. “Data Tables on data.census.gov”. Available: https://www.census.gov/acs/www/data/data-tables-and-tools/american-factfinder/ [Accessed 2020 Mar 03].

Andersen, Z.J., Raaschou-Nielsen, O., Ketzel, M., Jensen, S.S., Hvidberg, M., Loft, S., Tjønneland, A., Overvad, K. and Sørensen, M. 2012. Diabetes incidence and long-term exposure to air pollution: a cohort study. *Diabetes Care;* 35(1), pp. 92–98.

Babey, S.H., Wolstein, J., Diamant, A.L. and Goldstein, H. 2016. Prediabetes in california: nearly half of california adults on path to diabetes. *Policy brief (UCLA Center for Health Policy Research);* (PB2016-1), pp. 1–8.

Bowe, B., Xie, Y., Li, T., Yan, Y., Xian, H. and Al-Aly, Z. 2018. The 2016 global and national burden of diabetes mellitus attributable to PM2·5 air pollution. *The lancet. Planetary health;* 2(7), pp. E301–E312.

Boyd-Barrett, C. 2019. People of Color and the Poor Disproportionately Exposed to Air Pollution, Study Finds. *California Health Report*. Available: www.calhealthreport.org/2019/02/08/people-of-color-and-the-poor-disproportionately-exposed-to-air-pollution-study-finds/.

Canadian Centre for Occupational Health. “(None).” *Canadian Centre for Occupational Health and Safety*, 19 Mar. 2020, www.ccohs.ca/oshanswers/chemicals/how_do.html.

CARB. “Particulate Matter Program,” 2015. Available: https://ww3.arb.ca.gov/pm/pm.htm. [Accessed: 15 December 2019].

CARB. “Fine Particulate Matter Monitoring Program,” 2019. Available: https://ww2.arb.ca.gov/sites/default/files/2019-02/pm25-monitoring-2019.pdf. [Accessed: 11 December 2019].

Census Bureau. “2017 California Wildfires,” 2017. Available: https://www.census.gov/topics/preparedness/events/wildfires/2017-ca-wildfires.html. [Accessed: 10 March 2020].

Dendup, T., Feng, X., Clingin, S., and Astell-Birt, T. 2018. Environmental risk factors for developing type 2 diabetes mellitus: A systematic review. *International Journal of Environmental Research and Public Health;* 15(1).

Diamant, A.L., Babey, S.H., Brown, E.R. and Chawla, N. 2003. Diabetes in California: nearly 1.5 million diagnosed and 2 million more at risk. *Policy brief (UCLA Center for Health Policy Research);* (PB2003-1), pp. 1–8.

Dimakakou, E., Johnston, H.J., Streftaris, G. and Cherrie, J.W. 2018. Exposure to environmental and occupational particulate air pollution as a potential contributor to neurodegeneration and diabetes: A systematic review of epidemiological research. *International Journal of Environmental Research and Public Health;* 15(8).

Division of Diabetes Translation, N.C. for C.D.P. 2017. *National Diabetes Statistics Report, 2017*. Center for Disease Control.

EPA. 2017. Criteria Air Pollutants. Retrieved December 15, 2019, from https://19january2017snapshot.epa.gov/criteria-air-pollutants_.html

EPA. 2019. Policy Assessment for the Review of the National Ambient Air Quality Standards for Particulate Matter, External Review Draft. Retrieved December 19, 2019, from Available:https://www.epa.gov/sites/production/files/2019-09/documents/draft_policy_assessment_for_pm_naaqs_09-05-2019.pdf.

Feng, S., Gao, D., Liao, F., Zhou, F. and Wang, X. 2016. The health effects of ambient PM2.5 and potential mechanisms. *Ecotoxicology and Environmental Safety;* 128, pp. 67–74.

Golden, S.H., Brown, A., Cauley, J.A., Chin, M.H., Gary-Webb, T.L., Kim, C., Sosa, J.A., Sumner, A.E., and Anton, B. 2012. Health disparities in endocrine disorders: biological, clinical, and nonclinical factors--an Endocrine Society scientific statement. *Journal of Clinical Endocrine Metabolism*; 97, pp. E1579-639.

Green, C., Hoppa, R.D., Young, T.K., and Blanchard, J.F. 2003. Geographic analysis of diabetes prevalence in an urban area. *Social Science and Medicine;* 57, pp. 551–560.

He, D., Wu, S., ZHao, H., Qiu, H., Fu, Y., Li, X., and He, Y. Association between particulate matter 2.5 and diabetes mellitus: A meta-analysis of cohort studies. *Journal of diabetes investigation* 8(5), pp. 687–696.

Health Rankings 2018. Explore Diabetes in California - 2018 Annual Report. Available at: https://www.americashealthrankings.org/explore/annual/measure/Diabetes/state/CA. [Accessed: 10 April 2019].

Health Rankings 2019. Explore Diabetes in California - 2019 Annual Report. Available at: https://www.americashealthrankings.org/explore/annual/measure/Diabetes/state/CA. [Accessed: 10 April 2019].

Hime, Neil J., Mark, G., and Cowie, C.. 2018. A Comparison of the Health Effects of Ambient Particulate Matter Air Pollution from Five Emission Sources. *International Journal of Environmental Research and Public Health*; 15(6), pp. 1206.

Huang, Chun Fa et al. 2011. Arsenic and diabetes: current perspectives."*The Kaohsiung Journal of Medical Sciences;* 27(9), pp. 402–410.

Jerrett, M., Burnett, R.T., Beckerman, B.S., Turner, M.C., Krewski, D., Thurston, G., Martin, R.V., van Donkelaar, A., Hughes, E., Shi, Y., Gapstur, S.M., Thun, M.J. and Pope, C.A. 2013. Spatial analysis of air

pollution and mortality in California. *American Journal of Respiratory and Critical Care Medicine;* 188(5), pp. 593–599.

Jung, C., Yu-Ting, L., and Hwang, B. 2015. Ozone, particulate matter, and newly diagnosed Alzheimer's disease: a population-based cohort study in Taiwan. *Journal of Alzheimer's Disease;* 44(2), pp. 573-584.

Juntarawijit, C., and Yuwayong, J. 2018. Association between diabetes and pesticides: a case-control study among Thai farmers. *Environmental Health and Preventive Medicine;* 23(1).

Kelly, F.J. and Fussell, J.C. 2015. Air pollution and public health: emerging hazards and improved understanding of risk. *Environmental Geochemistry and Health;* 37(4), pp. 631–649.

Kouznetsova, I., Chwieralski, C.E., Bälder, R., Hinz, R., Braun, A., Krug, N. and Hoffmann, W. 2007. Induced Trefoil Factor Family 1 Expression by Trans-Differentiating Clara Cells in a Murine Asthma Model. *American Journal of Respiratory Cell and Molecular Biology;* 36(3), pp. 286–95.

Kouznetsova, M., Huang, X., Ma, J., Lessner, L., and Carpenter, D.O. 2007. Increased rate of hospitalization for diabetes and residential proximity of hazardous waste sites. *Environmental Health Perspectives* 115(1), pp. 75–79.

Linou, N., Beagley, J., Huikuri, S. and Renshaw, N. 2018. Air pollution moves up the global health agenda. *BMJ (Clinical Research Edition);* 363.

Miller, L., and Xu, X.. 2018. Ambient PM2.5 Human Health Effects-Findings in China and Research Directions. *MDPI*; *9*(11), 424.

National Emissions Inventory. “2014 National Emissions Inventory (NEI) Data.” 2014. *EPA*, Environmental Protection Agency, 7 Feb. 2020, www.epa.gov/air-emissions-inventories/2014-national-emissions-inventory-nei-data.

Navas-Acien, A., Silbergeld, E.K., Pastor-Barriuso, R. and Guallar, E. 2008. Arsenic exposure and prevalence of type 2 diabetes in US adults. *The Journal of the American Medical Association;* 300(7), pp. 814–822.

Office of Minority Health. 2016. Diabetes and Hispanic Americans [Online]. Available at: https://minorityhealth.hhs.gov/omh/browse.aspx?lvl=4&lvlid=63 [Accessed: 5 May 2019].

Pearson, J.F., Bachireddy, C., Shyamprasad, S., Goldfine, A.B., and Brownstein, J.S. 2010. Association between fine particulate matter and diabetes prevalence in the U.S. *Diabetes Care;* 33, pp. 2196–2201.

Rajagopalan, S., Al-Kindi, S.G. and Brook, R.D. 2018. Air Pollution and Cardiovascular Disease: JACC State-of-the-Art Review. *Journal of the American College of Cardiology;* 72(17), pp. 2054–2070.

Rivera-González L.O., Zhang Z., Sánchez B.N., Zhang K., Brown D.G. andRojas-Bracho L. 2015. An assessment of air pollutant exposure methods in Mexico City, Mexico. Journal of Air and Waste Management Association 65(5), pp. 581–591.

Rodriquez, M. and Zeise, L. 2017. *CalEnviroScreen 3.0*. OEHHA.

Rücker, l R., Ibald-Mulli, A., Koenig, W., Schneider, A., Woelke, G., Cyrys, J., Heinrich, .J, Marder, V., Frampton, M., Wichmann, H.E. and Peters, A. 2005. Air pollution and markers of inflammation and coagulation in patients with coronary heart disease. American Journal of Respiratory and Critical Care

Medicine 173(4), pp. 432–441.

Saldana, T.M., Basso, O., Hoppin, J.A., Baird, D.D., Knott, C., Blair, A., Alavanja, M.C.R. and Sandler, D.P. 2007. Pesticide exposure and self-reported gestational diabetes mellitus in the Agricultural Health Study. *Diabetes Care;* 30(3), pp. 529–534.

Sergeev, A.V. and Carpenter, D.O. 2005. Hospitalization Rates for Coronary Heart Disease in Relation to Residence Near Areas Contaminated with Persistent Organic Pollutants and Other Pollutants. *Environmental Health Perspectives;* 113(6), pp. 756–61.

Steinmaus, C., Yuan, Y., Liaw, J. and Smith, A.H. 2009. Low-level population exposure to inorganic arsenic in the United States and diabetes mellitus: a reanalysis. *Epidemiology;* 20(6), pp. 807–815.

Tran HT, Garcia C, Motallebi Z, Miyasato L and Vance W. 2008. Methodology for Estimating Premature Deaths Associated with Long-term Exposure to Fine Airborne Particulate Matter in California. California Environmental Protection Agency: Air Resources Board, pp. 1–48.

Valari, M., Martinelli, L., & Chatignoux, E. (n.d.). Time scale effects in acute association between air-pollution and mortality. Retrieved March 12, 2020, from https://www.lmd.polytechnique.fr/~mvalari/Files/VALARIGRL11.pdf

Vogelsang, T. 2012. Heteroskedasticity, autocorrelation, and spatial correlation robust inference in linear panel models with fixed-effects. *Journal of Econometrics;* 166(2).

Vrijheid, M. 2000. Health Effects of Residence Near Hazardous Waste Landfill Sites: a Review of Epidemiologic Literature.*Environmental Health Perspectives;* 108 Supplement 1, pp. 101–12.

World Health Organization. 2016. Global Report on Diabetes. Retrieved December 10, 2019, from https://www.who.int/diabetes/global-report/en/

Wu Y.H. and Hung M.C. 2016. Comparison of spatial interpolation techniques using visualization and quantitative assessment. InTech.

Xing, Y.-F., Xu, Y.-H., Shi, M.-H., & Lian, Y.-X. (2016). The impact of PM2.5 on the human respiratory system. Journal of thoracic disease, 8(1), E69-74.

Yang, Bo-Yi et al. 2019. Ambient air pollution and diabetes: A systematic review and meta-analysis. *Environmental Research* 180: 108817.