Causal Impacts of Climate Change on Human Health

A Technical Report submitted to the Department of Biomedical Engineering

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Zayyad Siddiqui

Spring, 2021

Technical Project Team Members

Annisa Elbedour

Prachi Yadav

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Timothy Allen, Department of Biomedical Engineering

Shannon Barker, Department of Biomedical Engineering

Causal Impacts of Climate Change on Human Health

Annisa Elbedour[a,1]. Zayyad Siddiqui[b,2], Prachi Yadav[c,3], Pam Bhattacharya[d]

[a] Fourth Year Biomedical Engineering Undergraduate, University of Virginia
[b] Fourth Year Biomedical Engineering Undergraduate, University of Virginia
[c] Fourth Year Biomedical Engineering Undergraduate, University of Virginia
[d] Lead Data Scientist, MITRE Corporation
[1] ae3es@virginia.edu
[2] zs5fe@virginia.edu
[3] psy9yd@virginia.edu

Abstract

Climate change is an accelerating issue, bringing the global environment closer than ever before to irreversible damage due to anthropogenic actions. Air pollution is just one contributor to a contaminating environment, and this project aims to shed light on air quality's impact on human health through computational modeling in R. Air quality data from the EPA consisting of pollutants $PM_{2.5}$, PM_{10}, SO_2, NO_2, CO, and ozone were acquired and cleaned. Respiratory-related mortality time-series data was also acquired from the CDC. Temperature data was also included as a confounding variable. The data was derived from six geographic locations in the DC-Maryland-Virginia area and spanned over 19 years.

Data transformation techniques were conducted on the time-series data for input into the causal model code. Granger Causality and Bayesian network principles were the basis for the two models created. Results demonstrated temperature to be a major causal contributor to mortality in both general population models. For race, age, and Chronic Obstructive Pulmonary Disease (COPD) stratification, racial minorities, older age groups, and those with underlying respiratory conditions were disproportionately affected by negative changes in air quality. The two models were validated through a variety of statistical tests and forecasting accuracy metrics. A pseudo-data set with defined causal links was also created in order to evaluate how well the Granger and Bayesian algorithms we used, could learn known causal relationships from the data.

These models, while non-exhaustive, provide a small contribution to elucidating the long-term effects of climate change on public health outcomes. Further research in causal modeling by the scientific community is needed to help verify the severity of air pollution and ultimately lead to more concrete adaptation and mitigation efforts by policymakers.

Keywords: climate change, air pollution, causality, Granger, Bayesian

Introduction

Climate change is deemed the largest global health threat of the 21st century. Climate change is defined as the long-term alteration of temperature and typical weather patterns in a given place and includes events such as average increases in the frequency of natural disasters, global temperatures, and exposure to vector-, food-, and water-borne infectious diseases. Between 2030 and 2050, climate change is expected to cause approximately 250,000 additional deaths per year, from malnutrition, malaria, diarrhea and heat stress[1]. Impacts of climate change extend well beyond increases in global temperature; things that we depend upon and value — water, energy, transportation, wildlife, agriculture, ecosystems, and human health — are experiencing the adverse effects of a changing climate.

Air pollution is one of many drivers of climate change, and negatively impacted air quality is also an exposure pathway affected by upstream climate change drivers. Changes in the quality of air, water, and food affect our health through multiple pathways, increasing respiratory and cardiovascular disease, injuries and premature deaths, as well as threats to mental health, just to name a few. When specifically referring to air quality, the annual U.S. average levels of fine particulate matter (PM 2.5) was on the decline by 24% between 2009 to 2016, and then experienced a 5% increase between 2016 and 2018 which can be attributed to the rollbacks of the Clean Air Act. Air pollution causes an estimated 7 million deaths worldwide every year and has been on the rise[2].

Due to the complex processes involved in climate change, the exact extent to which human health outcomes may be influenced by the changing climate remains unclear. In fact, several different interpretations of the severity of climate change exist. There is currently a strong divide in the US population with polarized beliefs in regards to the negative health impacts that climate change causes. As of 2020, the bipartisanship experienced the greatest divide in whether or not climate change should be a top priority, with 85% of Democrats and only 39% of Republicans stating yes, meaning a 46% disparity[2].

Recent studies show strong temporal correlations between climate change and societal public health crises. However, specific causal mechanisms underlying this relation have not been addressed and are often swept aside due to the presumed impracticality of establishing causality.

Our approach aimed to bridge the gap in knowledge regarding climate-induced health outcomes previously known to have a weak causal relationship. Although such literature has implied correlations between public health outcomes and climate change, the two have not been combined in a comprehensive *causal* model thus far. Separate climate change and health outcome trends have been analyzed, but few causal models have merged the two together. In order to further decipher the impacts of climate change on human health, we have developed a causal model between a climate change driver (and also exposure pathway) and a public health metric. As a long-term goal, this model can help support the prioritization of climate change preventative programs and adaptive resilience plans. It can also help estimate associated costs and benefits in mitigating human impacts attributed to climate change. Understanding the implications of climate change now will help foster the development of proactive and responsive policies and predictive models to reduce future risks and impacts of climate change.

The established aims for this project are as follows:

Aim 1**: To collect, process and integrate historical climate and public health time series data**

For historical climate time-series data on this topic, we acquired temperature and air pollution data gathered from sources such as the CDC, the Census, and the EPA's Environmental Justice Screening and Mapping Tool (EJSCREEN), with a focus on monthly time series data. The health outcome we chose was respiratory-related mortality, and time-series data was gathered for the same time period and at equivalent levels (monthly) as the climate time-series data in order to compare the two variables. Literature served as a basis for choosing the specific air quality and mortality variables. The climate and health outcome data was refined to only include relevant and useful time points, integrated from different sources, and combined into standardized headers using data pre-processing techniques in R. The final climate change driver chosen was air pollution, and data for 6 pollutants was incorporated in the models. Data was stratified also by geographic location (with an emphasis on the DC-Maryland-Virginia area), race, age, and COPD. The final health metric incorporated in the causal models was respiratory-related mortality.

Aim 2: **To develop a causal model of a specific pathway between a climate change factor and public health metric**

After leveraging information in temporal data to assist with causal linking between a factor of climate change and a health outcome metric, a series of models were developed utilizing computational modeling libraries in R. This included various Granger Causality models and Dynamic Bayesian Networks in order to compare and contrast results.

Aim 3: **To validate the causal model against data using time-series cross-validation**

The model was evaluated using the train-test split method by partitioning the time-series data into chronological subsets with the earlier observations (20-80% of the data) reserved for the training set and later observations (the final 80-20%) reserved for the testing set. To assess the model's accuracy, we determined how well the model utilizes the training set to predict the testing set through statistical errors such as mean absolute error[3]. Another form of validation included software validation, in which a "test" data set with predefined causal relationships, was created in MATLAB that was similar in characteristics to the climate/health data, and was used to evaluate whether Granger and Bayesian algorithms were capable of detecting the causal relationships embedded in the dataset.

Assumptions

In the beginning of the project, there were several hypotheses and assumptions that had to be made. A major assumption that was made during the data acquisition process was that the air pollution sensors in various counties worked precisely and thus the data was a reflection of the true pollution values. Another assumption that had to be made when referring to the respiratory-related mortality data was that the reportings of the crude death rate were accurate and that any missing data points would not drastically alter any trends and model outputs. As for hypotheses, links between all pollutants and respiratory-related mortalities were expected, since the six pollutants acquired have been indicated to cause health issues in previous literature.

Methods

Data Acquisition

Historical climate time series data was gathered from the EPA's Environmental Justice Screening and Mapping Tool, with a focus on monthly data from January 1999 to December 2018. Data for 6 different air pollutants was acquired, including $PM_{2.5}$ and PM_{10} in units of ug/m^3, and CO, ozone, SO_2, and NO_2 in units of parts per million (ppm). Monthly temperature averages (Fahrenheit) were also acquired from the EPA as temperature acts as a confounding variable when detecting mortality trends. General respiratory-related mortality time-series data which fall under the J00-J99 ICD-10 codes was then gathered for the same time period and at the equivalent monthly level in order to align temporal and geographical scales. Mortality was then stratified by race, age, and sub-categories of disease using the CDC's Wondertool. More specifically, race was broken down into white and Black populations, age groups into 65-74, 75-84, and 85+ populations, and disease into the COPD population. Limited data was available for other races (Latinx, Native American and Asian) as well as for younger age groups <65 years of age. Units were given in terms of "Crude Death Rate" which is the total number of deaths to residents in a specific geographic region per hundred thousand residents.

The data was gathered for 6 different counties within the D.C.-Virginia-Maryland region based on data availability: Washington D.C., Fairfax County (VA), Richmond City (VA), Montgomery County (MD), Prince George's County (MD), and Baltimore County (MD). Counties with very few time points were excluded, as well as those with very large gaps in data points in which interpolation would yield inaccurate estimations. Counties such as Loudon and Arlington county were omitted due to the lack of monthly data for the given time frame. The CDC's Social Vulnerability Indices (or SVIs) of each county were determined in order to identify demographic groups and geographic locations that are more vulnerable to external stresses on human health, including air pollution and disease. The higher the SVI (on a scale of 0-1), the more vulnerable an area is to environmental and public health hazards.

Data Cleaning

The data was integrated from the different sources and combined into standardized headers for data pre-processing. If daily data was available for certain counties, the monthly average was taken using Excel. The climate and health outcome data were then refined to only include relevant time points in R. Several counties required interpolation and extrapolation due to sparse data. Using the *dplyr* package in R, interpolation had to be used to fill in gaps in the time-series data in order to have consistent monthly time intervals for each year. Some datasets had to be extrapolated using Excel's curve fitting tool in order to have set start and end time points for the data. The optimal extrapolated values were determined based on R^2 values from either linear or polynomial regression. What was noticed as we began gathering and cleaning data was that data for non-white populations typically required much more interpolation and extrapolation than data for white population. For example, Prince George's County's white population only makes up about 17% of the total population, but the Black population still required more interpolation[4]. This lack of data may stem from the fact that some counties and communities are more socially vulnerable and may not have the resources for consistent and accurate data collection, indicating a need to enforce better and more consistent data collection and management across communities, at a policy level.

In addition to those two primary forms of data cleaning, we also conducted time-series decomposition using the *stats* package in R in which the monthly pollutant and mortality data was broken down into seasonality, trend, and random components. Seasonality would be defined as repeating short-term cycles found in the data, and the trend component refers to the overall increasing or decreasing pattern in data. The random residual output was used for Granger causality modeling in order to remove seasonality patterns and confounding trends when determining causal relationships, and also address requirements for data to be stationary. The trend component was calculated based on a 12-day moving average; therefore, 6 points were omitted at the beginning and end of the data set.

Model Formation

Two types of causal models, including the Granger-Causal model and the Bayesian Network model were used in model formation and then compared and contrasted. The Granger-Causal model is a prediction-based statistical concept of causality in which empirical data is used to find patterns of correlation. Specifically, it states that if variable X causes variable Y, then past values of X should help predict Y beyond just the previous values of Y alone. Granger causality is also known as a "bottom-up" procedure in which the assumption is that the two time-series variables are independent, and then the data sets are analyzed for correlation. This is different from "top-down" procedures which would be assuming a link between the time series variables, and then analyzed to see if they are generated independently from each other.

The *vars* package in R allows one to develop a Granger causal model using vector autoregression, a statistical model used to capture the relationship between all the pollutants and respiratory-related mortalities as they change over time. Time-decomposed variables for pollutants, temperature, and mortality for each county were imported as time-series objects. Then, the VAR() function created multivariate time-series equations where the endogenous variables in the system are functions of the lagged values of all endogenous variables. A lag order of 1, indicating a lag of one month, was determined using the VARselect() function which outputted the optimal lag based on different information criterion (IC). The Akaike Information Criterion (AIC) tests how well our model fits the data set without overfitting it, and therefore was used in determining the IC score.

After a VAR equation was obtained for each potentially causal variable (pollutants and temperature), the equation and its associated parameters were inputted into the causality() function. The causality function outputted p-values that represented the probability of statistical significance of Granger causality for each pollutant. After having acquired the p-values for each individual pollutant vs mortality, pollutants which showed significant causality were then paired and grouped together to investigate the potential relationships between multiple pollutants and their combined effect on mortality. Tables were then generated to organize the p-values for each model for later downstream analysis.

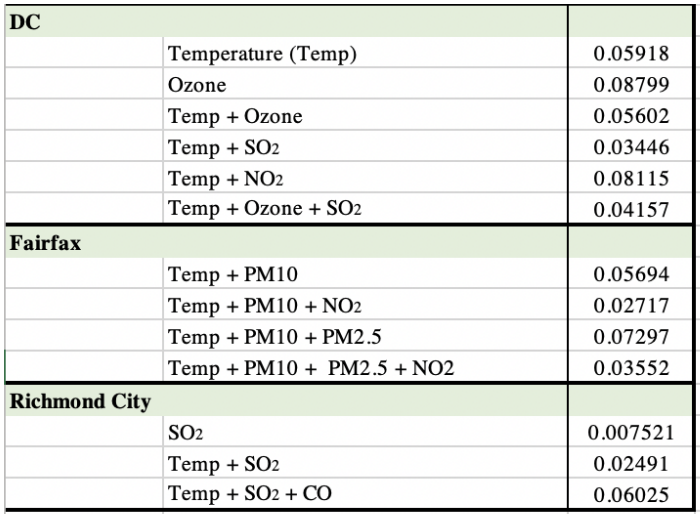
The second type of causal model was the Dynamic Bayesian Network model. A Bayesian network is defined as a directed acyclic graph that utilizes conditional probabilities to reflect causal links between parent and child nodes. The parent nodes are defined as the original input variables (i.e. pollutants/temperature) whereas the child nodes are nodes extending from other nodes and therefore are dependent. Furthermore, one approach to learning Bayesian networks is using score-based structure learning algorithms and more specifically the hill climbing algorithm. Hill climbing is an optimization technique that starts with an arbitrary solution to a problem and then makes incremental changes to the solution to find a better solution. The algorithm continues to make incremental changes until no further improvements can be made. Because we are using time series data representing a dynamic system, we learned a Dynamic Bayesian Network using the dynamic max-min hill climbing (dmmhc) algorithm. At their core, these algorithms are based on Bayes Theorem and conditional probabilities between pollutants, temperature, and mortality, in order to measure how likely variable X caused variable Y or the probability that evidence of Y occurs given that knowledge about X already exists (P (Y | X)).

To formulate a Dynamic Bayesian network, raw (non-time-decomposed) data from each county was imported into R using the *bnlearn, bnstruct, dbnR,* and *Rgraphviz* packages. The node sizes were set to five meaning five time-series points were compared. A node size of five was chosen because it balanced factors such as computational time, feasibility, and model accuracy for each county. Furthermore, certain arcs such as those protruding from mortality had to be blacklisted to ensure meaningful causal relationships were maintained.

The original data set was set to 80% training and 20% test but was altered afterwards for cross-validation testing. The 80% training set was subsequently inputted into the learn_dbn_struc() function and plotted to formulate a visual network for each county. Accessory functions such as fit_dbn_params() were then used to calculate the the regression coefficients linking parent and child nodes in the network. The equation with mortality as the dependent variable was used to generate conditional density tables to capture the strengths of each connection between pollutants and temperature.

Results

Granger General Population



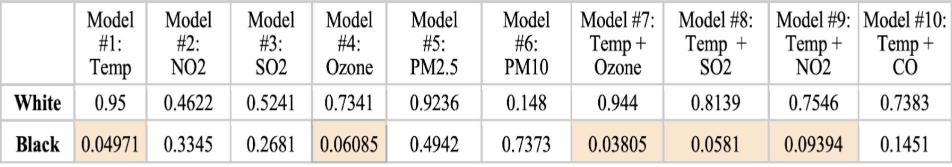
DC		
	Temperature (Temp)	0.05918
	Ozone	0.08799
	Temp + Ozone	0.05602
	Temp + SO_2	0.03446
	Temp + NO_2	0.08115
	Temp + Ozone + SO_2	0.04157
Fairfax		
	Temp + PM10	0.05694
	Temp + PM10 + NO_2	0.02717
	Temp + PM10 + PM2.5	0.07297
	Temp + PM10 + PM2.5 + NO2	0.03552
Richmond City		
	SO_2	0.007521
	Temp + SO_2	0.02491
	Temp + SO_2 + CO	0.06025

Table 1: Granger Causal Model P-Value Table for General Population. The causality function in R generated p values for each variable combination, describing the relationship they each have with respiratory-related mortality. Those which are less than 0.1 are displayed, indicating a potentially significant Granger causal link between the pollutants and respiratory-related mortality. P values were often found to be smaller when paired with temperature.

The VAR() function in R was used in conjunction with the causality() function to generate several models using the residual data combining various pollutants and temperature with respiratory-related mortality for each county. However, only DC, Fairfax County, and Richmond City yielded models with p-values under 0.1, our chosen critical value (See Table 1). Prior to model generation, it was hypothesized that temperature would act as a proxy for a seasonality component and would strengthen the causal links by decreasing the p-values. This hypothesis was taken from many pieces of literature linking seasonality-mortality relationships, such as a case study in subtropical China in which morbidity burden increased with exposure to extreme temperatures[5]. This hypothesis was supported by the results in Table 1, as eleven of the thirteen significant models included temperature. Significant causal pollutants included PM_{10} in Fairfax, SO_2 in Richmond City, and ozone in DC. NO_2-significance seemed to be evenly dispersed between the geographic regions, which makes sense because NO_2 is one of several gas pollutants produced by road traffic and fossil fuel combustion processes, which is present in all regions. A limitation for Granger Causal models is that although p-values are generated, it is less straightforward to visualize relationships intuitively with a network structure. This is a shortcoming that can be mediated by the Bayesian Network model.

Granger Stratification by Race

As for stratification by race, when inputted into the causality function, it was seen that a greater amount of p values were significant for the Black population, indicating more potentially causal links between air particulate matter and general respiratory-related mortality (see Table 2). For D.C., five causal variables and their pairwise groupings were seen to be significant in the Black population, and none were significant for the white population. This racial disparity may be due to the effects environmental racism has on the health and well-being of marginalized communities, where environmental racism is defined as the environmental decisions, actions, and policies that disproportionately affect minorities[6].



	Model #1: Temp	Model #2: NO2	Model #3: SO2	Model #4: Ozone	Model #5: PM2.5	Model #6: PM10	Model #7: Temp + Ozone	Model #8: Temp + SO2	Model #9: Temp + NO2	Model #10: Temp + CO
White	0.95	0.4622	0.5241	0.7341	0.9236	0.148	0.944	0.8139	0.7546	0.7383
Black	0.04971	0.3345	0.2681	0.06085	0.4942	0.7373	0.03805	0.0581	0.09394	0.1451

Table 2: P-value Table for Racially-Stratified Models. Granger Causal Models were generated for both the Black and white population in Washington, D.C. for multiple combinations of pollutants and mortality. The respective p-values are shown, with those less than 0.1 highlighted in orange. Based on statistical hypothesis testing, each model suggested whether there was a potential causal relationship between one or more air pollutant variables, and the health outcome of interest (mortality). The air pollutants that were considered for each model, are stated in the column titles.

Granger Stratification by Age

After completing the Granger causal models for the general population, the same air pollutant and respiratory-related mortality data was stratified by age for each geographic location. Certain regions, such as Fairfax County, DC, Montgomery County, and Prince George's County had available data for both 75-84 and 85+ age groups, while Richmond City had very sparse data that did not meet the chosen interpolation and extrapolation limit (at least half the time points). Baltimore County, however, had available data for three age groups, including 65-74, 75-84, and 85+ populations. The results for the Granger models stratified by age indicated that if one model was significant within a location for a given age group, the subsequent combinations of pollutants and temperature with mortality were found to be significant as well. This was shown with the DC 75-84 age group as well as the Montgomery 75-84 age group (See Table 3). The same level of consistency was not seen for the general population Granger models, and one reason for this may be because individuals in a certain age group are more likely to have similar backgrounds and medical history, and thus any existing causal links would be strengthened.

a)

DC Age 75-84	Model #1: TEMP	Model #2: TEMP + NO2	Model #3: TEMP+ SO2	Model #4: TEMP+ OZONE	Model #5: TEMP+ PM2.5	Model #6: TEMP+ PM10	Model #7 TEMP + CO	Model #8 TEMP +CO +PM10	Model #9 TEMP+CO+PM10+ OZONE+ NO2
p-Value	0.03216	0.09505	0.09959	0.09461	0.09886	0.07152	0.01022	0.02325	0.08755

b)

Richmond City COPD	p-Value
Model #1: Temp + SO2	0.02499
Model #2: NO2	0.0411

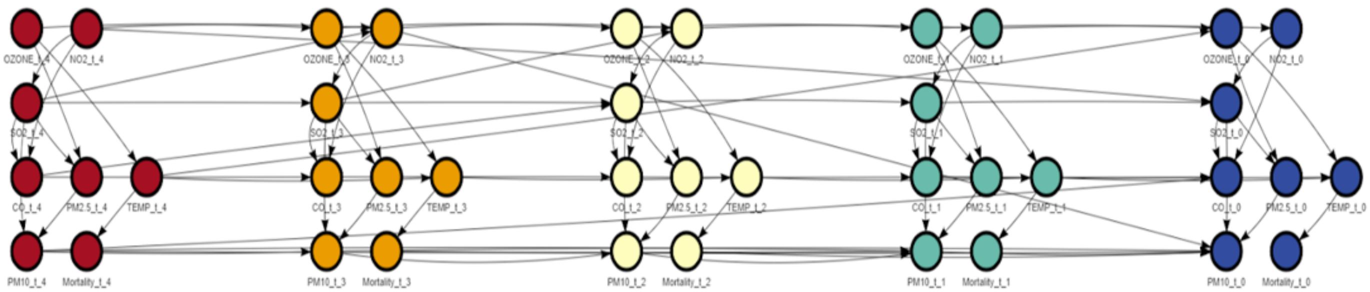
Table 3: Granger Causal Model P-value Table Stratified by Age and COPD: a) All attempted models for the 75-84 age population resulted in low p-values, with the ones below .05 indicated in dark red and the ones below 0.1 indicated in light red. **b)** For COPD stratification, only two models in Richmond City appear to show significant causality to mortality with a p-value <0.1. In model 1, temperature and SO_2 are linked to mortality while in model 2, NO_2 was linked to mortality which is feasible considering that NO_2 and SO_2 have been known to lead to characteristic airway remodeling and changes in mucus secretion. Therefore, NO_2 and SO_2 are major risk factors for COPD.

Granger Stratification by COPD

Granger models were also stratified by COPD, which is caused by long-term exposure to irritants that can damage lungs and airways. Therefore, we thought it would be beneficial to evaluate if there were similar causal links found in the general population. Given the high p-values, there were no significant causal relationships present in Washington DC, Fairfax, Montgomery, Prince George's, or Baltimore County. However, in Richmond City County, two models showed significant p-values under α= 0.1 One of those models demonstrated that temperature and SO_2 may be causally linked to COPD-related mortalities with a p-value of 0.02499. The other model demonstrated that NO_2 may be causally linked to COPD-related mortality with a p-value of 0.0411. This disparity is surprising as Richmond City is seen as one of the nation's least congested urban areas. Given Richmond City's sparse population and that the average commuter in Greater Richmond spends only 35 hours in traffic annually compared to the national average of 54 hours per year, it was originally hypothesized that the county would not have SO_2 or NO_2, pollutants mainly associated with traffic, closely associated with mortality[7]. However, based on the Bayesian results, this was not the case.

Bayesian General Population

For the general population, Bayesian network models were constructed for each of the 6 counties. As stated before, the node size was set to five, which representations relationships spanning 5 months, and blacklisting was utilized for arcs extending from mortality to ensure meaningful causal relationships were being maintained. Furthermore, the 80-20 train-test set was utilized for all general population models. For all counties, temperature had causal links present with mortality. In addition, all models had a slightly negative coefficient indicating that as temperature increases, mortality decreases. An example Bayesian network and coefficient table for DC's general population is provided in Figure 1. The negative relationship between temperature and mortality may actually be a false-positive association that occurs given the seasonal component present in the non-time-decomposed input data. Due to computational limitations and the *dbnR* package's sensitivity, the Bayesian models may not have been able to detect all causal links with mortality and therefore this may be another reason for temperature's negative relationship with mortality. In addition, for Montgomery, Baltimore, and Prince George's County, ozone also had causal links present with mortality. In the mortality equation, the weighted coefficients for ozone were slightly positive indicating that as ozone levels increase, the mortality rate increases.



DC	Mortality = Intercept + Coeff *(Temperature)	
Time Series	**Intercept**	**Coefficient**
0	31.6236	-0.11558
1	31.6557888	-0.1155261
2	32.1188493	-0.1217935
3	34.2179818	-0.127755
4	33.042456	-0.133801

Figure 1: General DC Population Bayesian Modeling. A Bayesian network model of D.C. was generated using the *dbnR* package, and the node size was designated to be five, which meant five time series points were compared. As seen across all networks, there seems to be a strong relationship between temperature and mortality. There are also many links between the pollutants themselves, such as ozone to NO_2, which, rather than indicating direct causal relationships, may be due to common sources like automobile emissions.

This aligns with the current literature that states that elevated concentrations of ozone are associated with an increased number of deaths from respiratory causes. Furthermore, the biological mechanisms behind this relationship is largely unknown but preliminary research is currently investigating the effects of ozone on the autonomic control of the cardiovascular system, on coagulation mechanisms, and on vasoactive substances in the blood[8].

In Richmond City, NO_2, alongside with temperature, also had causal links present with mortality. The weighted coefficients for NO_2 were slightly positive indicating that as NO_2 levels rise, the mortality rate increases. The reason may be due to the fact that NO_2 is a highly reactive, poorly water-soluble gas that deposits peripherally in the lungs. A major target site for the action of NO_2 is the terminal bronchioles where NO_2 has been linked to inducing airway inflammation, worsening coughing and wheezing, increasing asthma attacks, and reducing overall lung function[9].

Bayesian Stratification by Race

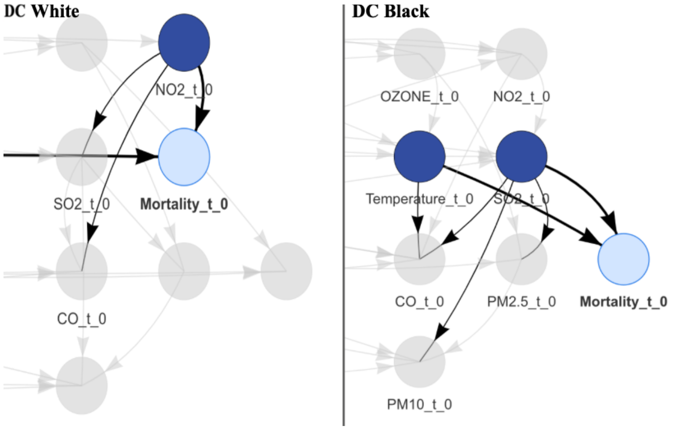


Figure 2: Bayesian Models for Racially-Stratified Models. Bayesian Models were created for both races. The nodes and connections at t 0 are highlighted, signifying that SO_2 and temperature are causally linked to mortality for the Black population, and NO_2 is causally linked to mortality for the white population.

Atmospheric ozone has two effects on the temperature balance of the Earth. It absorbs solar ultraviolet radiation, which heats the stratosphere[10]. It also absorbs infrared radiation emitted by the Earth's surface, effectively trapping heat in the atmosphere, and therefore it makes sense for ozone to be linked to temperature in the Bayesian model (see Figure 2). Based on Figure 2, temperature has a potentially causal relationship with mortality only in the Black population. This may be due to historical redlining and residential segregation. Redlining is the discriminatory practice that puts financial services, including housing mortgages, out of reach for racial and ethnic minorities. In fact, formerly redlined neighborhoods are, on average, five degrees warmer than whiter, more affluent neighborhoods where redlining never occurred[11].

SO_2 and NO_2 are both linked to mortality in the Black and white populations respectively. Both of these pollutants are emitted by the burning of fossil fuels — coal, oil, and diesel — from power plants and automobiles[12]. Taking a look at D.C., it is unsurprising to see these two pollutants act as potentially major causes of respiratory-related mortality as D.C. has 51 major roadways, each a major avenue that serve as the city's principal traffic arteries, with more than 3.8 million vehicles registered in the District[13]. This number does not take into account the number of commuters from neighboring states. Most other counties located near major cities and urban centers (Baltimore, Richmond City) also have NO_2 and SO_2 linkages to mortality. The EPA recognizes that NO_2 and SO_2 are inextricably linked from both atmospheric chemistry and environmental effects perspectives as deduced by their typical linkage in several Bayesian models[14]. However, the discrepancy between races may be due to the fact that each race may generally live nearby different industrial facilities within the D.C. metropolitan area due to differential land zoning[11]. Although, it is important to note that the coefficients linking mortality to NO_2 for the white population are lower than that of the links to SO_2 for the Black population, indicating a potentially stronger causal relationship for the Black population.

It was also observed that only Black populations, particularly in Baltimore and Prince George's counties, had Bayesian models in which $PM_{2.5}$ or PM_{10} had direct nodal connections to mortality. Disparities in residential proximity to pollution sources may account for this difference[6].

Bayesian Stratification by Age

When examining data stratification by age, it was seen that older age populations (85+) generally had more connections linking air pollutants to mortality. It was also seen that older ages (75+) typically had temperature linked to mortality, as seen with Baltimore County in Figure 3a. This aligns with the current literature which states that older adults are more susceptible to air pollutants compared to younger adults due to decreased physiological, metabolic and compensatory processes as well as a greater incidence of cardiovascular and respiratory disease[1].

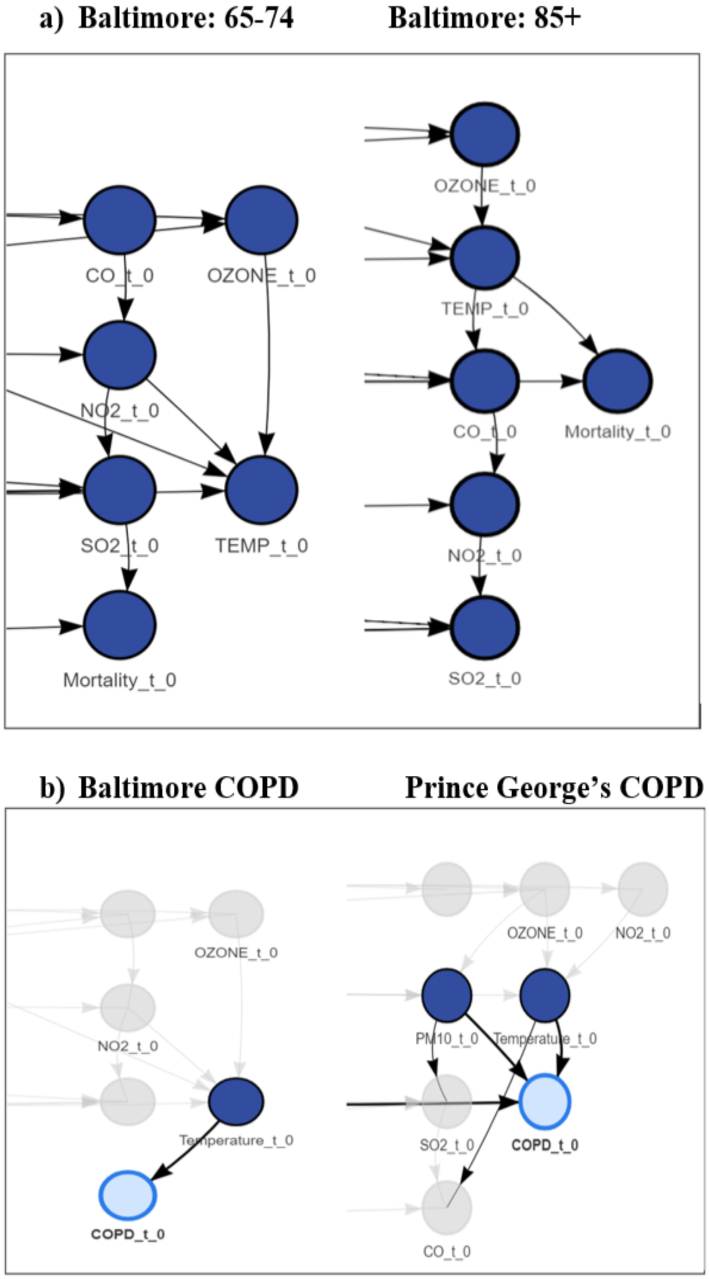


Figure 3: Bayesian Models for Age-Related and COPD-Related Mortality Models. Bayesian models were stratified by a) age groups and b) Chronic Obstructive Pulmonary Disease related mortality. Younger populations were less susceptible to pollutants. PM10 and temperature were linked to mortality for the COPD populations.

Bayesian Stratification by COPD

For COPD stratification, all county models except Baltimore had autocorrelation present between past and future COPD-related mortality time points. All county models except DC and Richmond City demonstrated casual links present between temperature and COPD-related mortalities, which was similar to the general population trends (see Figure 3b). In the COPD equation for all counties, the weight of the temperature coefficient was slightly negative showing that there may be false-positive artifacts present in the models and that additional data transformation techniques may be required.

In particular, Prince George's County also had a causal relationship between PM_{10} and COPD-related mortalities which was unlike the general population trends. In the COPD equation, there was also a slightly negative coefficient for PM_{10} which may again be a false-positive artifact. However, this finding does not align with the current literature where a study found that for every 10 ug/m^3 increase in PM_{10}, there was 1.1% increase in COPD-related mortalities[17]. This may be due to the fact that air pollutants like PM_{10} are seen as risk factors for COPD and have the potential to cause inflammation in the lungs. PM_{10} is also known to lead to destruction of cells and tissue through oxidative stress where there is an imbalance in the body's ability to neutralize certain molecules such as free radicals and reactive species. If the body's antioxidants are depleted, then this can result in impaired cellular function thereby contributing to diseases such as COPD[15].

Social Vulnerability Indices

Generally, counties with higher SVIs tended to have higher Black populations and tended to require greater data interpolation. Fairfax county is the third most affluent county in the country, with the lowest SVI of the 6 we examined. However, Fairfax did not have any available data for its Black population[16]. The two counties with highest SVIs, Prince Georges and Richmond City, required the most interpolation and extrapolation during data cleaning. The Black population in these counties also make up the majority, with Prince George's being 62.67% Black and Richmond City being 49% Black[4]. D.C. also has a Black majority, making up 46.31% of the population compared to 41% of white people[18].

When modeling Granger causality, it was evident that counties with higher SVIs had many more significant p-values outputted from the models compared to counties with lower SVIs. In fact, Montgomery and Baltimore counties had no significant p-values of the pairwise groupings tested. It was also observed that the Bayesian models for counties with the three lowest SVIs only had temperature and/or ozone linked to mortality, whereas the other counties with higher SVIs had NO_2, SO_2, CO, $PM_{2.5}$ and/or PM_{10} linked to mortality as well.

Data and Model Validation

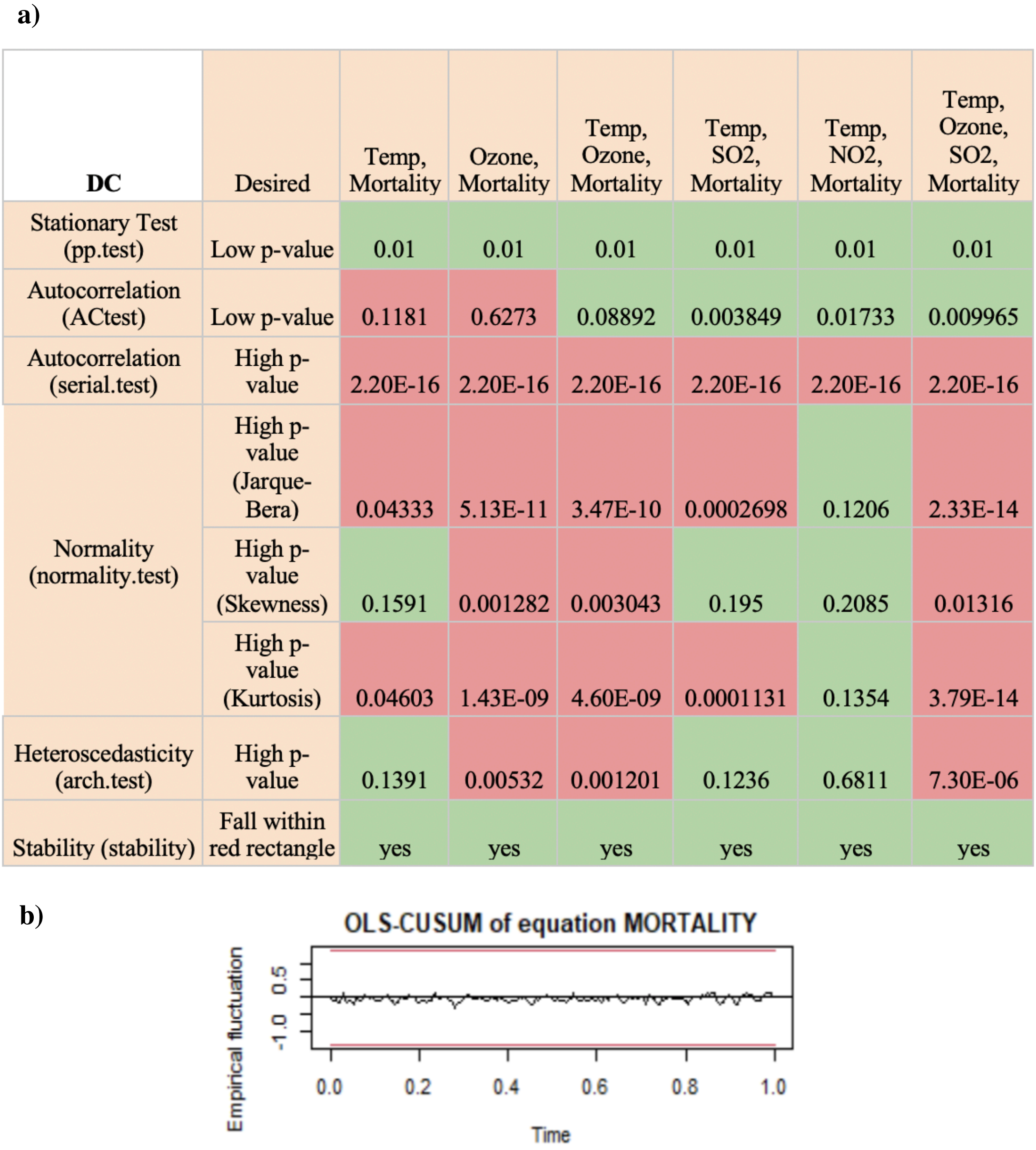
Granger Statistical Testing

To conduct validation as part of Aim 3, the group utilized the *vars, tseries,* and *forecast* packages in R in order to conduct diagnostic statistical testing on models that had already demonstrated Granger causality in Aim 2 (p-value < 0.1). The stationary test was first utilized prior to inputting the data into the model to determine if the statistical properties of the time-series data do not change over time. For all counties, it was found that the general population models passed the stationarity test and therefore the data was deemed stationary. An example Granger validation table for the DC general population is shown in Figure 4a.

The next test, autocorrelation, was utilized to determine if there was a similarity present between observations as a function of the time lag between them. Across all counties' general populations, approximately half of the models from each county passed this metric with a p-value < 0.1. Furthermore, a normality test was conducted to determine if the residuals of the data follow a normal distribution. Multiple statistical tests such as Jarque-Bera, Skewness, and Kurtosis, were utilized to measure normality. Across all counties' general populations, again, only a few models passed at least one of these tests.

Another metric to consider is heteroscedasticity or the volatility of the changing variance in data. In time-series data, it is desirable to have constant variance, but unfortunately, across all counties' general populations, only a handful of models demonstrated non-heteroscedastic behavior. The last metric, stability, tests for the presence of structural breaks which is important considering that if structural breaks go undetected in modeling, then the accuracy of the model degrades. In order to fail the stability tests, the data would need to fall outside of the red boundaries in the generated plots. An example plot of the stability test for Fairfax County is shown in Figure 4b. For all counties' general population, all models passed this metric demonstrating stable data. For Granger stratification by race, models for both the Blacks and whites in all counties exhibited similar trends to those of the general population models.

One should note, however, that the Black population models tended to fail more heteroskedasticity tests compared to white population models, which may arise from the extra interpolation that was needed for the Black population thereby increasing its volatility. For Granger stratification by age and COPD, models for all counties demonstrated similar trends to those of the general population.


a)
b)
OLS-CUSUM of equation MORTALITY
Empirical fluctuation
-1.0
0.5
0.0
0.2
0.4
0.6
0.8
1.0
Time

DC	Desired	Temp, Mortality	Ozone, Mortality	Temp, Ozone, Mortality	Temp, SO2, Mortality	Temp, NO2, Mortality	Temp, Ozone, SO2, Mortality
Stationary Test (pp.test)	Low p-value	🟩 0.01	🟩 0.01	🟩 0.01	🟩 0.01	🟩 0.01	🟩 0.01
Autocorrelation (ACtest)	Low p-value	🟥 0.1181	🟥 0.6273	🟩 0.08892	🟩 0.003849	🟩 0.01733	🟩 0.009965
Autocorrelation (serial.test)	High p-value	🟥 2.20E-16	🟥 2.20E-16	🟥 2.20E-16	🟥 2.20E-16	🟥 2.20E-16	🟥 2.20E-16
Normality (normality.test)	High p-value (Jarque-Bera)	🟥 0.04333	🟥 5.13E-11	🟥 3.47E-10	🟥 0.0002698	🟩 0.1206	🟥 2.33E-14
	High p-value (Skewness)	🟩 0.1591	🟥 0.001282	🟥 0.003043	🟩 0.195	🟩 0.2085	🟥 0.01316
	High p-value (Kurtosis)	🟥 0.04603	🟥 1.43E-09	🟥 4.60E-09	🟥 0.0001131	🟩 0.1354	🟥 3.79E-14
Heteroscedasticity (arch.test)	High p-value	🟩 0.1391	🟥 0.00532	🟥 0.001201	🟩 0.1236	🟩 0.6811	🟥 7.30E-06
Stability (stability)	Fall within red rectangle	🟩 yes	🟩 yes	🟩 yes	🟩 yes	🟩 yes	🟩 yes

Figure 4: Diagnostic Testing for DC Granger Model. a) The Granger model output for each significant relationship was tested for stationarity, autocorrelation, normality, heteroscedasticity and stability. The stationarity test was conducted prior to Granger causality testing, whereas the subsequent diagnostic tests were performed on the model's outputs. The red boxes indicate a failed test, and the green represents a passed test. The p values are shown for the applicable tests, with a 0.1 critical value. **b)** The stability() function in R generated the following plot for one of the models in Fairfax County. As seen from the figure, mortality passed the test since none of the data exceeded the red boundaries.

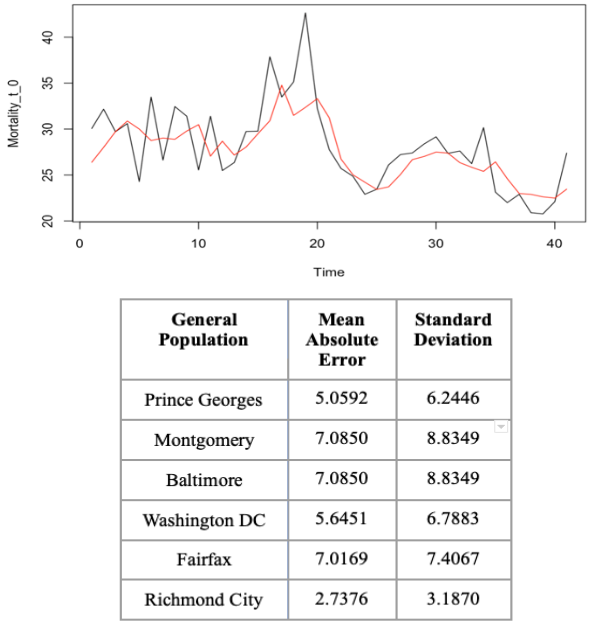
Test Data Set

Another large portion of validation for this project was the software validation portion. We wanted to test the ability of Granger and Bayesian algorithms implemented in R, to learn known causal relationships from a ground truth dataset that we constructed to have those causal links. In order to input pseudo-data that had distinct causal links, four variables were generated in MATLAB — Y, Z, N, and P. In order to generate these variables, Y was first defined as the first 300 values of a sin() function. Then, additional trend, random, and seasonality components were added in order to replicate similar data characteristics that were found in actualmortality and air pollutant data. An upward trend was added by increasing each Y data point by a random value between 0.01 and 1.5. Then, a random component was added by adding a random number within -1.5 and 1.5 to each Y value. Next, the variables Z, N, and P were made with respect to Y and a time delay of (t-1) was added. The variables Z, N, and P are defined as follows: $Z_{t-1} = 2Y^3 - Y^2 + 5Y$, $N = \sqrt[3]{3|Y|}$, and $P = 3Y + 2$. Once the test data set was completed, it was time-series decomposed in R and then subsequently run through the Granger causality code. All combinations of Y, Z, N, and P models were tried, and all models resulted in a p-value of 2.2E-16, which is how R outputs p-values less than 0.0001[19]. This indicates that we were able to learn the true causal links from the test data. In addition to running the test data set through the Granger causal code, the same statistical VAR tests done for the pollutant and mortality data was also completed. Similar to the air quality/mortality data, the test data passed about half of the statistical VAR tests. For the most part, the models passed the AC autocorrelation, stationarity, and the stability tests, and did not pass the serial autocorrelation, the normality, and the heteroscedasticity tests. This is a good sign because it demonstrates that even models built on data with known causal links can fail statistical tests.

The same data set, except the raw data instead of the residual data, was inputted into the Bayesian Network Model to test the algorithm's ability to learn the manufactured causal links from the data. The result was that the learn_dbn_struc function was only able to pick up on some of the relationships between the four variables, including a consistent causal link between variables Z and P. This is interesting because variables Z, N, and P were defined all in terms of Y, and Y did not have many causal links. This demonstrates a limitation with the *bnstruct* and *bnlearn* packages and their ability to decipher complex relationships. These links, or the lack thereof, may also be due to the noise and trend/seasonality components added to the data after variable definition.

Bayesian Mean Absolute Error

For Bayesian validation, the predict function, which uses past data to perform inference over test data was utilized to assess forecast accuracy for each general population model from each county. Figures illustrated the predicted data in red superimposed with the raw data outlined in black. An example figure for DC is seen in Figure 8. The average mean absolute error (MAE) and standard deviation (SD) were also calculated in R. As part of cross-validation, training and testing datasets for the general population models were altered and varied to include an 80-20 train-test split, a 60-40 split, a 40-60 split, and a 20-80 split, in which the MAEs and SDs were averaged. In an 80-20 train-test split, 80% of the data points are reserved for the training set and 20% are reserved for the testing set. As the training data set decreased from 80 to 20, the MAEs and SDs were observed to be higher on average given that less training data was being used to formulate the models thereby introducing the potential for more error. These MAEs and SDs were then recorded for each county's general population model as seen in Figure 8. Counties that required less interpolation and extrapolation tended to have the lowest MAEs and SDs.



General Population	Mean Absolute Error	Standard Deviation
Prince Georges	5.0592	6.2446
Montgomery	7.0850	8.8349
Baltimore	7.0850	8.8349
Washington DC	5.6451	6.7883
Fairfax	7.0169	7.4067
Richmond City	2.7376	3.1870

Figure 8. Bayesian Validation with Forecasting Accuracy. For all models, the predict function was utilized to plot mortality vs time with the predicted data in red superimposed with the raw data depicted in black as seen in the top graph. Mean absolute error (MAE) and standard deviation (SD) values were recorded for each of the counties' general population on the bottom table.

For Bayesian stratification by race, the MAE for each race's Bayesian models was also strikingly different - the white population of D.C. had a MAE of 0.38, with a standard deviation of 1.04. The Black population had a MAE of 4.70, with a standard deviation of 5.22, despite D.C. requiring very little interpolation for both races. This was consistent among every county. As for Bayesian stratification by age, the MAE and SD values decreased as the age groups increased. For example, in Prince George's 75-84 age group, the MAE and SD was 1.09 and 1.51, respectively. The MAE and SD for the 85+ age group was 0.9830276 and 1.238806. This trend was commonly observed across all counties. The lower error amongst older ages aligns with the fact that older age groups had more available data requiring less interpolation and thereby reducing any potential error. For Bayesian stratification by COPD, the MAEs and SDs were much lower for all counties compared to their general population model counterparts. For example, in DC's COPD population, the MAE and SD were 1.57 and 2.06, respectively. This aligns with the fact that a smaller portion of the population is being used, and therefore more specific subsets of data were analyzed. As a result, the *dbnR* package uses a less computationally intensive approach and is more accurate in detecting causal links.

Discussion

Bayesian and Granger Comparison

When comparing Granger and Bayesian models, one can see that temperature is repeatedly causally linked to mortality in both general population models. In both the Bayesian and Granger models, it can be seen that temperature is a major actor in causing respiratory-related mortalities in only the Black population. When combining temperature with an additional pollutant, the Granger models for only the Black population often outputted a p value less than 0.1, indicating the influence temperature has on respiratory health. SO_2 was also causally linked to mortality in both the Bayesian and Granger models for the Black population. For stratification by age and COPD, temperature was also causally linked to mortality in both models. However, these two models cannot necessarily be compared with each other as Granger utilized a combination of variables to demonstrate causality as opposed to relationships between pairs of variables in the Bayesian networks.

While we have established *potential* causal links with our models, this list is not exhaustive. Our data driven causal inference serves as a suggestion for further research to be conducted as more experimentation is needed to establish definite causality. However, compared to randomized control trials and cohort studies which can be unethical, time-consuming, and full of unaccounted for extraneous determinants, computational-based studies may serve as a more efficient way of testing air pollutant effects on health.

Limitations

Many limitations stemmed from the lack of data availability. Had daily raw data been available for all counties from 1999 to 2018, the models would have more successfully captured the complexities of climate epidemiology as opposed to more commonly collected monthly or yearly data. For example, public hospital admission data (HCUP) would have been a better public health outcome indicator than mortality due to respiratory disease because it would include a larger, more comprehensive set of patients and would detail the patterns of a consistently maintained hospital database on a daily timescale. If HCUP data was acquired, asthma data could have been analyzed as an effect of air pollutants. However, very few asthma cases lead to death and thus most geographic locations did not have enough data to analyze on an acceptable time scale. HCUP data could not be acquired due to funding restraints and legal hurdles.

In addition, Latinx, Native American and Asian populations lacked sufficient data and were omitted from analysis, presenting itself as an issue related to data diversity and model inclusivity. This could partially be due to such small populations of these races in certain DMV areas, but likely points to an issue of inaccurate demographic representation. It was seen that Black mortality data had to be interpolated much more than that of the white population, with the Black populations of some counties having very little or no data in the CDC Wonder Tool database. Since these sub-population groups required much more interpolation than others, their respective models may be less accurate than others, indicating major racial disparities in terms of analytical accuracy and model validity. Younger age groups also had large breaks in data.

Even though statistical p-values were used to detect causality, our definition of significance is subjective based on our critical value of 0.1. In addition, there is also a chance of type-1 error, in which the statistical test concludes causality when in actuality there is not, and therefore, further research should be conducted to verify our findings.

The failure of the models, including the test dataset, to pass many VARS statistical tests also presents itself as a computational limitation.

As for the Bayesian models, decreasing the node size loses some of the short-term connections, signifying the sensitivity of the dynamic Bayesian graph connections. In addition, blacklisting unwanted linkages introduced different linkages between a few variables and readjusted nodal relationships that were originally presented, demonstrating the sensitivity of the dynamic Bayesian model. Whitelisting may be used for further model advancement to link nodes that are already known to have a causal relationship. It may also be beneficial to incorporate low-pass filtering of the data which would remove excess noise and outliers.

Societal Impact

The development of these causal models helps explore the reality of climate change's impact on health outcomes, and more specifically, the effect of air quality on respiratory disease outcomes. The models allow for enhanced predictability of respiratory disease based on geographical location, environmental exposures and prior medical conditions. They can improve the predictability of higher respiratory disease rates in certain geographical areas with higher pollution rates, thus allowing for better allocation of resources and proactive risk management. Furthermore, the models serve to incentivize more research in terms of what behavioral adaptations should be taken to reduce the extent of climate change and its incontestable threats to health. With an improved understanding, the models will also foster the implementation of preventative action and adaptive resilience programs with prime focus on the reduction of carbon emissions, water resource management, and urban planning[20]. It is hoped that stricter regulations of power plants, industrial urbanized centers, and automobile pollutants, as well as better monitoring of construction sites, unpaved roads, fields, and fires will be prioritized[21]. Not to mention, through successful implementation, this model could promote higher standards in medical/climate data collection across states by highlighting the importance of acquiring frequent and readily accessible data.

Amidst the country's struggle to bring climate change under control, an even deeper, entrenched problem of health inequity in society is brought to light. Not only do our models transform how we view the boundaries and determinants of human health, but they provide evidence for the disproportionate effects of climate change and air particulate matter on socially vulnerable, marginalized communities. By quantifying disparities of climate change impacts by modeling racial subpopulation groups and counties of varying social vulnerability indices, our model investigates the socioeconomic and granular demographic factors that affect the resilience of communities[22]. However, for future studies, it would be beneficial to probe systems that produce and perpetuate inequalities in exposure to particulate matter and how these can persist by measuring particulate-matter emitting facilities in residential areas and comparing differences between places of varying demographics. Additional climate change drivers (extreme weather conditions, greenhouse gas emissions, biodiversity loss) and public health outcomes (infectious disease, mental health issues) could also be explored and implemented into the same model. Given its robustness and usage of universal parameters, the model can also be applied to other geographical areas to better understand the health impacts of climate change.

In conclusion, with our project, understanding the implications of climate change now will hopefully help foster the development of adaptation and mitigation efforts and predictive models to reduce future risks and impacts of climate change for all demographic groups. Pathways to resilient public health sectors are possible, building on the growing evidence-based understanding of the pronounced, persistent and pervasive threats climate change imposes on human health.

End Matter

Author Contributions and Notes

A.E., Z.S., and P.Y. wrote the paper, conducted research, and constructed the causal models using computational R software. A.E. created and analyzed the race-stratified models, Z.S. created and analyzed the COPD models, and P.Y. created and analyzed the age-stratified models. P.B. provided feedback and resources, and assistance in coding. The authors declare no conflict of interest.

Acknowledgments

We would like to thank Pam Bhattacharya, lead data scientist from the MITRE Corporation for defining the problem scope and approaches, and for her guidance and advice in our research and development throughout this project.

References

1. Climate change and health. https://www.who.int/news-room/fact-sheets/detail/climate-change-and-health.
2. The challenging politics of climate change. https://www.brookings.edu/research/the-challenging-politics-of-climate-change/.
3. Bronshtein, A. Train/Test Split and Cross Validation in Python. Medium https://towardsdatascience.com/train-test-split-and-cross-validation-in-python-80b61beca4b6 (2020).
4. Prince George's County, MD | Official Website. https://www.princegeorgescountymd.gov/.
5. Morbidity burden of respiratory diseases attributable to ambient temperature: a case study in a subtropical city in China | Environmental Health | Full Text. https://ehjournal.biomedcentral.com/articles/10.1186/s12940-019-0529-8.
6. Environmental Justice & Environmental Racism – Greenaction for Health and Environmental Justice. http://greenaction.org/what-is-environmental-justice/.
7. Region remains nation's least-congested large urban area | Greater Richmond Partnership | Virginia | USA. https://www.grpva.com/blog/2019/11/27/region-remains-nations-least-congested-large-urban-area/.
8. Health Effects of Ozone Pollution | Ground-level Ozone Pollution | US EPA. https://www.epa.gov/ground-level-ozone-pollution/health-effects-ozone-pollution.
9. Nitrogen Dioxide | American Lung Association. https://www.lung.org/clean-air/outdoors/what-makes-air-unhealthy/nitrogen-dioxide.
10. Shindell, D. T. Separating the influence of halogen and climate changes on ozone recovery in the upper stratosphere. J. Geophys. Res. 107, 4144 (2002).
11. America's formerly redlined neighborhoods have changed, and so must solutions to rectify them. https://www.brookings.edu/research/americas-formerly-redlines-areas-changed-so-must-solutions/.
12. [Traffic regulation and environmental pollution by particulate material (2.5 and 10), sulfur dioxide, and nitrogen dioxide in Metropolitan Lima, Peru] - PubMed. https://pubmed.ncbi.nlm.nih.gov/30183915/.
13. Vehicle Registration Historical Data | dmv. https://dmv.dc.gov/page/vehicle-registration-historical-data.
14. Nitrogen Dioxide (NO2) and Sulfur Dioxide (SO2) Secondary Air Quality Standards | Reviewing National Ambient Air Quality Standards (NAAQS): Scientific and Technical Information | US EPA. https://www.epa.gov/naaqs/nitrogen-dioxide-no2-and-sulfur-dioxide-so2-secondary-air-quality-standards.
15. The relationship between particulate matter (PM10) and hospitalizations and mortality of chronic obstructive pulmonary disease: a meta-analysis - PubMed. https://pubmed.ncbi.nlm.nih.gov/23323929/.
16. U.S. Census Bureau QuickFacts: Fairfax County, Virginia. https://www.census.gov/quickfacts/fairfaxcountyvirginia.
17. Hansel, N. N., McCormack, M. C. & Kim, V. The Effects of Air Pollution and Temperature on COPD. COPD 13, 372–379 (2016).
18. Washington, DC - Profile data - Census Reporter. https://censusreporter.org/profiles/16000US1150000-washington-dc/.
19. R Handbook: Hypothesis Testing and p-values. https://rcompanion.org/handbook/D_01.html.
20. Climate Change Prevention and Preparation --- Tracking California. https://trackingcalifornia.org/climate-change/prevention-and-preparation.
21. US EPA, O. Particulate Matter (PM) Basics. US EPA https://www.epa.gov/pm-pollution/particulate-matter-pm-basics (2016).
22. Flanagan, B. E., Gregory, E. W., Hallisey, E. J., Heitgerd, J. L. & Lewis, B. A Social Vulnerability Index for Disaster Management. J. Homel. Secur. Emerg. Manag. 8, (2011).