Saisriram Gurajala
DATA 512
Final Report
12/4/2024

Introduction

Resolving the connection between wildfire smoke and respiratory disease incidence is a salient and essential investigation for current times. In 2022 Chronic lower respiratory diseases, including asthma, have caused over 147,000 deaths at a rate of 44.2 per 100,000 people in the United States[1]. Additionally, the EPA has reported that prolonged exposure to wildfire smoke exacerbates lung disease in the short term and reduces lung function in the long term[2]. Reinforcing the potentially dire consequences of this connection, wildland fire smoke $PM_{2.5}$ exposure has been linked to adverse health outcomes and mortality in the United States[3]. The rising frequency and intensity of wildfires driven by climate change makes this potential causal relationship even more pressing. This analysis seeks to uncover how wildfire smoke impacts the incidence of respiratory diseases in the community of Dearborn, Michigan. This research unites the fields of ecological research and public health, orienting current knowledge towards an interdisciplinary phenomenon. In investigating this research question, I aim to establish a data-driven framework connecting natural disasters with chronic respiratory health outcomes.

The real-world implications of this work include enablement of public policy officials to design targeted interventions, effectively allocate resources, and mitigate the health impacts of wildfire smoke exposure. Moreover, the citizens of Dearborn may benefit from actionable findings which lead to better understanding of their real-world health risks. This work also contributes to a growing body of interdisciplinary research linking environmental phenomena and public health. Escalating climate-related disasters in recent years necessitate through and rigorous evaluation of this question. Refining our understanding of wildfire smoke's impact on human health empowers effective planning, containment, and proactivity from impacted communities.

Background/Related Work

Previous efforts in this project focused on creating a framework to quantify wildfire smoke impacts on Dearborn, Michigan, based on geospatial and temporal wildfire data. To source wildfires and generate an initial metric, I utilized United States Geological Survey data of wildland fires in the US and territories from 1800 to the present. Using this data, I generated annual estimates of wildfire smoke impacts over 60 years (1961–2021) and compared these estimates with available Air Quality Index (AQI) data. This work developed initial predictive and statistical models exploring the relationship between wildfire smoke and chronic respiratory

disease incidence. Fires within a 650-mile radius of Dearborn during the fire season (May 1–October 31) were considered. The metric considered two factors: fire size (measured in acres burned) and proximity to the city. Larger and closer fires were assumed to contribute more significantly to smoke exposure. These estimates were evaluated by comparison with AQI data from the U.S. EPA's Air Quality System (AQS). These comparisons provided initial insights into the alignment of modeled smoke impacts and observed air quality trends.

The initial smoke metric, calculated as the ratio of fire acreage burned to the squared distance from Dearborn, served as a first estimate. A 2018 study found wildfire smoke trended with the Palmer drought severity Index in California[4]. Therefore, I sought to refine the smoke metric via incorporation of local climate metrics in Wayne County, Michigan. The National Oceanic and Atmospheric Administration's (NOAA) "Climate at a Glance" dataset, provided by the National Center for Environmental Information (NCEI), was utilized for this purpose. This dataset furnished county-level coverage of temperature metrics such as average, maximum, and minimum temperatures; precipitation metrics including monthly and annual precipitation data; and drought metrics such as the Palmer Drought Severity Index. The key variables of average temperature, average precipitation, and palmer-drought severity index were selected for inclusion in the refined smoke metric (**Figure 1**).

After formulating a refined smoke metric, I proceeded to focus on assessing the metric's broader implications for public health. I specifically selected the incidence of chronic respiratory diseases and disregarded mortality since the number of deaths were observed to be essentially a linear function of time. Information about chronic respiratory disease incidence was sourced from the Global Burden of Disease (GBD) dataset from the University of Washington's Institute of Health Metrics and Evaluation (IHME). The dataset provided incidence (number of new cases) with state-level specificity for Michigan from 1990 to 2021. To predict the future incidence of chronic respiratory disease in Michigan, I needed to predict the wildfire smoke metric. I employed Holt-Winters exponential smoothing, a time series forecasting technique accounting for seasonality, trends, and noise in the data, to this end. I used cubic polynomial regression of the smoke metric to predict future incidence values, incorporating a linear term for the year to capture long-term temporal trends.

Saisriram Gurajala
DATA 512
Final Report
12/4/2024

Methodology

In designing this study, I sought to integrate robust analytical methods with a human-centered approach. I tried to ensure that the outcomes would be scientifically valid, accessible, and actionable for the affected Dearborn community. The goal of this project was to understand the long-term health impacts of wildfire smoke on populations in Dearborn, Michigan. I placed a particular focus on chronic respiratory diseases like asthma and chronic obstructive pulmonary disease (COPD). To achieve this, I selected methods which provide clear, understandable insights for health officials, policymakers, and the public. These methods ensure the findings can be effectively communicated and used to inform decisions such as public health interventions and resource allocation. Additionally, I filtered the data to wildfires within 500 miles to select for fires more impactful to the citizens of Dearborn, Michigan.

To forecast the impact of wildfire smoke, I employed Holt-Winters Exponential Smoothing. This forecasting method was chosen for its ability to capture both the short-term fluctuations and long-term trends in wildfire data. In my dataset wildfires were found to exhibit periodic (approximately three years in length) spikes, making Holt-Winters particularly well-suited to model the smoke impact. This method allowed me to generate reliable estimates of future smoke levels. In turn, these estimates arm health professionals with the critical foresight to anticipate periods of elevated risk. Proactive public health strategies, such as the deployment of medical resources and public advisories became accessible. Additionally, model results are simple to communicate with stakeholders without a deep technical background, enabling accessibility and high impact.

To model the relationship between wildfire smoke and respiratory disease incidence, I utilized polynomial ordinary least squares regression. I selected this approach for its interpretability, which is crucial when communicating complex relationship. Additionally, polynomial regression allows for both linear and non-linear relationships. This method provided me with the flexibility to capture intricate effects of wildfire smoke on public health over time. The linear year term helped assess whether the health impacts of wildfire smoke have evolved in a predictable pattern. The interpretability of the results made it easier to translate the results into actionable insights, enabling policymakers to understand the specific impact of smoke exposure on health outcomes and plan accordingly.

Ethical considerations were central to the design and execution of the work. For instance, the health data sourced from the Global Burden of Disease (GBD)

Saisriram Gurajala
DATA 512
Final Report
12/4/2024
dataset was handled in strict accordance with the terms set by the University of Washington's Institute for Health Metrics and Evaluation (IHME). I ensured in wrangling these data that the privacy and confidentiality of individuals represented in the dataset were maintained. In addition to ethical data handling, I prioritized transparency and clarity in communicating the results. I focused on making findings not only scientifically sound but also accessible to communities who might be directly affected by wildfire smoke. Therefore, I generated clear visualizations and avoided complex jargon. Additionally, I supplied as much data and code as possible under an open MIT license to make the study reproducible. The design of the study, therefore, attempted to provide the Dearborn community with understandable tools to anticipate and mitigate the impacts of wildfire smoke.

Findings

The integration of climate measurements into the smoke metric formulation improved correspondence of the metric with air quality data. The refined metric extended the original inverse square law-based formulation. It accomplished this by incorporating the squared ratio of temperature to the product of average precipitation and drought severity index. Therefore, the original metric was scaled by average climate conditions in the year the fire occurred (**Figure 1**). This enhanced metric demonstrated a significant improvement in its alignment with annual Air Quality Index (AQI) values, achieving a Pearson correlation coefficient of $R=0.425$ and a p-value of 0.008. The previous metric, when filtered to fires within 500 miles of Dearborn, achieves a Pearson correlation coefficient of -0.04 and a p-value of 0.828. Therefore, the new smoke metric exhibits a statistically significant correspondence with air quality trends over time and represents a substantial improvement to the previous metric (**Figure 2**).

Holt-Winters exponential smoothing was employed to forecast the smoke metric over time, and a five-year rolling average was used to examine temporal trends. The results revealed an overall upward trajectory in the smoke metric, with intermittent spikes corresponding to periods of heightened wildfire activity. Figure 3 illustrates these trends, with the blue line representing computed respiratory disease incidence from observed data and the red line depicting predicted incidence based on the forecasted smoke metric. The predicted incidence mirrored the temporal patterns of the observed values, showing a general increase over time punctuated by specific peaks.

Polynomial regression modeling was used to predict current chronic respiratory disease incidence, incorporating year, and exponentiated terms of the

moving average of smoke metric as predictors. Several different versions of the smoke metric were tested, including a moving average, an exponentially weighted moving average, and the raw smoke metric. Of these the moving average performed the best, and the model achieved an adjusted R-squared value of 0.669. Additionally, the results showed that moving average of smoke metric and year accounted for a large portion of the variance in the observed incidence data. The p-value of 2.10×10^{-5} further confirmed the statistical significance of the model (**Figure 4**). The model, therefore, captures the complex relationship between these variables and chronic respiratory health disease incidence (**Figure 5**).

Projections of chronic respiratory disease incidence from 2021 to 2050, derived from the forecasted smoke metric, revealed a steady upward trend over the analyzed period. After a slight dip post-2020, incidence values began to rise consistently, surpassing 2020 levels by the mid-2020s. The model generally predicts a year-over-year increase in incidence rates. Therefore, the results predict a continuous upward trajectory throughout the projection period. Figure 6, displaying these forecasts, shows clear evidence of rising incidence values of chronic respiratory disease as time progresses. Taken together, these results establish a clear link between chronic respiratory disease and a moving average of smoke metric. They additionally suggest rising chronic respiratory disease incidence in coming years.

Discussion/Implications

The findings from this study suggest clear and actionable links between wildfire smoke, air quality, and public health for citizens and agencies in Dearborn, Michigan. The refined smoke metric integrates local climate variables such as temperature, precipitation, and drought severity with the much broader USGS datasets. Therefore, it has enabled a more nuanced understanding of how these factors collectively influence air quality in the region. By incorporating these measurements into the smoke metric, I was able to improve its predictive power. Additionally, the improved metric resulted in more accurate forecasting of the impact of wildfire smoke on air quality. These results establish this connection between respiratory disease and smoke impact clearly. The projected rise in wildfire smoke intensity over the next several years is pressing. Specifically, I found that the cumulative effect of high levels of exposure to wildfire smoke will result in a steady increase in chronic respiratory diseases.

Given the projections of increasing wildfire smoke impact, it is imperative that the city of Dearborn take immediate action to address these findings. City

Saisriram Gurajala
DATA 512
Final Report
12/4/2024
council members, the city manager, and the mayor must implement strategies to mitigate the effects of wildfire smoke on public health. Short-term measures may include expanding air quality monitoring, improving air filtration, and issuing public health advisories during periods of high smoke exposure. Public education campaigns to inform residents during wildfire season could also be an immediate step. However, long-term actions will likely have the most substantial impact. Healthcare infrastructure supporting chronic respiratory conditions, air quality regulations, and urban planning will all be significant in minimizing the future health burden. The city has a limited window of time to act, and the implementation of concrete plans is required.

The human-centered data science principles applied in this project helped ensure that my findings were scientifically robust as well as practically relevant to the people of Dearborn. One such principle was exhibited by my decision to model the smoke metric as a rolling average. The effects of wildfire smoke are not always immediate, and the cumulative impact over several years may be more relevant than the specific impact of a single year. By using a rolling average with a five-year window, I was able to capture the delayed and compounded effects of repeated smoke exposure. Additionally, I selected Holt-Winters exponential smoothing for forecasting because of its ability to handle seasonal and cyclical patterns, which are inherent in both wildfire occurrences and air quality trends. The flexibility of this approach is crucial from a human-centered perspective. It provides city officials and residents with practical, forward-looking insights, informing decision-making in the face of an uncertain future.

Another human-centered design principle central to this project was the use of polynomial ordinary least squares (OLS) regression to predict chronic respiratory disease incidence. This model captured both linear and non-linear relationships between the smoke metric, year, and incidence. The flexibility of polynomial regression made it a powerful tool for interpreting and modeling complex phenomena such as human health. The model's transparency and ease of understanding ensured that stakeholders could grasp the key drivers of respiratory disease trends and make informed decisions. This interpretability is critical for making findings accessible and actionable, allowing for effective communication of complex health risks.

Additionally, as much as possible I have integrated principles of reproducibility and access of my work, code, and datasets. This effort has improved the accuracy and reliability of my findings. In the context of enabling accessibility of my results, I have ensured the findings are relevant, actionable, and

understandable for the people most affected. My work focused on long-term impacts and ensured the flexibility of my models while also prioritizing explainability. I have thereby developed a framework that can help Dearborn's leadership and residents respond to the growing challenge of wildfire smoke in a way that is both effective and equitable. My sincere hope is that the citizens and municipality of Dearborn swiftly and strategically plan and execute mitigation efforts for the rising risk of wildfire smoke.

Limitations

Valuable insights into the relationship between wildfire smoke and chronic respiratory diseases have been uncovered. However, several key limitations and assumptions must be considered when acting upon results. First, there are licensing restrictions associated with the use of the Global Burden of Disease (GBD) dataset, limiting accessibility and potential for broader dissemination. According to the terms of use, users are prohibited from copying, reproducing, or redistributing the dataset outside of their organization. Additionally, actions like decrypting, reverse-engineering, or compiling a similar dataset from the GBD data are prohibited. Therefore, I am unable to share the respiratory disease data used for this work, significantly impacting reproducibility of the study.

Data-related limitations also exist in terms of the temporal and spatial scope of the datasets used. The GBD dataset provides health data from 1990 onward, making exploration of long-term trends infeasible. The wildfire data, while useful, comprises only about 1,500 fires with an uneven distribution across years. The number of fires per year ranged from a minimum of four to a maximum of 130. This source of variability resulted in uncertain approximation of health events for a given year when modeling with smoke metric. This uneven distribution could potentially skew the analysis and may result in poor modeling for years with fewer fire events. Another challenge encountered was the need to rescale the drought severity index when constructing the smoke metric. The original drought index values included negative numbers, which could result in inconsistency for the smoke metric. To address this, I rescaled the drought severity index to lie within zero and one. While this approach preserved the relative distance between points, it may have resulted in loss of nuance for negative drought severity values, thereby affecting the accuracy of the smoke metric in certain regions or years.

From a statistical modeling standpoint, both the polynomial regression and Holt-Winters exponential smoothing methods have their own assumptions and limitations. Polynomial regression shares its assumptions with ordinary least

square regression. Errors are assumed to be normally distributed with constant variance, which may not be true for complex and non-linear temporal phenomena. Additionally, it assumes that the relationship between the predictors and the outcome can appropriately modeled as a linear function. Finally, the polynomial regression model does not account for potential autocorrelation between observations. Autocorrelation is a common phenomenon occurring in time series datasets, affecting the reliability of predictions[5].

Holt-Winters exponential smoothing also has inherent limitations. The exponential smoothing method assumes that future values are a weighted average of past observations: more recent data is therefore weighted more. While this method is useful for forecasting trends, it may not fully capture sudden, unforeseen changes or outliers. Events such as extreme wildfire seasons or shifts in environmental conditions would therefore disrupt the reliability of the model. Moreover, exponential smoothing is sensitive to the choice of smoothing parameters, and improper selection can lead to inaccurate forecasts. Additionally, the method assumes that the time series is stationary in terms of its overall trend. This stationary nature may not be the case in a dynamic and rapidly changing environment like wildfire smoke and air quality. The model itself, finally, can provide erroneous long-term forecasting, so predictions 30 years into the future may not be reliable[6].

Conclusions

This study reveals several important insights into the relationship between wildfire smoke and chronic respiratory disease incidence in Dearborn, Michigan. The improved smoke metric, which integrates climate variables such as temperature, precipitation, and drought severity, significantly enhanced its alignment with the Air Quality Index (AQI). The improved smoke metric achieved a statistically significant Pearson correlation ($R=0.425$, $p=0.008$) with AQI per year. Forecasting the smoke metric with Holt-Winters exponential smoothing resulted in a general upward trend in smoke exposure. These trends were reflected in the predicted respiratory disease incidence, showing a consistent increase over time. The model, with an adjusted R-squared of 0.669, accurately captured the complex relationship between smoke exposure and disease outcomes. Projections indicate a steady rise in chronic respiratory disease incidence from 2021 to 2050, surpassing 2020 levels by the mid-2020s.

The findings underscore the importance of integrating environmental factors into public health models and emphasize the growing impact of wildfire smoke on

respiratory health. Through human-centered data science principles, this study prioritized interpretability, reproducibility, and accessibility. Decisions such as employing rolling averages for the smoke metric reflect the heuristic knowledge that people experience and react to environmental changes over time. Similarly, the use of polynomial regression highlights the need for models that can capture the complexities of human life while making the results meaningful for stakeholders.

Incorporating these human-centered approaches enables facile communication, and in term actionability, with city officials, public health planners, and residents alike. The model's transparency and explainability contribute to informed decision-making, understanding, and trust of the results of my work. Conducting this study has been a reminder of the importance of keeping people at the heart of data science. Data science is not just about numbers, algorithms, or metrics. Fundamentally, it is about the lives, health, and well-being of the individuals who make up a community like Dearborn. Chronic respiratory diseases aren't just statistics. They represent the struggles of individuals who face tangible challenges in their daily lives, exacerbated by factors like wildfire smoke. Therefore, this study has helped me ground abstract data science work in the lived experience of real people.

Saisriram Gurajala
DATA 512
Final Report
12/4/2024

References

1. CDC. (2019). *FastStats - Chronic Lower Respiratory Disease*. Center for Disease Control and Prevention. https://www.cdc.gov/nchs/fastats/copd.htm
2. US EPA. “Health Effects Attributed to Wildfire Smoke.” *US EPA*, 13 Aug. 2019, www.epa.gov/wildfire-smoke-course/health-effects-attributed-wildfire-smoke.
3. Ma, Yiqun, et al. “Long-Term Exposure to Wildland Fire Smoke PM $_{2.5}$ and Mortality in the Contiguous United States.” *Proceedings of the National Academy of Sciences*, vol. 121, no. 40, 24 Sept. 2024, https://doi.org/10.1073/pnas.2403960121.
4. Chen, Angela. “Evaluating the Relationships between Wildfires and Drought Using Machine Learning.” *International Journal of Wildland Fire*, vol. 31, no. 3, 3 Mar. 2022, pp. 230–239, https://doi.org/10.1071/wf21145.
5. Frost, Jim. “7 Classical Assumptions of Ordinary Least Squares (OLS) Linear Regression - Statistics by Jim.” *Statistics by Jim*, 15 Mar. 2019, statisticsbyjim.com/regression/ols-linear-regression-assumptions/.
6. “Exponential Smoothing for Time Series Forecasting.” *GeeksforGeeks*, 27 May 2024, www.geeksforgeeks.org/exponential-smoothing-for-time-series-forecasting/.

Data Sources

USGS wildfire dataset

“Combined Wildland Fire Datasets for the United States and Certain Territories, 1800s-Present | U.S. Geological Survey.” *Usgs.gov*, 2022, www.usgs.gov/data/combined-wildland-fire-datasets-united-states-and-certain-territories-1800s-present. Accessed 5 Dec. 2024.

EPA AQS API

“Obtaining AQS Data | US EPA.” *US EPA*, 26 June 2015, www.epa.gov/aqs/obtaining-aqs-data. Accessed 5 Dec. 2024.

NCEI Climate at a Glance

NCEI.Monitoring.Info@noaa.gov. “Climate at a Glance | County Time Series | National Centers for Environmental Information (NCEI).” *Noaa.gov*, 2024, www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/county/time-series/MI-163/pcp/1/0/1965-2024?base_prd=true&begbaseyear=1901&endbaseyear=2000.

IHME GBD Dataset

The Institute for Health Metrics and Evaluation. “Global Burden of Disease (GBD).” *Www.healthdata.org*, 2020, www.healthdata.org/research-analysis/gbd.

Saisriram Gurajala
DATA 512
Final Report
12/4/2024

Figures

Figure 1: Improved Smoke Metric Formulation

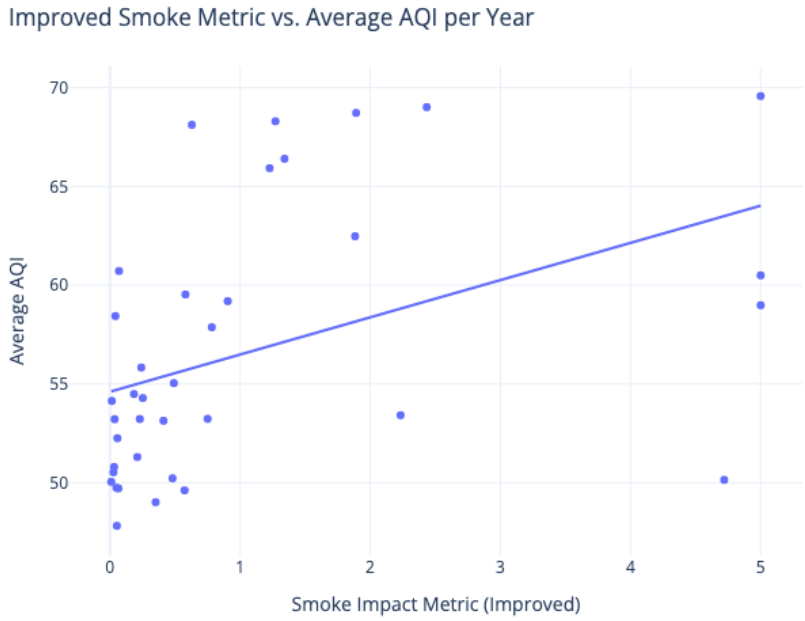
$$SmokeMetric_{fire} = f(acreage_{fire}, distance_{fire}, temperature_{fire}, dsi_{fire}, precipitation_{fire})$$

$$f(acreage_{fire}, distance_{fire}, temperature_{fire}, dsi_{fire}, precipitation_{fire}) = \frac{acreage_{fire}}{distance_{fire}^2} * (\frac{temperature_{fire}}{precipitation_{fire} * dsi_{fire}})^2$$

$$SmokeMetric_{year} = \sum_{x=0}^{n} f(acreage_x, distance_x, temperature_x, dsi_x, precipitation_x)$$

Description: The improved smoke metric leverages the inverse square law formulation devised in previous work. It additionally incorporates the ratio of temperature to the product of average precipitation and drought severity index. It squares this ratio and multiples by the original metric.

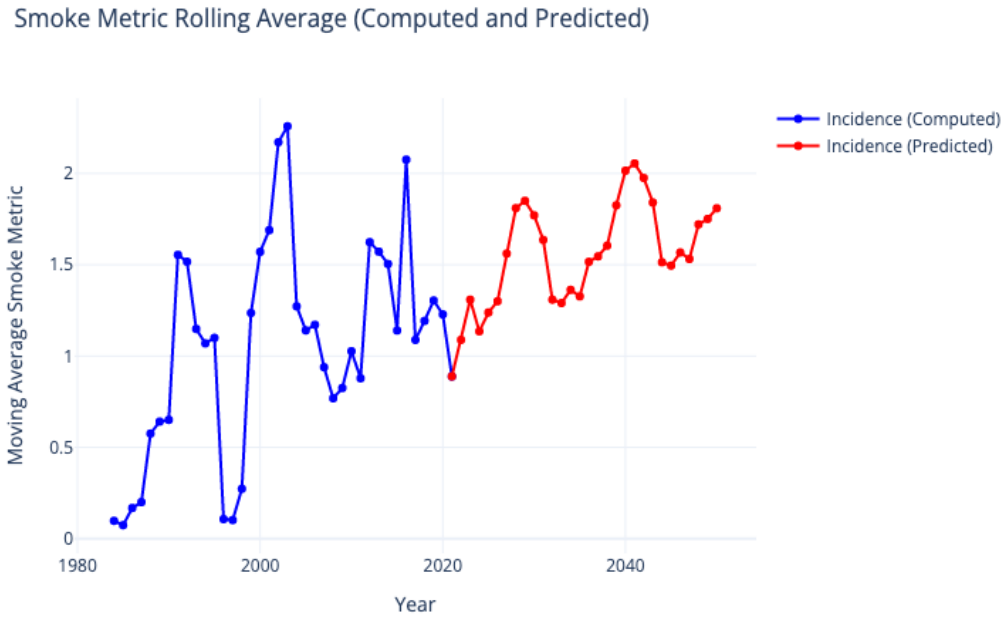
Figure 2: Correspondence of improved smoke metric with Air Quality Index



Description: The improved smoke metric showed significantly higher correlation relative to the previous version. A Pearson correlation between the two was highly significant at a p-value of 0.008 and R = 0.425.

Saisriram Gurajala
DATA 512
Final Report
12/4/2024

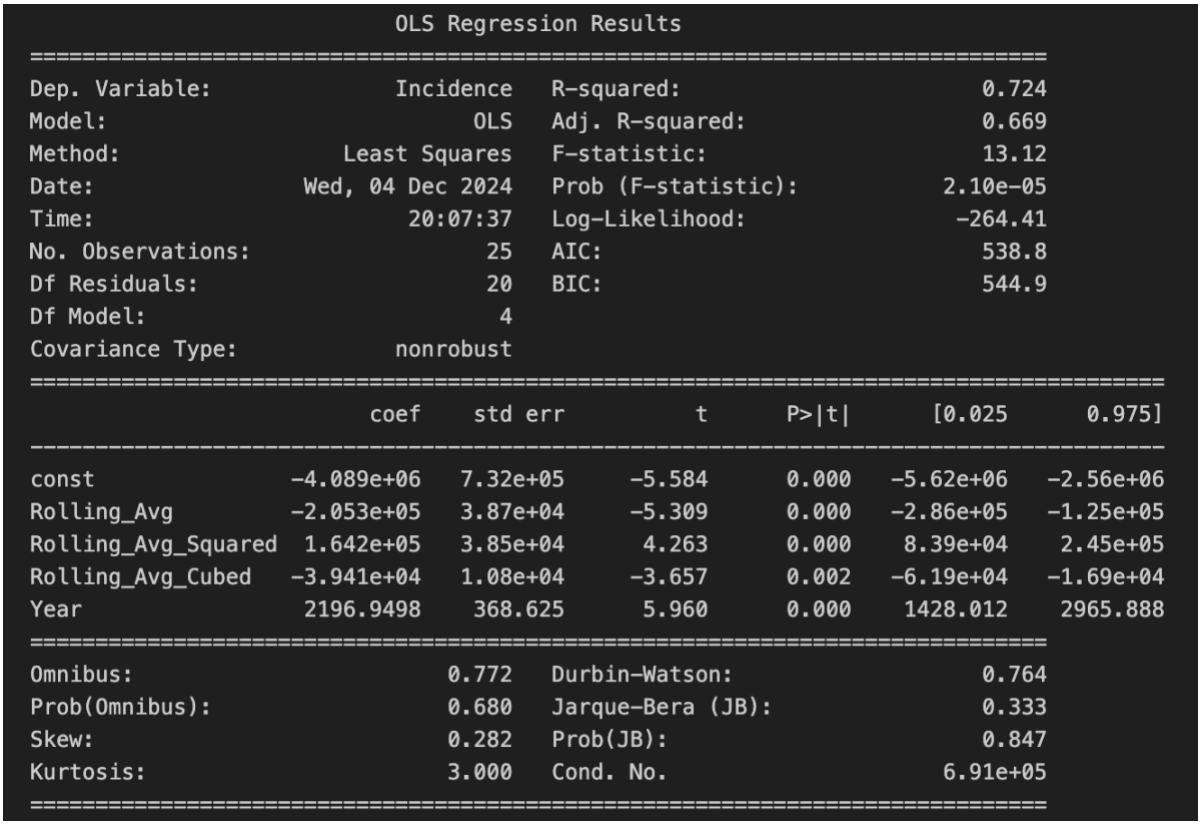
Figure 3: Rolling Smoke Metric Forecasting



Description: The smoke metric was forecasted with Holt-Winters forecasting, and a rolling average was computed with a window size of eight years. The computed incidence (from the source data) is represented via the blue line, whereas the predicted incidence is represented via the red line. The red line replicates the trend seen in the blue line, which is a general increase punctuated with occasional spikes.

Saisriram Gurajala
DATA 512
Final Report
12/4/2024

Figure 4: Summarizing Polynomial Regression Results of Best Performing Model



OLS Regression Results

Dep. Variable:	Incidence	R-squared:	0.724
Model:	OLS	Adj. R-squared:	0.669
Method:	Least Squares	F-statistic:	13.12
Date:	Wed, 04 Dec 2024	Prob (F-statistic):	2.10e-05
Time:	20:07:37	Log-Likelihood:	-264.41
No. Observations:	25	AIC:	538.8
Df Residuals:	20	BIC:	544.9
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-4.089e+06	7.32e+05	-5.584	0.000	-5.62e+06	-2.56e+06
Rolling_Avg	-2.053e+05	3.87e+04	-5.309	0.000	-2.86e+05	-1.25e+05
Rolling_Avg_Squared	1.642e+05	3.85e+04	4.263	0.000	8.39e+04	2.45e+05
Rolling_Avg_Cubed	-3.941e+04	1.08e+04	-3.657	0.002	-6.19e+04	-1.69e+04
Year	2196.9498	368.625	5.960	0.000	1428.012	2965.888

Omnibus:	0.772	Durbin-Watson:	0.764
Prob(Omnibus):	0.680	Jarque-Bera (JB):	0.333
Skew:	0.282	Prob(JB):	0.847
Kurtosis:	3.000	Cond. No.	6.91e+05

Description: Summary of ordinary least squares polynomial regression results for the best performing model (using rolling average) for prediction. This model resulted in an adjusted R-squared of 0.669 with an F statistic of 2.10e-05. We can see a clear and significant positive correlation with the squared rolling average with the incidence of diseases.

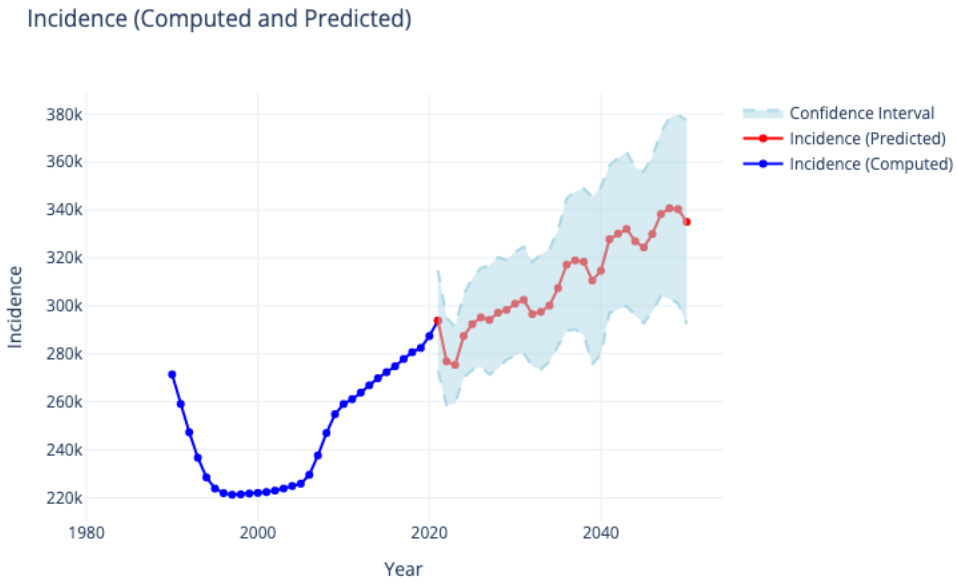
Saisriram Gurajala
DATA 512
Final Report
12/4/2024

Figure 5: Predicting Incidence from Smoke Metric and Year

Description: Using a polynomial ordinary least squares regression model, we predicted incidence from a linear combination of year, and various exponentiated forms of the smoke metric. Here we found these variables serve as a high-quality prediction of existing incidence values, with an adjusted r-squared of 0.669 and a p-value of 2.10e-05 when modeling this relationship.

Saisriram Gurajala
DATA 512
Final Report
12/4/2024

Figure 6: Predicting Future Chronic Respiratory Disease Incidence from Forecasted Rolling Smoke Metric



Description: Using polynomial regression, I predicted the incidence of chronic respiratory diseases for Michigan via the forecasted smoke impact metric values. While the model predicts a slight drop in 2020, it predicts a consistent increase year over year from 2021 onwards. Eventually the incidence will catch up to 2020 values and exceed it.