Wildfire Smoke and Population Health: A Case Study in Community Prioritization and Risk Assessment in Gainesville and Alachua County, Florida

Derek Tropf[1]

[1]MSDS Program, University of Washington

Abstract

This analysis uses Gainesville and Alachua County, Florida as a case study for a community looking to understand and respond to wildfire smoke affects on population health, specifically asthma and heart attacks. Using publicly available data from the United States Geological Survey, Environmental Protection Agency, and the Florida Department of Health, a "smoke metric" is created used to estimate air pollution impacts due to wildfire smoke. A physics informed model inconjunction with a linear regression model was leveraged to determine that wildfire smoke does not have a significant effect on the associated health conditions in the region of Alachua County. The implications of this study include the need to collect data that is more temporally granular and to incorporate more datasets and subject matter experts to create a more useful, accurate, and trustworthy model. The findings of this analysis suggest that having more confidence in the risks a community faces associated with wildfire smoke allows for better resource allocation and legislative prioritization.

1 INTRODUCTION

Wildfire events have increased in frequency, intensity, and area burned globally [1]. Although it may be difficult to notice trends of frequency without scientific analysis, millions of people living through the deleterious consequences of wildfires over the last decade are acutely aware of how extreme wildfires have become. The 2019 Kincade fire of Sonoma County, California showcased how much vulnerable populations (e.g., populations with lower income or disabilities) are disproportionately impacted by wildfires [2]. Losing a home due to an extreme natural event is devastating and instantaneous – however, smoke due to wildfires is also responsible for sinister health effects that are note quite as obvious. Wildfire smoke is comprised of a mixture of gaseous pollutants (for example, carbon monoxide), hazardous air pollutants (such as chemicals from lithium-ion batteries), water vapor, and particle pollution [3]. Ultimately, trees are not the only things that burn in a wildfire, everything from cars to chemicals are fair game as well. Breathing in such toxic fumes over the course of as little as a few days has been directly attributed to aggravation of lung disease,

bronchitis, heart failure, stroke, as well as many other respiratory and cardiovascular conditions [4]. According to the Natural Resources Defense Council, wildfire smoke in just the United States was responsible for an estimated 6,200 respiratory hospital visits and 1,700 deaths totaling $16 billion in health care costs in 2009 [5]. With extreme wildfires such as those that occurred in Eastern Canadian in 2023 – responsible for producing "Martian-like" conditions in New York City – it is becoming increasingly evident that all factors of life from people to cities to countries are susceptible to the dangerous consequences associated with wildfires [6]. With such well understood effects, it is imperative that local communities assess their risk to wildfire smoke exposure to be adequately prepared for potential population health implications.

Typically associated with football and hurricanes, the city of Gainesville – home to the University of Florida – in Alachua County, Florida is one of the fastest growing metros in the United States and has an increasingly diverse population [7]. Medical research suggests that diverse populations in terms of age, race, and socioeconomic status present unique challenges [8]. In an attempt to respond to changing wildfire trends and demographics, publicly available data is leveraged alongside modern Machine Learning (ML) methods to inform government and public health officials in their decision making processes. First, an analysis is conducted to determine the current state of air quality, wildfire risk, and public health for Gainesville and greater Alachua County, Florida focusing on asthma and heart attacks. Second, an ML based model is developed to help create a novel "smoke metric" which provides a quantitative estimate of wildfire smoke on public health. Finally, a recommendation is made to the community to improve upon data collection and regularly conduct "public health checkups" including wildfire smoke assessments relating to public health.

2 Background, Related Work, & Exploratory Data Analysis

To understand the intersection of wildfire smoke and population health, with use from community officials in mind, publicly available data was procured from federal and state level government organizations.

2.1 Wildfire Data

Wildfire data was obtained from the United States Geological Survey (USGS). The "Combined wildland fire datasets for the United States and certain territories, 1800s-Present (combined wildland fire polygons)" contains information dating back to the mid 1800s of fires that have occurred across the United States [9]. Nested JSON objects makeup the dataset with each fire sharing the same attributes. Data processing and cleaning was conducted to reduce the total size for purposes of efficient analysis and to account for changes in modern technology and practices used in measuring wildfire data. The dataset includes fire during the designated fire season which occurs between the months of May and October. To distill the dataset, two filtering methods were used. First, all fires a distance greater than 650 miles from Gainesville, FL were removed. To determine a fire's distance from Gainesville, the distance from Gainesville of each Polygon object that makes up the fire perimeter was computed and averaged. Second, fires that occurred before 1961 were removed

due to the advent of satellite technology used for wildfire data collection. The final attributes of the cleaned dataset can be found in **TABLE 1**.

Attribute	Datatype	Description
year	integer	Year of fire
name	string	Name of fire
size	float	Size of fire in acres
type	string	Type of fire (e.g Prescribed Fire)
ring	Shapely MultiPolygon Object	Lat/Lon positions of perimeter of fire
distance	float	Average distance of fire perimeter Polygon from Gainesville, FL

Table 1: Attributes from cleaned subset of USGS wildland fire dataset. Original attribute names have been changed for simplicity to the ones shown in this table. The subset was comprised of data from years 1961 to 2021 and fires within 650 miles of Gainesville, FL.

To obtain an initial intuition of the wildland fire data near Gainesville, the total number of acres burned annually was plotted as seen in **FIGURE 1**. Other than a particularly troublesome fire season in 2011, there is a clear and steady increase in acres burned in and around Gainesville since the mid 1980's.

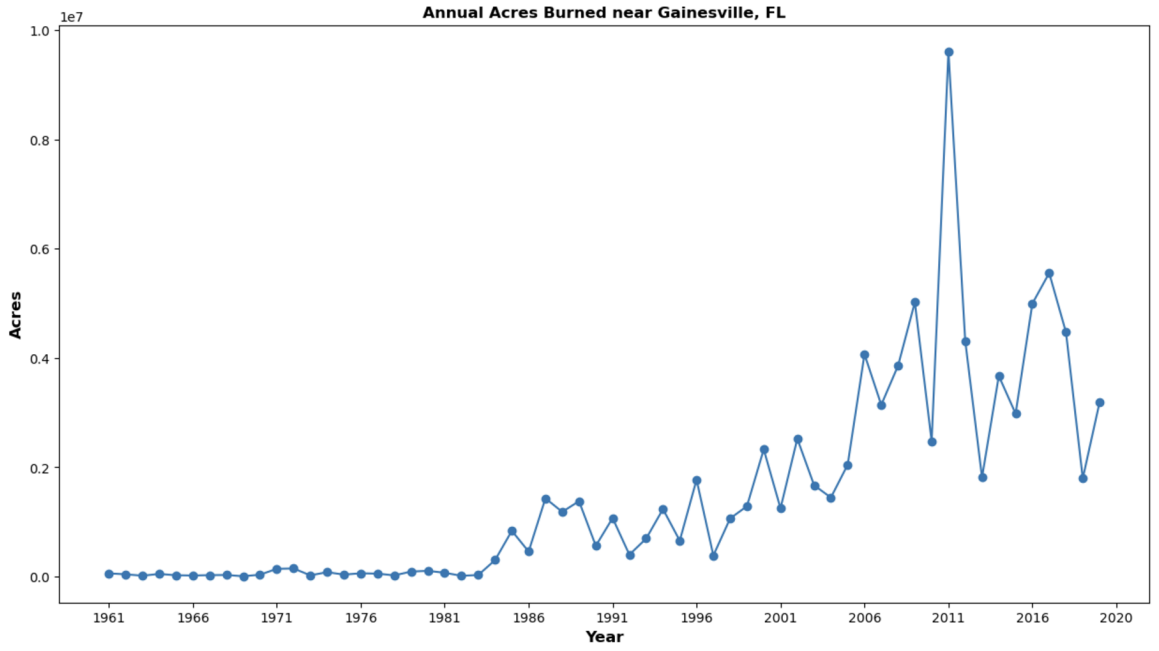

Figure 1: Annual acres burned within 650 miles of Gainesville from 1961 to 2020. Y-axis is 10^7

A geographical distribution of fires from 2015-2020 in **FIGURE 2** shows a consistent trend of fires clustered in the same areas each year – however some years exhibit the evidence of substantial burning in areas with increased fire fuel such as what occurred in 2016 with far more fires in the Appalachia region of Tennessee.

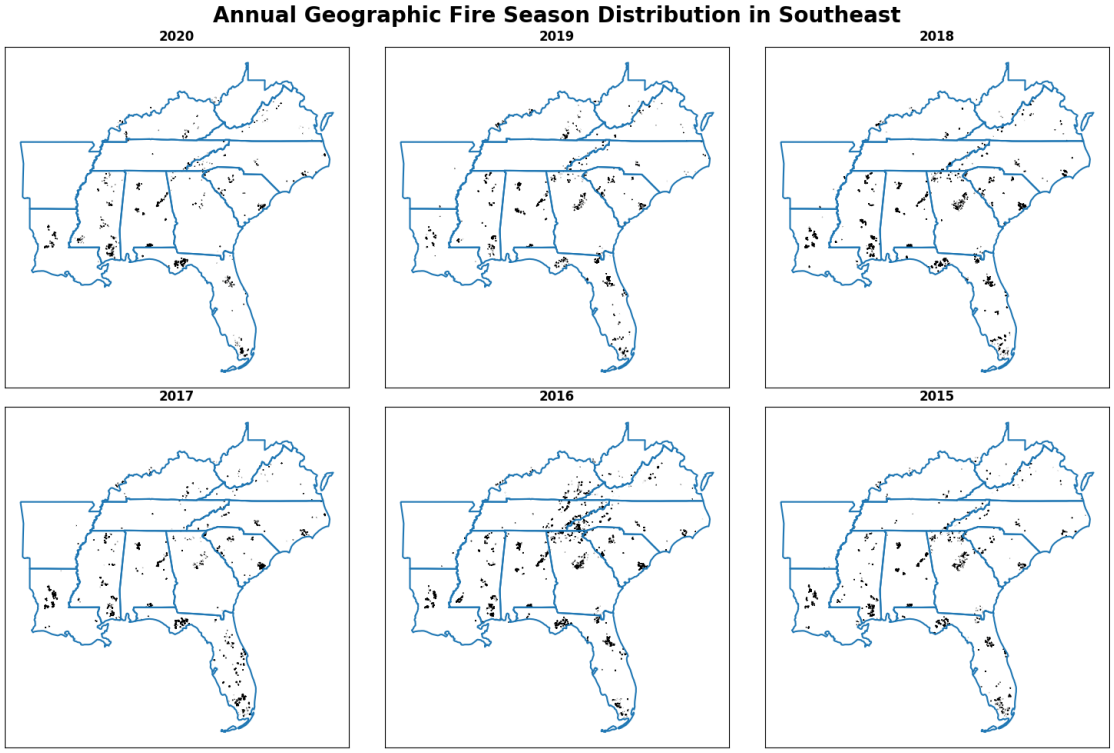

Figure 2: Distribution of all fires events within 650 miles of Gainesville, FL according to the USGS wildfire dataset from 2015-2020. Notice the consistent presence of certain clusters that appear each year. The year 2016 shows more activity in the Appalachia region of Tennessee where there exist far more fuel for fires.

2.2 Air Quality Data

The Air Quality Index (AQI) is a quantitative estimate of air quality developed by the United States Environmental Protection Agency (EPA) that measures gaseous and particulate pollutants [4]. EPA Air Quality System (AQS) API was used to access air quality monitoring station data throughout the United States. The closest AQS, roughly 10 miles from Gainesville city center, is located in Paynes Prairie Farm and monitors Ozone (O3) as well as Particle Pollution (PM2.5) [10]. The resulting data returned from the AQS API is extremely nested JSON which required substantial cleaning. First, only data from the Alachua County AQS was extracted. Second, the measurements were filtered down to a daily frequency and measurements taken outside of the regular fire season (to match USGS wildfire data) were removed. Finally, the data was split into two subsets: gaseous and particulate pollutants. The resulting attributes are the same for each dataset and can be seen in **TABLE 2**.

To get an idea of air quality trends annual combined AQI was plotted as seen in **FIGURE 3**. The calculation of annual AQI was made by taking AQI measurements for both particulate and gaseous pollutants and summing together the calculate AQI for all daily measurements. Although it appears that air quality started to improve since roughly 2004, there does seem to be a steady

Attribute	Datatype	Description
date_local	string	year-month-day of measurement
sample_duration	string	Interval of time between each measurement
aqi	float	Air Quality Index

Table 2: Attributes of cleaned data from EPA AQS API. Each day contains multiple measurements with seperate AQI score for each measurement. The same schema above applies to both the gaseous and particulate datasets.

state in the amount of pollution in the air near Gainesville within the last decade. Note that AQI measurements only go far back as 1973 due to the creation of the EPA of the same year.

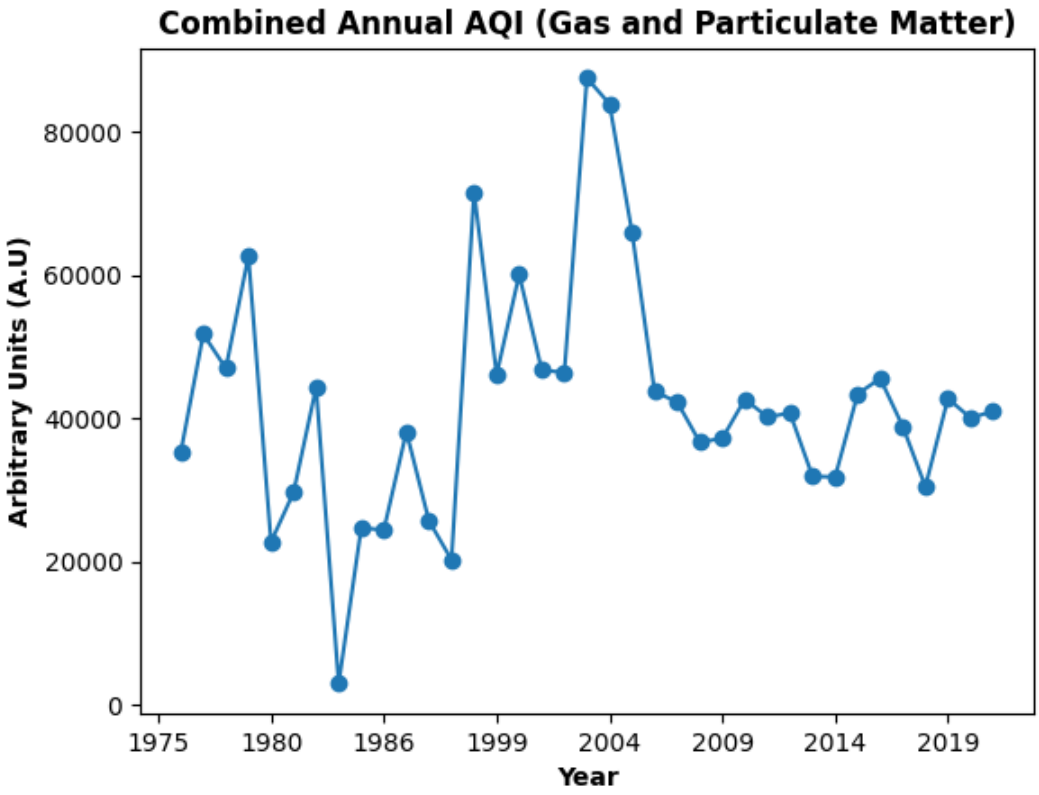

Figure 3: Total annual AQI for particulate and gaseous pollutants from all daily measurements. Note the y-axis is in Arbitrary Units (A.U) due to the unit-less nature of AQI.

2.3 Health Data

Those that suffer from asthma or are susceptible to cardiovascular problems are particularly vulnerable to the dangers of wildfire smoke and air pollution [11]. To be as actionable as possible for governing officials, hospitalization data related to asthma and heart attacks was acquired from the

Florida Department of Health for Alachua County. The data is an annual aggregation and is the most temporally granular dataset publicly available for the geographic region of Gainesville and Alachua County. Other than slight transformations for data processing purposes, the health data seen in **TABLE 3** and **TABLE 4** has not been altered in any way.

The similarities in health data go beyond their respective schemas as seen in both **FIGURE 4** and **FIGURE 5**. Each plot, comprised of six total subplots, reveals hospitalization trends for asthma and heart attack related events. Each condition shares two primary features: a sustained increase of events in the mid 2010's and a massive drop off in the year 2020 followed by far lower event trends compared to previous years. The 2020 drop off is hypothesized to be from the COVID-19 pandemic which overwhelmed hospitals across the country and world. As a result of the inconsistent data collection from 2020 on, all analysis leveraging health data will be truncated to 2019.

Attribute	Datatype	Description
Year	integer	Year of data aggregation
Hospitalizations per 10,000 (CRUD Rate)	float	Count÷Mid-year Population*Const
Emergency Department Visits per 10,000 (Age-adjusted Rate)	float	Rate account for age
Emergency Department Visits per 10,000 (CRUD Rate)	float	CRUD for emergencies
Number of Emergency Department Visits	integer	Annual count of all visits
Hospitalizations per 10,000 (Age-adjusted Rate)	float	Age-adjusted for hospitalizations
Number of Asthma Hospitalizations	integer	Annual count of all hospitalizations

Table 3: Attributes for asthma related events in Alachua County from 2005 to 2023. All data attributes are annual.

Attribute	Datatype	Description
Year	integer	Year of data aggregation
Hospitalizations per 10,000 (CRUD Rate)	float	Count÷Mid-year Population*Const
Emergency Department Visits per 10,000 (Age-adjusted Rate)	float	Rate account for age
Emergency Department Visits per 10,000 (CRUD Rate)	float	CRUD for emergencies
Number of Emergency Department Visits	integer	Annual count of all visits
Hospitalizations per 10,000 (Age-adjusted Rate)	float	Age-adjusted for hospitalizations
Number of Heart Attack Hospitalizations	integer	Annual count of all hospitalizations

Table 4: Attributes for heart attack related events in Alachua County from 2005 to 2023. All data attributes are annual.

2.4 Model Selection

Application of modern ML methods and models, including Deep Learning (DL), has been of keen interest across all domains. As popular as some of these methods have become, DL models are not a panacea and factors such as computational limitations, interpretability, and explainability should be considered in model selection. This is especially important when considering the implications of public decision making from model results. Studies in medical related data have shown promise of DL models in subdomains with tremendous amounts of heterogenous data – however, there are also known limitations and needs for improved methods development and applications, especially

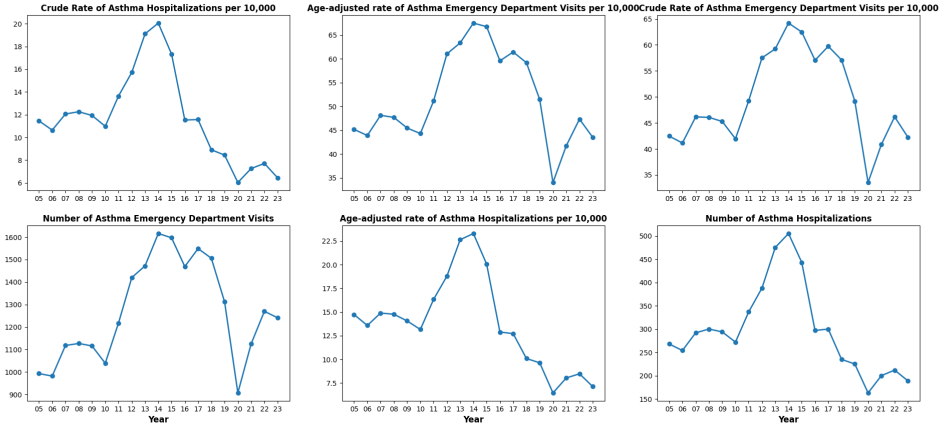

Figure 4: Various heart attack related health data from 2005-2023 in Alachua County, FL. Each plot was made using raw data provided by the Florida Department of Health. The data is given as an annual aggregate. Notice the massive dip across all metrics in 2020 due to the COVID-19 pandemic.

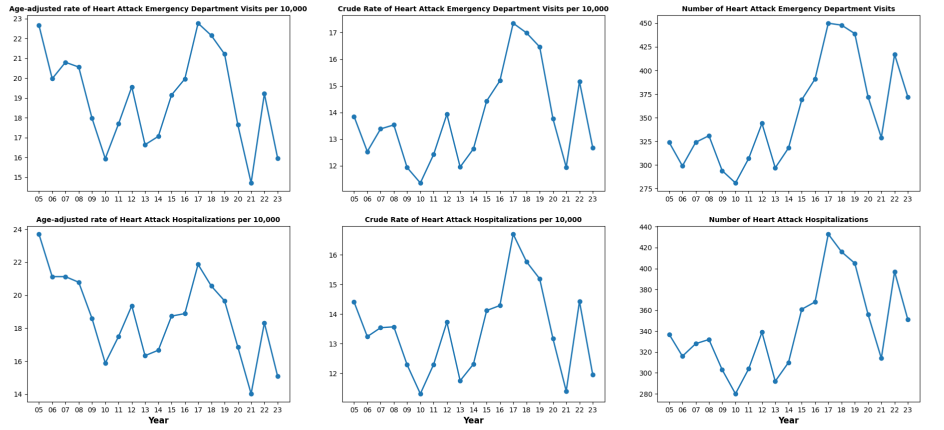

Figure 5: Various heart attack related health data from 2005-2023 in Alachua County, FL. Each plot was made using raw data provided by the Florida Department of Health. The data is given as an annual aggregate. Notice the massive dip across all metrics in 2020 due to the COVID-19 pandemic.

in terms of ease-of-understanding for domain experts and citizen scientists [12]. Considering the actionable goals of the model developed for this study, whatever modeling methodology should error on explainabilty and interpretability. Further, the model should suitable for the given quantity of data at hand.

3 Methodology

The primary goal of this study is to produce a trustworthy and understandable methodology for quantifying wildfire smoke effects on specific health conditions which can inform governing bodies with policy making decisions. Considering the discussion in **SECTION 2.4**, a model that is physics informed with the well understood ML model of linear regression would be the most understandable to those that do not have Data Science expertise but still leverages data collected from publicly available sources.

The physics informed portion of the model makes the assumption of how smoke dissipates from the origin of a fire evenly in 3-D space much as a point source of light does. Ubiquitous throughout physics, the inverse square is applicable to many point sources that radiate a physical quantity such as a force (for example, gravitational) [13]. Also similar to how the universe expands, the particles produced from a wildfire could be thought of as being on the surface of a hypothetical balloon that is being blown up. The particles spread out evenly from each other at the same rate creating a gap between each particle that is larger the further the particles are from the source of the fire. An intial physics based smoke metric would then take the form of the following:

$$\text{smoke metric} = \frac{\text{Total Acres Burned}}{(\text{Fire Distance})^2}$$

Breaking the metric down, the numerator "Total Acres Burned" takes into account the total amount of fuel that turned to smoke while the denominator "Average Fire Distance" takes into account the further a fire is away then the lower the smoke density.

To incorporate health metrics, linear regression of annual AQI against number of hospitalizations of asthma and heart attacks was conducted. The resulting parameter value from the regression algorithm will be used as the exponent (dubbed "health constant") to an empirical factor that is tuned by a multiple of 10 and then multiplied against the smoke metric to adjust the estimated impacts of wildfire smoke in the region. The full metric will then be the following:

$$\text{smoke metric} = \text{Empirical Factor}^{\text{health constant}} * \frac{\text{Total Acres Burned}}{(\text{Fire Distance})^2}$$

To test the results of the final health smoke metric, AQI will be used as ground truth. The rough magnitude and shape of the model compared to AQI will determine the efficacy of the health smoke metric.

4 Findings

The findings from the modified "smoke metric" can be seen in **FIGURE 6**. The first thing to point out is data for the smoke metric extends beyond any health data available. This was made possible by setting the "health constant" equal to zero for years where health data did not exist. Because the "health constant" is an exponential, then it did not set the entire "smoke metric" to zero. Second, **FIGURE 6** shows a normalized version of both the AQI and "smoke metric". This is to try and emphasize the rough scale and overall shape instead of magnitude since the units for

both are unit-less.

A few key findings include the "shifted" look of the "smoke metric" relative to the AQI. The normalization of AQI greatly flattened AQI during the mid 1990s. Referencing **FIGURE 1**, it is clear that the number of acres burned was starting to greatly increase during the same period of time, thus we expect "smoke metric" to be going up during that same time period due to the linear relationship the "smoke metric" has with acres burned. This makes more sense looking at the year 2011 which has the most number of acres burned near Gainesville in the entire dataset. Another finding is that it the chosen health metrics of "Number of Asthma Emergency Department Visits" and "Number of Heart Attack Hospitalizations" were negatively correlated to AQI as seen in **FIGURE 7** and **FIGURE 8**. From this experiment, a value of health constant = -.965 which was obtained by averaging the slope of the regression lines from both linear regressions. So it appears that the more important factor for the "smoke metric" is the how many acres were burned in a fire and how far the fire was away from Gainesville. For the time being, it appears that wildfires are not associated with health issues associated with polluted air in Gainesville and Alachua County.

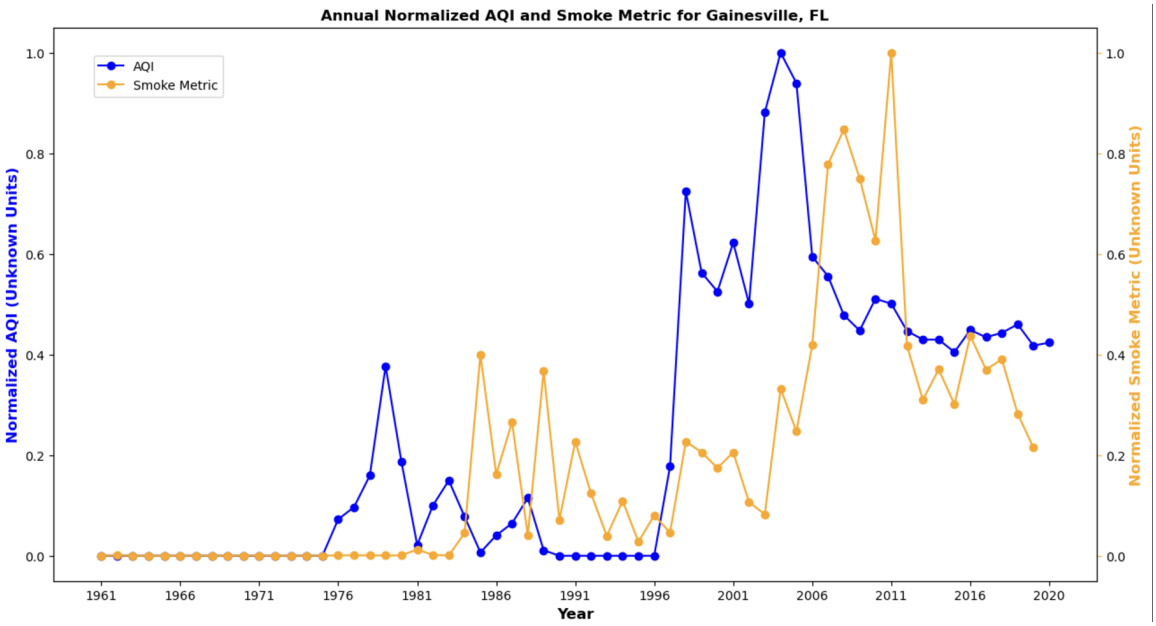

Figure 6: Normalized comparison of "smoke metric" and AQI. The two largely showcase the same information, but challenges still exist to make a more accurate model to produce a "smoke metric". Further, using AQI as a form of "ground truth" is troublesome. Doing so will likely result in a "smoke metric" that is essentially just AQI, which is why normalized versions of these metrics was compared instead of magnitude.

5 Discussion & Implications

There were some surprising conclusions to the analysis and model derived for the "smoke metric". The primary surprise is the lack of a positive correlation between air quality and health problems

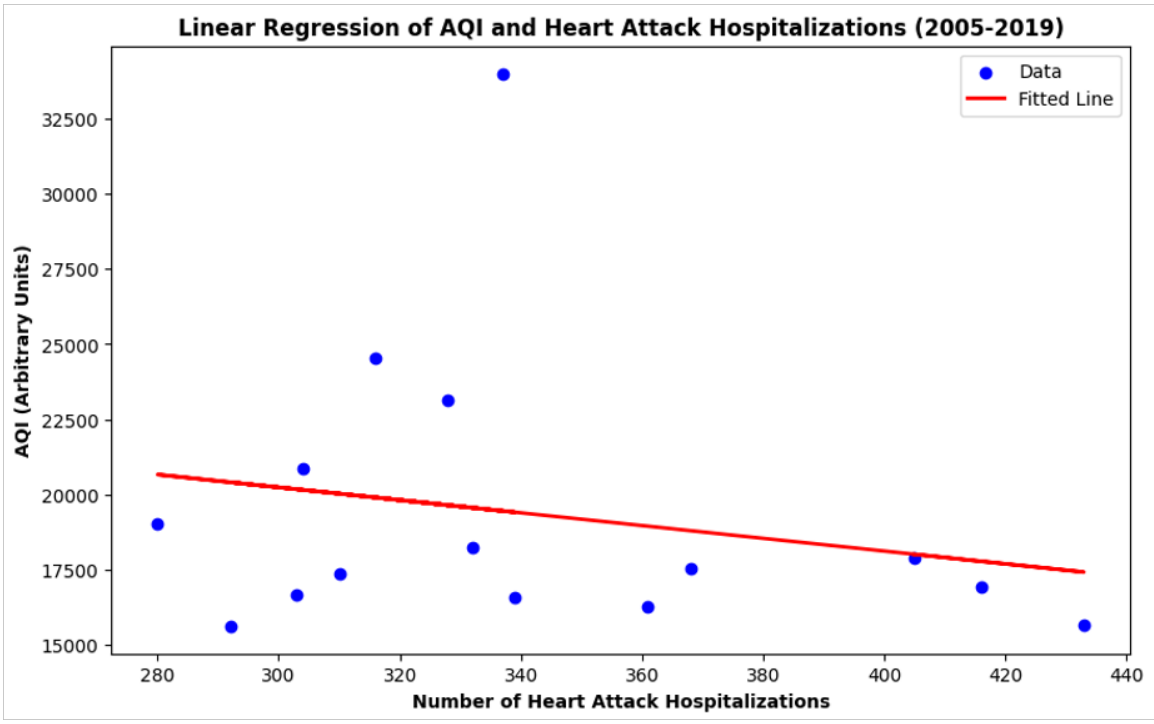

Figure 7: Linear Regression of Annual Heart Attack Hospitalizations to Annual AQI from 2005-2019 in Alachua County, FL. Note that regression spans this temporal range due to health data only going back to 2005. Truncation of data at 2019 due to pandemic affecting data collection in 2020 forward. There is a negative correlation between air quality (AQI) and heart attack hospitalizations.

typically exacerbated by polluted air. Further, it appears that the "smoke metric" is modeling the rough shape of air quality (that is, changes in air quality), but is not providing any obvious information that factors in population health. Forecasting leveraging the smoke metric model and the "health constant" would predict a roughly flat air quality trajectory and a flat rate of hospitalizations associated with asthma and heart attacks.

Ultimately, government officials should take with confidence that, for the time being, wildfire smoke is not necessarily affecting the health of the community in an extreme way. That is not to say this could change in the future, but according to the results of this analysis, it would be best if officials leveraged resources and time to address more pertinent issues. Further, officials should conduct this analysis as part of a regular "health checkup" which includes a revised "health constant" as well as an updated model which includes better data and subject matter expertise. This leads to the final point which is the importance of quality data collection. The more data that can be collected the better – from both perspectives of frequency and granularity. No matter how sophisticated ML models might be, they can only be as good as the data used to train them. To be better informed officials, more data must be collected.

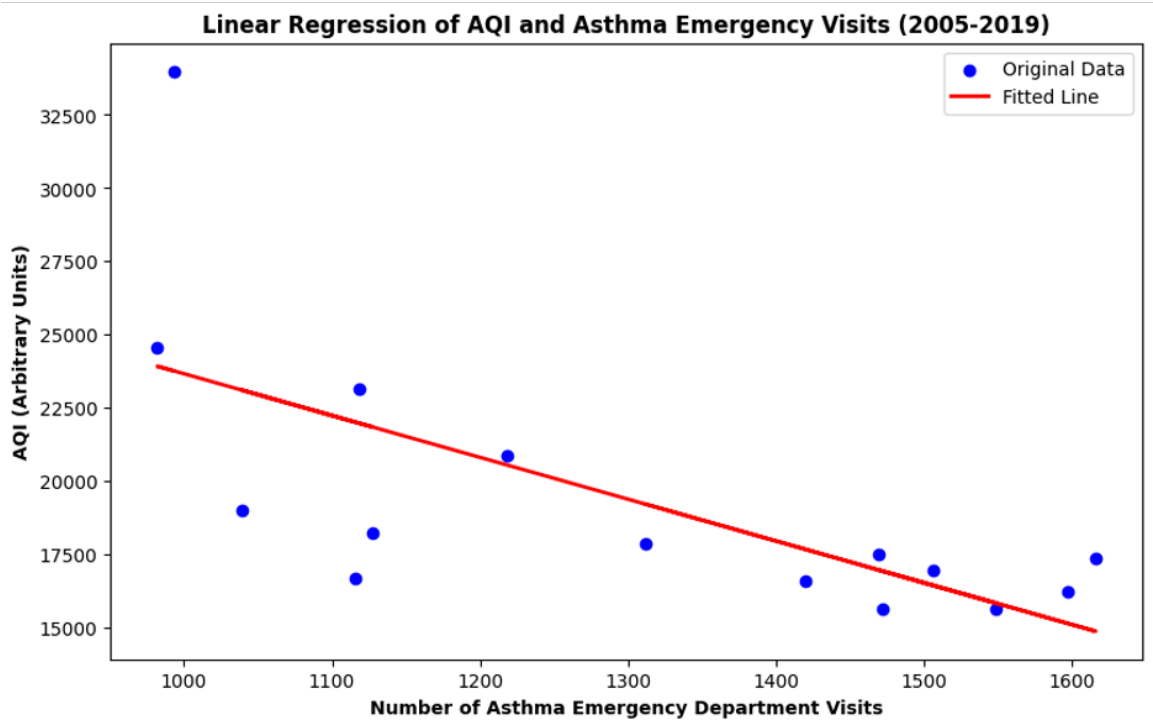

Figure 8: Linear Regression of Annual Asthma Emergency Visits to Annual AQI from 2005-2019 in Alachua County, FL. Note that regression spans this temporal range due to health data only going back to 2005. Truncation of data at 2019 due to pandemic affecting data collection in 2020 forward. There is a negative correlation between air quality (AQI) and asthma emergency visits.

6 Limitations

There were major limitations to much of the data leveraged in the above analysis which was outlined most of **SECTION 2**. The primary limitation that affects all metrics (AQI and "smoke metric") is the wildfire and AQI data collected without incorporating other physical and temporal factors. For example, a major assumption made about smoke dissipation in the above analysis was a constant distribution of smoke particles radiating from the point of the source of a fire. This is an extreme oversimplification and the way smoke flows is most likely to be affected by wind patterns. Having data regarding atmospheric attributes such as the jet stream would produce a far more accurate analysis in reality. From a temporal perspective, the health data was only found to be provided at an annual frequency. This may suffer from averaging out nuances throughout the year. The wildfire data was also only available during the fire season (May to October each year). This is quite limiting and is hard to correlate with data that covers an entire year.

More time should also be spent on potentially more adequate datasets such as those from FEMA or other organizations. As an example, FEMA provides their own fire based risk assessments for regions and that could be incorporated to smooth out some of the temporal and spatial issues with the data used in this analysis.

Finally, a more appropriate ground truth data source must be determined. This should be

developed with a subject matter expert.

7 Conclusion

There is tremendous evidence that shows wildfires are becoming more extreme with their effects globally reaching. Since all communities are possibly vulnerable to the deleterious consequences of the toxic fumes from such events, it is imperative that local communities think the risks associated to these wildfire events. Of all the ways wildfire smoke causes problems, population health is one of the most dire. Health conditions such as asthma are particularly affected by polluted air. Using Gainesville and Alapucha County Florida as a case study, data regarding heart attacks and asthma for the region was collected and used with a regression model against AQI data to determine if wildfire smoke within a 650 mile region of Gainesville was causing increased health risks. The results showed that at the moment, wildfires are not associated with any cardiovascular issues in the region. Government officials should take this information and choose to leverage their resources at a more appropriate venue under the assumption that a similar analysis is regularly conducted as part of a health checkup. Finally, government officials should improve data collection from both temporal and spatial perspectives (city, county, and state levels) to improve model development and forecasting.

8 Data Sources

1. Gainesville and Alachua County demographics [7].
2. Combined wildland fire datasets, US Geological Survey [9].
3. EPA AirNow API for Daily AQI data measurements (https://docs.airnowapi.org/)
4. Florida Department of Health (https://www.floridatracking.com/healthtracking/topic.htm?i=10)

References

1. Cunningham CX, Williamson GJ, and Bowman DMJS. Increasing frequency and intensity of the most extreme wildfires on Earth. Nature Ecology & Evolution 2024;8:1420–5.
2. Sun Y, Forrister A, Kuligowski ED, Lovreglio R, Cova TJ, and Zhao X. Social vulnerabilities and wildfire evacuations: A case study of the 2019 Kincade fire. Safety Science 2024;176:106557.
3. United States Environmental Protection Agency. Why Wildfire Smoke is a Health Concern. 2024. URL: https://www.epa.gov/wildfire-smoke-course/why-wildfire-smoke-health-concern#:~:text=Effects%20to%20Exposure-,Wildfire%20Smoke%3A%20A%20Complex%20Mixture,water%20vapor%2C%20and%20particle%20pollution..
4. United States Environmental Protection Agency. Health Effects Attributed to Wildfire Smoke. 2024. URL: https://www.epa.gov/wildfire-smoke-course/health-effects-attributed-wildfire-smoke.

5. Limaye V and De Alwis D. The Costs of Inaction: The Economic Burden of Fossil Fuels and Climate Change on Health in the U.S. Tech. rep. National Resource Defence Council, 2021. URL: https://www.nrdc.org/resources/costs-inaction-economic-burden-fossil-fuels-and-climate-change-health-us.

6. Byrne B, Liu J, Bowman KW, et al. Carbon emissions from the 2023 Canadian wildfires. Nature 2024;633:835–9.

7. Dataset. Data USA, 2024. URL: https://datausa.io/profile/geo/gainesville-fl/.

8. Brach C and Fraser I. Reducing disparities through culturally competent health care: an analysis of the business case. Qual Manag Health Care 2002;10:15–28.

9. Welty J and Jeffries M. Combined wildland fire datasets for the United States and certain territories, 1800s-Present. U.S. Geological Survey data release. 2021. DOI: 10.5066/P9ZXGFY3. URL: https://doi.org/10.5066/P9ZXGFY3.

10. Alachua County 2024 Air Monitor Site. 2024. URL: https://fldep.dep.state.fl.us/air/flaqs/SiteDetail.asp?SiteID=120013012.

11. Henry S, Ospina MB, Dennett L, and Hicks A. Assessing the Risk of Respiratory-Related Healthcare Visits Associated with Wildfire Smoke Exposure in Children 0-18 Years Old: A Systematic Review. Int J Environ Res Public Health 2021;18.

12. Miotto R, Wang F, Wang S, Jiang X, and Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform 2018;19:1236–46.

13. Jiang ZW, Ding H, and Li EQ. The Inverse-Square Law Force between Vapor-Mediated Droplets. 2021. arXiv: 2110.02338 [physics.flu-dyn]. URL: https://arxiv.org/abs/2110.02338.