Zachary Bowyer

Data 512

12/5/2023

Final report

**Exploring the impact of wildfire smoke on the health care industry of Leavenworth, Kansas**

**Introduction:**

In this work, the author was tasked with conducting a human-centered data analysis project concerned with measuring the impact that wildfire smoke has on a particular city, and from there creating a policy recommendation for the city. The author was specifically assigned to consider the city of Leavenworth, Kansas, and opted to measure the impacts on the healthcare industry. As expected, the primary motivation behind this work is to save Leavenworth money and resources through education. To answer the question of why this analysis is important, research [25] shows that wildfire smoke is dangerous to inhale over both the short and long term. This means that if the city of Leavenworth is unaware of current smoke exposure trends, it may be unable to mitigate additional, avoidable health and economic repercussions due to wildfire smoke. Tangentially, this work may be considered interesting from a technical standpoint because it served as an exercise that demonstrates the difficulty in solving real world problems. These challenges include suboptimal datasets, operating on multiple assumptions, and combining datasets for analysis. The remainder of this report will go into details about related work, methods used, findings, limitations, and finally, the proposed policy recommendation.

**Background/Related work:**

In this work, the first main research question that was asked was a requirement by the course staff for the first portion of work, which will be referred to as the “common analysis” section. Specifically, the question is, “What are the estimated smoke impacts on your assigned city for the last 60 years?” In this case, the assigned city was Leavenworth, Kansas. After the first smoke estimates were determined through the work done in the common analysis section, new research questions were then chosen by the author to be answered in the second portion of work called the “extended analysis” section. The research questions regarding the extended analysis are: “What impacts does wildfire smoke have on the healthcare industry in Leavenworth, Kansas”, “How do specific health conditions relate to emergency department visits”, and “What are the increased risks of emergency department visits for the chosen health conditions when wildfire smoke is involved”. Given the research questions, the next few paragraphs mention and discuss the use of various different research sources, data sources, and algorithms used to attempt to answer the proposed research questions.

In this section, each research or information source used will be discussed. First, in the creation of an annual smoke estimate, source [1] was utilized to determine the volume of smoke produced by burning one kilogram of wood, which according to the source was 87.5 meters squared per kilogram. Second, sources [2], [3], and [4] were used as justification to why specific health conditions such as asthma, cardiovascular disease (CVD), chronic obstructive pulmonary disease (COPD), myocardial infarctions (MI), and strokes were considered in this analysis. Next, the results from [5] were used to estimate that 7.7 percent of people in Leavenworth, Kansas have asthma regardless of the year, the results from [6] were used to estimate that 7.53411 percent of people in Leavenworth, Kansas have CVD regardless of the year, the results from [7] were used to estimate that 6.4 percent of the people in Leavenworth, Kansas have COPD regardless of the year, the results from [8] and [9] were used to estimate that 0.241 percent of the people in Leavenworth, Kansas have experienced a heart attack each year, and the results from [9] and [10] were used to estimate that 0.237 percent of the people in Leavenworth, Kansas have experienced a stroke each year. After that, sources [11], [12], [13], [14] were used to determine that asthma, CVD, COPD, heart attacks, and strokes account for 1.32 percent, 11.4 percent, 12.44 percent, 0.6 percent, and 0.6 percent of all emergency department (ED) visits in Leavenworth, Kansas regardless of the year. Similarly, the number of ED visits per year was estimated to be equivalent to 40 percent of the population of that year due to another research source [15]. Next, sources [16], [17], [16], [18], and [17] were used to determine that the number of ED visits for asthma, CVD, COPD, heart attacks, and strokes could increase by up to 2.2, 5, 1.39, 42, and 11 percent respectively depending on how severe the annual smoke estimate was. Finally, [19] was used to estimate that each ED visit cost 530 dollars.

This work utilized three different sources of data. First, a dataset titled “Combined wildland fire datasets for the United States and certain territories, 1800s-Present” was used. This dataset was published [here](https://www.sciencebase.gov/catalog/item/61aa537dd34eb622f699df81) on 12/08/2021. In short, the dataset is a combination of many smaller or incomplete United States wildfire datasets and stores various metadata for historical fires such as: Area, location, date, polygon coordinate geometry, etc. The second source of data utilized the [Air Quality System (AQS) API](https://aqs.epa.gov/aqsweb/documents/data_api.html), which contained “ambient air sample data collected by state, local, tribal and federal air pollution control agencies”. The contents of data called by the API include recorded station data such as: Particulate type, particulate amount measured, date of collection, location of the station, etc. Third, this work utilized United States census data, which can be found [here](https://data.census.gov/). United States census data is quite vast and can include many different schemas which cannot be fully described here. For this work specifically, population estimates, age estimates, healthcare coverage estimates, and employment by industry estimates were used.

In terms of specific algorithms, two main methods were employed. First, linear regression [20] was utilized to determine if a custom smoke estimate could be predicted using the number of acres a wildfire burned and the distance the wildfire was from Leavenworth, Kansas. The selection of this model was driven by course requirements stipulating the use of a predictive model in the common analysis portion of the work. Second, two autoregressive integrated moving average (ARIMA) models [21] were employed to make projections for both a custom smoke estimate and the number of healthcare, social, and educational workers in Leavenworth up to 2050. ARIMA was chosen specifically because it is a simple model that predicts future time series values based on previous values.

**Methodology:**

For the first part of this section, the first portion of work, called the common analysis will be summarized. This section of work was mainly concerned with answering the research question concerned with finding the estimated smoke impact on Leavenworth, Kansas for the last 60 years. To answer that question, multiple different steps were performed in a Jupyter notebook. First, the combined wildfire dataset was read in using the “geojson” python module due to the fact that data was in JavaScript Object Notation (JSON) format. Next, in order to reduce processing time and to eliminate fires that were not relevant to future analyses, the distance between each wildfire and Leavenworth, Kansas was calculated, where fires that occurred greater than 1250 miles away from Leavenworth were not considered for analysis. This calculation was done by calculating the average distance between the EPSG:4326 [22] coordinates of Leavenworth and each wildfire polygon coordinate. The decision to calculate distance as an average was made because it is simple to calculate and easy to understand. From there an annual smoke estimate was created using wildfire data and the research from [1]. Specifically, the formula used for calculating the smoke estimate for a specific wildfire is: , where A is the area in square meters burned by the wildfire, V is the volume of smoke in meters squared produced by burning one kilogram of wood, W is the amount of wood in kilograms per square meter of forest, and D is the average distance from the wildfire to Leavenworth. To extend that, the annual smoke estimate was simply a sum over all fires for that year, which can be expressed as: , where n is the number of fires that occurred that year and k represents each specific fire. The reason this method was used as a wildfire smoke estimate was due to the fact that the only relevant information available regarding smoke production from the wildfire dataset were the location and areas of the fires. Specifically, it was assumed that smoke dispersion is inversely proportional to the distance of the wildfire to the fifth power, where the fifth power was specifically used to keep the relative scale of the smoke estimate low so that it could be comparable to the AQI. Additionally, area, volume, and mass were used due to the fact that [1] had some relevant information about the volume of smoke produced by burning wood. In terms of the annual smoke estimate, the decision to make it a sum over all wildfires for that year was due to the fact that data was sparse, so using daily or weekly averages may have led to the estimate being near zero at all times. After calculating annual smoke estimates, the next step was to estimate monthly AQIs for Leavenworth, Kansas by utilizing the AQS API. Specifically, AQI was used because it was a requirement of this project made by the course staff. In terms of how the AQI estimate was derived, this work only considered using PM10 and PM2.5 data for the four nearest AQI measurement stations, as no station directly existed in the city. To get the AQI, the nearest four stations were looked at, their PM10 and PM2.5 values were averaged, and the largest average was chosen as the monthly AQI estimate for Leavenworth. The reason the nearest four stations were considered was arbitrary, the reason only PM10 and PM2.5 were chosen was because wildfires were a mentioned source of them according to [22], and the greatest average was chosen because it was simple to implement. Next, a linear regression model and ARIMA model were used to predict and forecast the annual smoke estimate to 2049. The reasons these models were specifically used was due to requirement by the course staff and to maintain simplicity. Finally, four visualizations showing the number of fires that occurred every 50 miles from Leavenworth, the annual acres burned by wildfires, the annual projected smoke estimate, and the AQI estimate were created for the sake of understanding wildfire trends.

The extended analysis portion of work required the answering of a new research question, which was “What impacts does wildfire smoke have on the healthcare industry in Leavenworth, Kansas?” To answer this question, the first step performed was loading in US census data using the Python module ‘Pandas’ due to its simplicity. The specific data that was loaded were the population sizes, old age dependency ratios, total number of people employed, total number of healthcare, education, and social workers, and the percentage of the population without insurance for Leavenworth, Kansas. From there, four basic time-series visualizations were created to show various population trends in Leavenworth. These visualizations showed the population over time, the percentage of working-aged people employed in healthcare, education, and social work over time, the old age dependency ratio over time, and the percentage of the population with no health insurance in Leavenworth. Next, in order to properly estimate future healthcare impacts in Leavenworth, the population of Leavenworth was projected to 2050 using a line equation, which again was chosen for the sake of simplicity. From there, the number of residents in Leavenworth with five specific health conditions were estimated using both the projected population counts and the results from [5], [6], [7], [8], [9], and [10]. Specifically, the population was multiplied by a factor which was described in the literature, where the decision to use a simple multiplier was to again maintain simplicity. Then, the number of estimated emergency department visits were calculated for the five specific health conditions using results from [11], [12], [13], [14], and the estimated visits were visualized to better understand trends. After that, the estimated smoke values from the common analysis section of work were combined with the total number of ED visits due to specific health conditions using simple multipliers from [16], [17], and [18] to get the number of ED visits attributed to specific health conditions when smoke was present. The reason this was done was to help answer the question concerning what the increase in ED visits would be due to wildfire smoke exposure. Finally, visualizations were made concerning the total number of projected ED visits, the extra occurring costs from ED visits, and the number of healthcare workers per projected ED visit to help understand trends, which would be used to make the final policy recommendation. In order to visualize the number of healthcare workers into 2050, another ARIMA [21] model was used for simplicity’s sake.

When it comes to human-centered design, two main considerations influenced how the project was conducted. The first human-centered design component considered is reproducibility, which was especially emphasized by the course staff, where including complete documentation, data descriptions, license files, etc., were deemed essential to having a complete project. With that in mind, this work is committed to being reproducible, which is evident through well-documented Jupyter notebook files, a comprehensive README.md file, and the inclusion of an Anaconda Python development environment file, which ensure that future researchers can quickly replicate the project setup. The second main human-centered aspect of this work pertains to algorithmic transparency. Specifically, this work ensured the use of simple models such as linear regression, ARIMA, and line equations to keep analyses interpretable.

**Findings:**

As previously mentioned, the analyses performed produced many visualizations which can be used to infer general findings that were then used to justify the final policy recommendations made to the city council. Below will discuss each finding and provide a visualization.

The first two findings in this work, which were from the common analysis portion of work, are that both the estimated AQI levels and estimate wildfire smoke exposure levels are increasing in Leavenworth, Kansas. For example, the AQI graph below show how the AQI levels have seen an upward trend from 1990 and that the AQI is starting to shift from the designated ‘Good’ safety level to the ‘Moderate’ safety level. Additionally, the projected smoke estimate graph shows how the smoke levels have drastically increased from 1990 to 2020 and the smoke levels are projected to be greater on average between 2020-2040 than they were between 2000-2020.

A graph showing a safety level

Description automatically generated A graph of smoke

Description automatically generated

The next finding in this work is that the old age dependency ratio, which is “the number of individuals aged 65 and over per 100 people of working age defined as those at ages 20 to 64” [24], has been increasing in Leavenworth, Kansas from 2010 to 2021. After that, using the ARIMA projection along with population data, the next finding was that the population of Leavenworth, Kansas has been rising since at least 2010 and is projected to keep linearly increasing until 2050.

A line graph with numbers and a line

Description automatically generatedA graph with a line

Description automatically generated

From there the total population employed in healthcare, education, or social work was plotted to reveal the trend that the number of healthcare workers is decreasing in recent years.

A graph of a number of people with numbers

Description automatically generated with medium confidence

The next finding utilized the research sources about ED visits related to specific conditions, the number of ED visits that would normally occur in Leavenworth, the increase in ED visits due to wildfire smoke, and the economic cost of an ED visit. Specifically, the research sources were used to find the increase in both the number of ED visits and the economic costs attributable to the increased number of ED visits caused by wildfire smoke. Naturally, the second visualization concludes that a small increase in costs from ED visits can be attributable to wildfire smoke.

A graph with a line and a line

Description automatically generatedA graph of a graph with numbers and a line

Description automatically generated with medium confidence

After that, another visualization was created to show the percentage of population with no health insurance in Leavenworth, Kansas. Specifically, the finding was that from 2018-2021, the number of people without insurance increased.

A line graph with numbers and text

Description automatically generated

Finally, using projected population and health care worker count data, the final finding shows that the ratio of health care, social, and educational workers to the general population is declining in Leavenworth.

A graph of health care workers

Description automatically generated

**Discussion/Implications:**

The findings mentioned above have some important implications that are used to shape the final policy recommendations. First, using the visualizations that show the population over time, old-age dependency ratio over time, and total civilian population employed in healthcare, education, and social work over time, we can conclude that the healthcare industry in Leavenworth, Kansas is going to have an increased relative burden due to the fact that there will be more people and less healthcare workers, which means there will likely be a lower standard of care for patients in the future due to wildfire smoke exposure. Second, the visualizations that show how both the AQI and the wildfire smoke estimates are increasing imply that the amount of wildfire smoke that Leavenworth receives is increasing and will likely continue to increase into 2050. This implication is important because it suggests that the issues relevant to wildfire smoke cannot be ignored forever, as eventually smoke exposure may consistently reach unhealthy levels if the current trends keep up for a long period of time. Finally, the third major implication from the findings is that since long term exposure to wildfire smoke exacerbates health issues, the increasing exposure to smoke will lead to increased Emergency department visits and associated costs. The visualizations relevant to this implication are the projected ED visits graph and the comparison of ED costs with and without smoke graph. This is important because it gives a monetary impact to the impact of wildfire smoke, which is generally universally understood and is often used to compare the impact of differing issues.

With that in mind, the findings and implications discussed in previous sections were used to craft a recommendation. In short, the policy recommendation made was mostly concerned with educating the public and monitoring the situation. For example, since the historical AQI levels of Leavenworth have been in a healthy range for a very long time, both the health and economical costs of wildfire smoke exposure are likely easy to miss because they are small. However, due to the upward trends of wildfire smoke, the issue cannot be entirely ignored. Thus, this work proposes that the city should at the very least continue to monitor the situation and educate the public about wildfire smoke. The goal of increasing public awareness for this specific issue should be sufficient enough to allow future conversations about the topic to be had when the issue becomes more relevant. Finally, the one, potentially radical recommendation made to the city council is to consider subsidizing the healthcare industry in Leavenworth, as the combination of an increasing populous paired with a decreasing number of healthcare workers is alarming regardless of smoke being present or not. However, including smoke into healthcare estimates of course makes the situation worse, and with the rising old age dependency ratio of Leavenworth, the city should want the standard of care to be higher, not lower, which is what the data suggests will be in the future. In terms of a timeline, due to the current low impacts of wildfire smoke, it is safe to say the city council can wait a decade or more before needing to take any drastic action. However, it is suggested that the city council start educating the public as soon as reasonably possible.

To include a short reflection that describes the specific ways that human centered data science principles informed the decision-making process, there are a few things that should be mentioned. First, this work was designed to be as interpretable as possible to accommodate for the fact that the people who are the recipients of the policy recommendation likely do not have a technical background. Second, work was intended to be reproducible so that the city council could appoint another data scientist in the future to easily pick back up where this work ended, which aligns with the notion that the situation should be continuously monitored.

**Limitations/assumptions:**

In this work, several issues were identified. This section is meant to list as many of them as possible. First, when calculating the distance between a wildfire and the city of Leavenworth, the average distance between all of the polygon point coordinates of the wildfire and the coordinates of Leavenworth was used. In reality, the entire area of a wildfire needs to be considered when estimating how smoke travels through the air/atmosphere. What this work is doing is condensing a wildfire down to one point, which is an oversimplification. Next, only fires that were within 1250 miles of Leavenworth were included in the analyses. However, there is no reason to suggest that wildfire smoke could not travel over 1250 miles. After that, the wildfire smoke estimate used has issues because the volume of smoke produced by wood depends on the types of vegetation being burned, more than just vegetation can burn in a wildfire, and the amount of wood on average per square meter in a forest varies on the type of flora present, the season, and other various factors. Additionally, smoke dispersion is much more complicated than what this work proposed, which was a simple proportion to distance. In the real world, atmospheric transport models are a much better way to estimate how smoke disperses. Next, only annual smoke estimates were calculated, instead of monthly, weekly, or daily estimates due to data sparsity. After that, the AQI estimate is not completely correct because there is no AQI measurement station in Leavenworth, which meant four different stations had to be combined to come up with an estimate. Also, the AQI estimate only considered PM10 and PM2.5, and considered them to be equally weighted, which is likely not true when it comes to weighing the impact of particular particulates on health conditions. Next, all of the ARIMA models used for projection were chosen arbitrarily and need to be further refined. For census datasets, this work assumes all of the data is perfectly accurate, which obviously is not true due to the fact that uncertainties are stated in the data files in some cases. Another issue is that the shown population projection used a line equation, and it’s clear that using many different aspects of life in Leavenworth together would allow for a much better understanding of population growth in Leavenworth, which would lead to a better estimate. Another assumption is that the demographic data of the various research papers perfectly matched Leavenworth’s, which is obviously false. It is very clear that in order to conduct better analyses, the specific age, health, and other socioeconomic factors of Leavenworth should be included in the analyses. However, this data may not exist, so this work had to resort to averages and generalizations. Another main limitation is that only five major health conditions were used, but there are likely hundreds of different health conditions that are adversely affected by wildfire smoke. Additionally, the average number of ED visits per population was estimated as well as the cost of an average ED visit. Again, these estimates could be adjusted to the specific sociodemographic data of Leavenworth.

**Conclusion:**

In conclusion, this work was concerned with answering four main questions, which are: “What are the estimated smoke impacts on your assigned city for the last 60 years?”, “What impacts does wildfire smoke have on the healthcare industry in Leavenworth, Kansas”, “How do specific health conditions relate to emergency department visits”, and “What are the increased risks of emergency department visits for chosen health conditions when wildfire smoke is involved?” To answer these questions, the findings of this work suggest that the amount of smoke that Leavenworth receives is increasing, the adverse effects of health conditions such as asthma, CVD, COPD, heart attacks, and strokes increase due to exposure to wildfire smoke, and exposure to wildfire smoke increases the number of ED visits and associated costs to the local economy. This study can be used as a good example of human centered data science because it exhibits key tenants of human-centered design such as reproducibility and interpretability through the use of simple model selection, extensive documentation, and easy to digest visualizations.

**References:**

1. <http://virtual.vtt.fi/virtual/innofirewood/stateoftheart/database/burning/burning.html>
2. <https://www.cdc.gov/asthma/most_recent_national_asthma_data.htm>
3. <https://www.cdc.gov/air/wildfire-smoke/default.htm>
4. <https://www.epa.gov/wildfire-smoke-course/health-effects-attributed-wildfire-smoke>
5. <https://www.cdc.gov/asthma/most_recent_national_asthma_data.htm>
6. <https://professional.heart.org/-/media/PHD-Files-2/Science-News/2/2022-Heart-and-Stroke-Stat-Update/2022-Stat-Update-factsheet-GIobal-Burden-of-Disease.pdf>
7. <https://www.cdc.gov/copd/basics-about.html>
8. <https://www.cdc.gov/heartdisease/facts.htm>
9. <https://www.census.gov/library/stories/2022/12/happy-new-year-2023.html>
10. <https://www.cdc.gov/stroke/facts.htm>
11. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8862306/>
12. <https://www.cdc.gov/nchs/data/nhsr/nhsr174.pdf>
13. <https://journal.copdfoundation.org/jcopdf/id/1103/Characteristics-of-COPD-Patients-Using-United-States-Emergency-Care-or-Hospitalization>
14. <https://hcup-us.ahrq.gov/reports/statbriefs/sb47.pdf>
15. <https://www.cdc.gov/nchs/products/databriefs/db452.htm/>
16. <https://www.healio.com/news/pulmonology/20231107/wildfire-pollution-linked-to-cardiopulmonary-ed-visits-among-vulnerable-populations>
17. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6015400/>
18. <https://newsroom.heart.org/news/wildfires-may-fuel-heart-health-hazards-smoke-exposure-increases-cardiovascular-risks>
19. <https://hcup-us.ahrq.gov/reports/statbriefs/sb268-ED-Costs-2017.jsp>
20. <https://www.ibm.com/topics/linear-regression>
21. <https://www.investopedia.com/terms/a/autoregressive-integrated-moving-average-arima.asp>
22. <https://en.wikipedia.org/wiki/EPSG_Geodetic_Parameter_Dataset>
23. [Air Quality Index - Washington State Department of Ecology](https://ecology.wa.gov/Research-Data/Monitoring-assessment/Air-Quality-Index)
24. <https://www.oecd-ilibrary.org/social-issues-migration-health/old-age-dependency-ratio/indicator/english_e0255c98-en>
25. <https://www.health.ny.gov/environmental/outdoors/air/smoke_from_fire.htm>

**Data sources:**

**Combined wildland fire datasets:**

[Combined wildland fire datasets for the United States and certain territories, 1800s-Present (combined wildland fire polygons) - ScienceBase-Catalog](https://www.sciencebase.gov/catalog/item/61aa537dd34eb622f699df81)

**AQS API:**

<https://aqs.epa.gov/aqsweb/documents/data_api.html>

**US Census Data:**

<https://data.census.gov>

<https://data.census.gov/table/ACSST5Y2021.S2407?q=Leavenworth%20city,%20Kansas&t=Class%20of%20Worker>

<https://data.census.gov/table/ACSSE2022.K201803?q=Leavenworth%20city,%20Kansas&t=Disability> -

<https://data.census.gov/table/ACSSE2022.K201801?q=Leavenworth%20city,%20Kansas&t=Health> -

<https://data.census.gov/table/ACSST5Y2021.S0101?q=Leavenworth%20city,%20Kansas&t=Age%20and%20Sex>

<https://data.census.gov/table/ACSST5Y2022.S2701?q=Leavenworth%20city,%20Kansas&t=Health%20Insurance>