Sarah Kilpatrick
Part 2 - Extension Plan

Motivation Statement

Estimating smoke impact on respiratory illness-caused mortality in Fresno, CA.

The linkages between increased particulate pollution due to wildfires and increased mortality in California are well-documented (Chen et al., Gwon et. al). This analysis will use a predictive estimate for the impact of smoke caused by wildfires within 650 miles of Fresno. Data on mortality caused by respiratory illness from the California Department of Health and Human Services will then be added into the predictive estimate to quantify future impacts of smoke from a public health perspective. This project will present a refinement of a smoke estimate predictive model by using non-periodic time series forecasting techniques such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) (Tensorflow). We hope to generate correlations between the refined predictive estimate for smoke impact and changes in Fresno's mortality rates caused by respiratory conditions. In turn, calling for the future study of causal connections between the two variables.

Data To Be Used

Data: Death Profiles by Zip Code (1989-2022)

Source: Hosted on the California Health and Human Services Open Data Portal, provided by the California Department of Public Health

Summary:
"This dataset contains counts of deaths for California residents by ZIP Code based on information entered on death certificates. Final counts are derived from static data and include out-of-state deaths of California residents. The data tables include deaths of residents of California by ZIP Code of residence (by residence). The data are reported as totals, as well as stratified by age and gender. Deaths due to all causes (ALL) and selected underlying cause of death categories are provided. See temporal coverage for more information on which combinations are available for which years."

Dataset:
CSV Files of:

- 2019-2022 Final Deaths by Year by ZIP Code
    - "This data table reports the annual number of deaths of California residents by the ZIP Code of residence regardless of where the death occurred (by residence). The cause of death categories are based solely on the underlying cause of death as coded by the International Classification of Diseases." (CalHHS)

- 2009-2018 Final Deaths by Year by ZIP CodeCSV Popular
- 1999-2008 Final Deaths by Year by ZIP CodeCSV Popular
- 1989-1998 Final Deaths by Year by ZIP CodeCSV Popular
- Data Dictionary - Deaths by ZIP CodeCSV Popular

License/Terms of Use: Open for public use contingent on following the CalHHS terms of use

---

Data: ICD-9 and ICD-10 Insurance Codes for Respiratory Illnesses

Source: Wikipedia

Summary: Unified and standardized codes attributed to "diseases, signs and symptoms, abnormal findings, complaints, social circumstances, and external causes of injury or diseases", used for medical or insurance purposes (Wikipedia).

Dataset: ICD-9 [460-519] and ICD-10 [J00-J99] codes

License/Terms of Use: Open for public use under the Creative Commons Attribution-ShareAlike 4.0 International License, additional terms may apply.

Model To Be Used

Model: The initial predictive model was a Holt-Winters Exponential Smoothing Time-Series Forecasting Model. This model is good for periodic and noisy data, but has limitations in predicting non-periodic outcomes. Therefore, this extension will seek to refine this predictive model by switching to predicting smoke estimates with Convolutional Neural Networks (CNNs) provided by Tensorflow. Time will also be dedicated to explore the success of Recurrent Neural Networks (RNNs) on this question.

The foundational theory and practice of these networks comes from a Jupyter Notebook named Time Series Forecasting, by an unnamed team of Tensorflow authors. This model is distributed under the Apache License, Version 2.0 (Tensorflow).

Smoke Estimate Model Adaptations: This extension will adapt the initial smoke estimate by creating a smoke IMPACT estimate on mortality. This will be a new variable measuring not smoke, but smoke impact on human-centered mortality rates from respiratory illnesses in Fresno, CA.

The correlation between increases in smoke estimate and increases in mortality will inherently carry some lag time. Therefore, a large part of updating my model will be finding the lag time, in

years, which maximizes the correlation between my initial smoke estimate and Fresno's mortality rate due to respiratory diseases.

An important subfield of mortality studies in demography is the study of excess mortality. This process involves proving causality: proving that some political regime, pandemic, or other large-scale change in a population caused an excess of deaths, compared to if that large-scale change hadn't occurred. This requires computing the counterfactual mortality rate, the rate of deaths in a population if that large-scale change hadn't occurred. The reason excess mortality will not be studied in this extension is that it is beyond the scope of this assignment; this is a human-centered data science assignment, not a demography paper. Ultimately, this extension will aim to provide evidence suggesting that research establishing such causality would be scientifically impactful in the California Central Valley.

Unknowns and Dependencies

Due to the techniques used and author's understanding of demographic methods, this analysis will be inherently restricted to building a refined smoke impact score informed by correlations with mortality rates. This integration of morality rates will also seek to identify the approximate lag in years between high-wildfire smoke years and spikes in mortality. Furthermore, there exist inherent limitations to this extension's ability to optimally tune the CNN, due to the timeline of this project. All forecasted smoke impact estimates will therefore be documented as such.

The data contains all relevant zip codes for Fresno, CA, but not all years from the USGS Wildfire dataset. This extension will be dependent on a narrower timeframe, from 1989 - 2019.

Furthermore, this analysis will be informed by hospital-reported mortality rates, as this is the primary source for the California Department of Public Health. This extension has a tradeoff between the availability of data capturing all of Fresno's deaths, informing their death rates, and the precision of the smoke estimate's impact on mortality; it will be assumed that not all of Fresno's deaths during this time period have been fully recorded.

Sarah Kilpatrick
Part 2 - Extension Plan

Timeline to Completion

Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
27	**28**	**29**	**30** Collect Data	**31**	**1**	**2** Build Initial Model
3	**4** *Class*	**5**	**6** Improve Mortality Rate Integration	**7**	**8**	**9**
10	**11** *Class* Finalize Mortality Rate-Informed Smoke Impact Score	**12**	**13**	**14**	**15** First Iteration Model Tuning	**16** Document Process
17	**18** *Class*	**19** Second Iteration Model Tuning	**20** Document Process	**21**	**22** Third Iteration Model Tuning	**23** Document Process
24 Final Model Tuning Done	**25** *Class*	**26** Visualizations Complete	**27** Project Part 3 - Presentation	**28**	**29**	**30** Document Process
Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
1 Finalize Documentation	**2** *Class* Project Part 5 - Presentation Feedback	**3**	**4** Project Part 4 - Final Repo	**5**	**6**	**7**

Sarah Kilpatrick
Part 2 - Extension Plan

Bibliography

California Department of Public Health. California Comprehensive Master Death File (Static), 2014-2022. Compiled by Center for Health Statistics and Informatics.

California Department of Public Health. Death Statistical Master File (Static), 1989-2013. Compiled by Center for Health Statistics and Informatics.

Chen, Gongbo, et al. "Mortality risk attributable to wildfire-related PM2·5 pollution: A global time series study in 749 locations." The Lancet Planetary Health, vol. 5, no. 9, Sept. 2021, https://doi.org/10.1016/s2542-5196(21)00200-x.

Gwon, Y.; Ji, Y.; Bell, J.E.; Abadi, A.M.; Berman, J.D.; Rau, A.; Leeper, R.D.; Rennie, J. The Association between Drought Exposure and Respiratory-Related Mortality in the United States from 2000 to 2018. Int. J. Environ. Res. Public Health 2023, 20, 6076. https://doi.org/10.3390/ ijerph20126076

The Tensorflow Authors, (2024) Time Series Forecasting. Tensorflow. https://www.tensorflow.org/tutorials/structured_data/time_series

United States Mortality DataBase. University of California, Berkeley (USA). Available at www.usa.mortality.org (data downloaded on 31 Oct. 2024).