Breakout Session 3: Track B

Approaches for AI / ML Readiness for Wildfire Exposures

Dr. Joan Casey

Assistant Professor of Environmental and Occupational Health Sciences, University of Washington School of Public Health

Ms. Michelle Audirac

Senior Data Scientist, Harvard University

Approaches for AI/ML Readiness for Wildfire Exposure and Health Analysis

Supplement Title: *Approaches for AI/ML Readiness for Wildfire Exposure (*RF1AG071024*)*

Speakers: Joan A. Casey (PI, University of Washington), Michelle Audirac (Senior Programmer, Harvard)

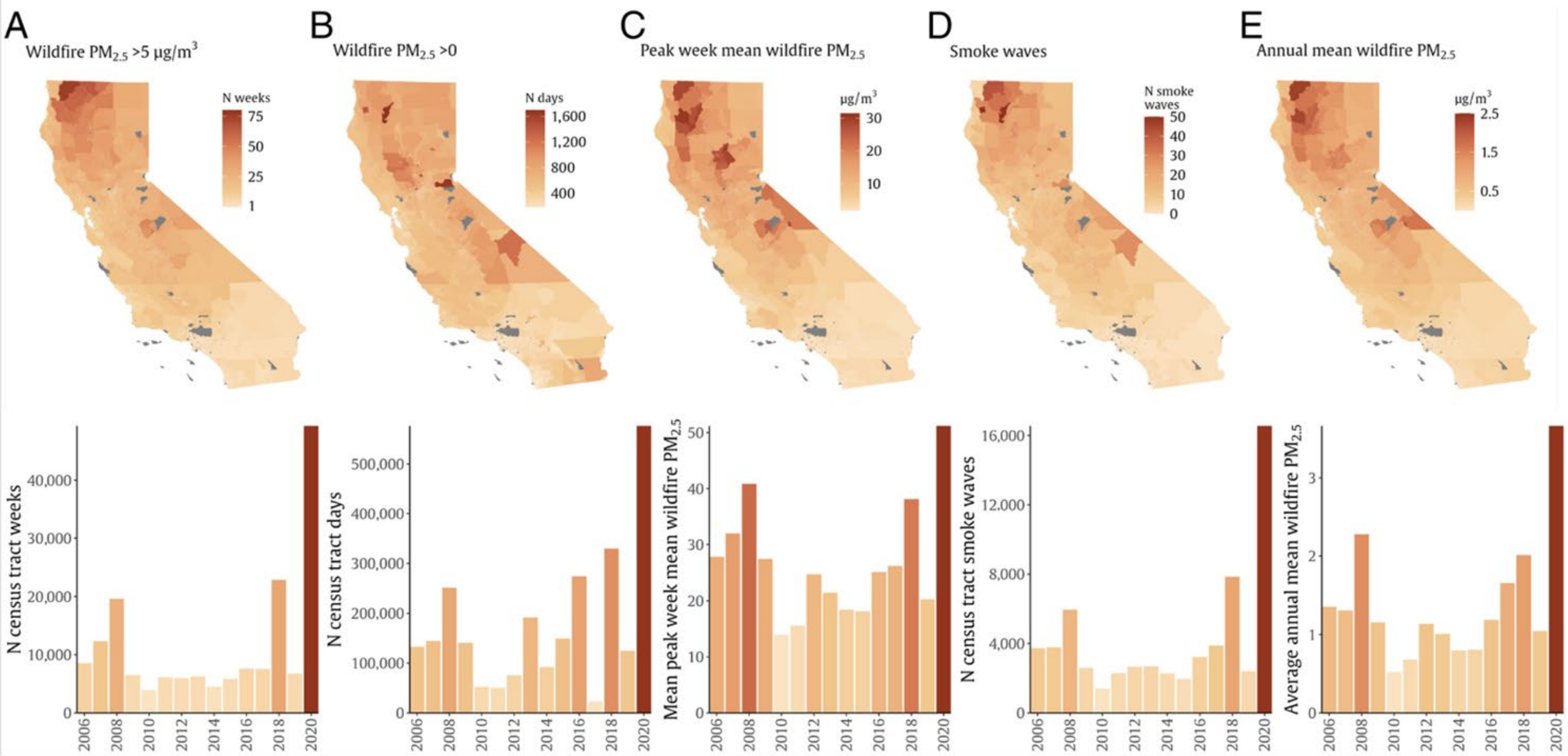
Summary of parent grant: Short and long-term consequences of wildfires for Alzheimer's disease and related dementias (RF1AG071024, PI: Casey)

Aim 1: Estimate the risk of mild cognitive impairment (MCI) and Alzheimer’s disease (AD) and AD-related dementias (ADRD) associated with wildfire $PM_{2.5}$ exposure

Aim 2: Identify individual and area-level susceptibility factors that exacerbate the association between wildfire $PM_{2.5}$ exposure and MCI and AD/ADRD

Aim 3: Estimate the risk of MCI and AD/ADRD that is associated with living in close proximity to the site of a wildfire disaster and the extent to which specific subgroups differ with respect to these outcomes

Example of wildfire $PM_{2.5}$ output



Casey et al. PNAS 2024

Motivation

- The data sources needed to do effective wildfire analysis are disparate, not very accessible, and unfriendly to AI/ML applications
    - These data often do not follow FAIR principles
- Although the data is rich and publicly available through US agencies, acquiring it and preparing it for analysis presents a significant investment by any researcher

Goals

- Our goal is to develop **reproducible pipelines** that can be harnessed by others
- Leverage **Harvard Dataverse**, a generalist repository, and **GitHub**, to ensure that our data is shared according to the latest research dissemination standards (such as FAIR and TRUST principles)

Challenges: working with gridded/raster data for linkable and inter-operable manipulation

- **Format Diversity** There's a wide range of file formats used to store raster data (e.g., TIFF, NetCDF, HDF, and more), each with its own specifications and intended use cases.
- **Data Size** Raster data, especially high-resolution imagery or extensive time series datasets, can be extremely large, making storage, transmission, and processing resource-intensive.
- **Spatial Reference Systems** Raster data can be represented in various spatial reference systems. Discrepancies between these systems can lead to misalignments when integrating data from different sources.
- **Scalability of Processing Tools** As the volume of raster data grows, existing processing tools may struggle to handle them efficiently.
- **Data Quality and Uncertainty** The quality of raster data can vary significantly depending on the source and collection methods, affecting its suitability for certain applications.

Challenges: aggregating gridded/raster data at a specified geographic level for health studies across years

Raster data inherently represent **continuous space**, while health data (MCI, AD/ADRD and other health outcomes) often correspond to residence at **discrete administrative units** (like counties or zip codes).

- **Spatial alignment** using existing aggregation solutions within gis-packages in R and Python
  - **failure/crash or excessively long processing times** is often encountered when dealing with very high-resolution raster data and/or intricate polygon shapes
  - **Missing data handling**
- **Temporal handling**
  - **changes in administrative units** adds additional complication for aggregations at various points in time

Challenges: fetching census data at a specified geographic level for health studies across years

- **Vast amount of surveys** U.S. Census Bureau data involves navigating a complex landscape of information collected through various surveys that takes time to understand
- **Vast amount of variables** Surveys such as the American Community Survey renders up to 60,000 variables
- **API's variable and time coverage** existing census packages and APIs fetch data for different subsets of variables and years, the ease-of-use of each package varies
- **Surveys geographic level coverage** not all surveys cover all geographic levels
- **Harmonization of variable codes across years** census variable codes change over time, complicating data comparability and usage across years
- **Changes in administrative units** Changes in geographic boundaries over time, such as those due to redistricting or the incorporation of new municipalities

Project stages

Spatial aggregations

- Assessing the performance of multiple GIS-packages in R and Python
- Determining the most appropriate GIS-object type to perform fast aggregations
- Understanding the differences between different raw gridded-datasets
- Identifying sources of GIS-files containing administrative boundaries across time (and their differences)
- Harmonized geographic ID across years

Census data

- Investigating and understanding key differences between US Bureau Census surveys and APIs
- Identifying key features such as time and spatial coverage of surveys
- Performing NLP analysis to simplify the identification of “variable themes” clusters
- Documenting variable code changes across years for time series fetching

Our unifying pipeline approach: **data-as-code containerized tasks**

- Identifying commonly used Data Science tooling for pipelines
  - workflow languages -> Snakemake, cwl
  - configuration parsers -> Hydra
  - container builders -> Docker
- Creating **Github repositories** for easy-to-use reproducible dataset generation
- Sharing the datasets in **Dataverse** within a collection that has metadata specific for environmental health studies

Finalized products

Climate types

Raw source
Köppen-Geiger climate classification from Beck et al
Github repository
https://github.com/NSAPH-Data-Processing/climate_types_raster2polygon
Dataverse doi
TBD

Satellite $PM_{2.5}$

Raw source
Atmospheric Composition Analysis Group V5.GL.04 model
Github repository
https://github.com/NSAPH-Data-Processing/satellite_pm25_raster2polygon
Dataverse doi
TBD

Census series

Raw source
api.census.gov
Github repository
https://github.com/NSAPH-Data-Processing/census_series
Dataverse doi
https://doi.org/10.7910/DVN/N3IEXS

Gridmet

Raw source
Gridmet from climatology lab
Github repository
https://github.com/NSAPH-Data-Platform/nsaph-gridmet
Dataverse doi
TBD

Finalized products, continued

Zip code smoke aggregations

Raw source
https://doi.org/10.7910/DVN/DJVMTV from Childs et al
Github repository
https://github.com/NSAPH-Data-Processing/census_series
Dataverse doi
https://doi.org/10.7910/DVN/VHNJBD

Zip2zcta x-year x-walk

Raw source
UDS mapper
Github repository
https://github.com/NSAPH-Data-Processing/zip2zcta_master_xwalk
Dataverse doi
https://doi.org/10.7910/DVN/HYNJSZ

$PM_{2.5}$ components

Raw source
Atmospheric Composition Analysis Group V4.NA.03 model
Github repository
https://github.com/NSAPH-Data-Platform/nsaph-gridmet
Dataverse doi
TBD

Future Work

- Continue to deposit and share data on Dataverse
- Currently in the process of conducting analysis using the processed AI/ML ready data to accomplish aims of the parent R01

Acknowledgements:
NIH/NIA: RF1AG071024