SCIENTIFIC DATA

Overleaf

Machine learning-derived daily wildfire and non-wildfire $PM_{2.5}$ concentration estimates in the western US, 2008-2018

Melissa M Maestas[1], Colleen Reid[2*], Ellen Considine[1], Gina Li[1]

July 3, 2019

1. Cooperative Institute for Research in Environmental Sciences, Earth Lab and 2. Department of Geography, University of Colorado Boulder, Boulder, Colorado, USA *corresponding author(s): Colleen Reid (Colleen.Reid@Colorado.edu)

Abstract

[Insert abstract text here.]

Background & Summary

Fine particulate matter (PM2.5) air pollution is increasingly associated with numerous adverse health outcomes including mortality ([?], respiratory and cardiovascular morbidity [?, ?], impacts on adverse birth outcomes [?], lung cancer [?], and more. Although PM2.5 concentrations have been declining in many parts of the United States due to policies to limit emissions of air pollutants [?], PM2.5 levels have been increasing in parts of the northwestern US [?] and this increase has been shown to be associated with wildfire smoke [?, ?], which can cause PM2.5 levels to exceed the EPA's NAAQS many times-fold in many places for periods of time [?].

Estimation of PM2.5 concentrations for health studies traditionally used monitoring data, most of which are placed and managed for regulatory purposes. In the US, the EPA's Federal Reference Monitors (FRMs) often only measure every third or sixth day and do not provide enough spatial coverage to get a good estimate of the air pollution exposures where every person lives. In fact most US counties do not contain a regulatory air pollution monitor [?]. Using solely monitoring data in health studies leads to exposure misclassification, which often drives effect estimates of the association between air pollution and health towards the null [?].

To improve population exposure assessment of PM2.5, epidemiological researchers in the past two decades have increasingly been using methods to estimate PM2.5 exposures in the temporal and spatial gaps between regulatory monitors using a variety of methods including interpolation, satellite data, or

air pollution models. Each of these methods had its own benefits and limitations, however. Increasingly, researchers are using hybrid modelling techniques to statistically "blend" information from a combination of data sources to better estimate PM2.5 in space and time. Various methods of blending have been used including spatiotemporal regression kriging (e.g., [**?**], geographically-weighted regression (e.g., [**?**], and increasingly machine learning methods (e.g., [**?**, **?**, **?**]. These machine learning methods train large auxiliary datasets, often including satellite AOD, meteorological data, chemical transport model output, and land cover and land use data to provide optimal estimates of PM2.5 where people breathe.

These models have been implemented in various locations throughout the world from city-level, regional-level, to country-wide [**?**]. The larger the spatial and temporal domain of the modeling, the more helpful the data can be for use in longitudinal epidemiological studies that cover long time periods and large spatial domains. Within the US, [**?**] and [**?**] have separately used machine learning algorithms to create fine-resolution daily PM2.5 estimates for the entire US. Both of these models, however, have performed poorly in the western US compared to the rest of the country. Given the increasing trends in PM2.5 in parts of the western US and the importance of wildfires as a source of PM2.5 there, it is important to have a model that is specifically focused on this region to capture the variability in space and time in this region.

The dataset we describe here improves upon previous daily estimates of PM2.5 from machine learning models in the following ways: (1) use of a more extensive monitoring station network than used in previous models that captures more spatial locations and also proximity to wildfires, a key driver of PM2.5 in the western US, (2) use of an ensemble of machine learning algorithms which have been shown to improve model performance (citation), (3) better temporal prediction through the use of a nonlinear function (cosine) on day of year, (4) allowing for different prediction models for fire-affected and non-fire affected days to better capture and predict exposures affected by wildfires, and (5) incorporation of errors in prediction back into daily estimates through spatial interpolation. We are making these data available as daily estimates of PM2.5 exposures averaged over ZIP codes and counties for future use in epidemiological investigations of the health impacts of air pollution exposure in the western US, where wildfires are a significant contributor to PM2.5 concentrations.

[insert Figure 1: monitor locations (points) and state boundaries]

[insert Table 1: list variables]

Methods

Study Area

Our study area includes 11 western US states: Arizona, California, Colorado, Idaho, Montana, Nevada, New Mexico, Oregon, Utah, Washington, and Wyoming.[What other descriptions should we put? - square kilometers? climate zones? topog-

raphy? other?]

Example of citation: [1]

$PM_{2.5}$ Measurements

[Write short description of each $PM_{2.5}$ data source.]

We downloaded the 2008-2018 pre-generated daily summary files for $PM_{2.5}$ (88101 and 88502 parameter codes) (https://aqs.epa.gov/aqsweb/airdata/download_files.html#Daily) as well as the spreadsheet listing all AQS monitors with datums (https://aqs.epa.gov/aqsweb/airdata/aqs_monitors.zip) from the United States Environmental Protection Agency (US EPA).

All available $PM_{2.5}$ data in the Fire Cache Smoke Monitor Archive (https://wrcc.dri.edu/cgi-bin/smoke.pl) was downloaded for the years 2008-2018.

$PM_{2.5}$ data from the Uintah Basin, Utah were provided by Seth Lyman at Utah State University (personal communication).

$PM_{2.5}$ data from the Persistent Cold Air Pool Study (PCAPS) [2] conducted in the Salt Lake Valley, Utah in January–February, 2011 were provided by Dr. Geoff Silcox in Chemical Engineering at the University of Utah

Predictors

[Write short description of each predictor data set and refer to Table 1]

Machine learning modelling and mapping

[Write description of ML modelling approach]

Code availability

[Insert brief description of how to access code on GitHub.] The code was written and annotated in R [version number] and Python [version number] and is available from GitHub [doi citation link]. The key package for implementing the ML model was [caretEnsemble?].

Data Records

All data are freely available from [repository name, data doi citation]. We provide ... [reference Figure 2]

[insert Figure 2: choropleths at zip code level - 4-panel: a) highest year PM2.5, Aug or Sept, b) highest year PM2.5, Jan/Feb, c) lowest year PM2.5, Aug or Sept, d) lowest year PM2.5, Jan/Feb.]

[Insert Table 3: list of files]

Technical Validation

[Write description of goodness of fit methods/metrics - out-of-bag data, RMSE, R2, models run on subsets of data, etc.]

[Insert Figure 4: a) out-of bag observed PM2.5 vs predicted, b) full model observed PM2.5 vs predicted, c-j) various subsets of data - oob and full model plots (see figure 5 of example paper)]

[Write discussion about variable importance, possibly referring to the suggested figure of variable importance panel figure. Could make an observation or two about the complexity of the variables, e.g., PM2.5 can be highest at highest and lowest temperatures (summer fire season and winter inversions), etc.]

[Thoughts - insert figure of predicted PM2.5 vs predictor variable for the 8 (or so) most important variables (panel figure)]

Thoughts: compare to PM2.5. Concerned comparing to HMS will take too long?

Usage Notes

[Write brief description of things the provided code can be adapted to do, such as making plots of specific years, use in health/pollution studies.]

Acknowledgements

[Write acknowledgements text here.]

Author contributions

[Write brief description of contribution from each author.]

Competing interests

The authors declare not competing interests.

Figures and figures legends

[All figures go here and are referred to in the text]

Tables

[All tables go here and are referred to in the text - read template text for tables]

SCIENTIFIC DATA

Overleaf

References

[1] Y. Liu, J. A. Sarnat, V. Kilaru, D. J. Jacob, and P. Koutrakis. Estimating ground-level PM2.5 in the eastern United States using satellite remote sensing. *Environ Sci Technol*, 39(9):3269–78, May 2005.

[2] Geoffrey D. Silcox, Kerry E. Kelly, Erik T. Crosman, C. David Whiteman, and Bruce L. Allen. Wintertime pm2.5 concentrations during persistent, multi-day cold-air pools in a mountain valley. *Atmospheric Environment*, 46:17 – 24, 2012.