Extensible database of validated biomass smoke events for health research

Ivan C. Hanigan * [1] Fay H. Johnston [2] Geoffrey G. Morgan [3] Grant J. Williamson [2] Farhad Salimi [2] Sarah B. Henderson [4]

[1]National Centre for Epidemiology and Population Health, Australian National University

[2]Menzies School of Population Health, University of Tasmania

[3]University Centre for Rural Health, University of Sydney

[4]School of Population and Public Health, University of British Columbia

* *Corresponding author: ivan.hanigan@anu.edu.au*

2016-03-01

DRAFT

Abstract

Background: Epidemiological studies of the health effects of biomass smoke events (such as bushfires or wood-heater smoke spikes due to inversion layers) have been hindered by the lack of available datasets that explicitly list the locations and dates of pollution events from these sources. Extreme air pollution events may also be caused by dust storms, fossil fuel induced smog events or factory fires, and so validation is necessary to ensure the events are from biomass sources. This paper presents an extensible database developed by the authors to identify historical spikes in air pollution and to evaluate whether they were caused by vegetation fire smoke or by other possible sources. The ability for this database to be extended by other researchers means that new events can be added, and new information for already identified events can be described. These methods provide a systematic framework for retrospective identification of the air quality impacts of biomass smoke. In this paper, we describe the database and data aquisition methods, as well as analytical considerations when validating historical events using a range of reference types.

Methods: Several major urban centers and smaller regional towns in the Australian states of New South Wales, Western Australia, and Tasmania were selected as they are intermittently affected by extreme episodes of vegetation fire smoke. Air pollution data was collated and missing values were imputed. Extreme values were identified and a range of sources of reference information were assessed for each date. Reference types included online newspaper archives, government and research agency records, satellite imagery and a Dust Storms database.

Results: This dataset contains validated events of extreme biomass smoke pollution across Australian cities. The authors have previously demonstrated the utility of this database in analyses of hospital admissions and mortality data for these locations to quantify the pollution-related health effects of these events.

Conclusions: The database was created using open source software and this makes the prospect for future extensions to the database possible. This is because if other scientists notice an ommision or error in these data they can offer an amendment. We believe that this will improve the database and benefit the whole biomass smoke health research community.

Epidemiological studies of outdoor air pollution

For decades, researchers have studied the public health impacts of ambient outdoor air pollution, particularly from the effects of particulate and gaseous pollutants, especially associated with the combustion of coal, petroleum and biomass used for cooking (Pope & Dockery 2006). Far fewer studies have examined the effect of intermittent smoke from biomass burning, such as that which occurs in bushfires, or from woodsmoke trapped by inversion layers during winter months as wood is burned for heating (Naeher *et al.* 2007).

There is a gap in the epidemiological literature of health effects from ambient outdoor air pollution relating to smoke from biomass burning such as that from bushfires or woodsmoke from heating. Most literature available that focuses on biomass smoke health impacts looks at indoor pollution from cooking (Smith 1993). Particles (and perhaps noxious gases) in outdoor pollution from biomass smoke might directly influence the respiratory system through their inhalation and lodgement in the lungs. Particles may then affect the cardiovascular system after their entry into the circulatory system from the alveolae. Indirect effects on mental health and wellbeing are also plausible.

Epidemiological studies that investigate the relationship between health and air pollution exposures have primarily used time-series methods that study variations of some health outcomes such as deaths or hospitalisations from specific disease groups (Peng & Dominici 2008). These outcomes are usually monitored by day across whole cities, and relationships with atmospheric variables estimated in regression models. These typically focus on daily levels of ambient air pollution measured by a network of monitoring sites scattered across a city, time matched to the health outcomes on the same day or a few days after. In general biomass smoke forms only a small part of the mixture of pollutants in the air, however when a bushfire or inversion layer event occurs there is often a concomitant spike in the pollution levels primarily composed of biomass smoke. There is then the ability to study statistical associations between these pollution spikes and the health outcomes around those days. Anomalous levels of pollution can be arbitrarily defined using a threshold such as the 95th percentile and these might be assumed to be biomass smoke days, however there are other events that might cause such as spike such as dust storms, factory fires or even sea salt being driven by certain wind events. There is a need then to validate the dates on which events are ascribed in any correlational study of pollution spikes and health that claims the high levels are due to biomass smoke.

The development of this biomass smoke events database

This open and extensible database was developed by the authors to identify historical spikes in particulate matter concentrations and to evaluate whether they were caused by vegetation fire smoke or by other means. A summary of the protocol for developing this database and a summary of the data we collated is published already as a descriptive paper (Johnston *et al.* 2011). This paper describes how the database has been extended to be able to be distributed in an open, extensible format that allows the research community to add to the history of these events.

System design

The system is described in Figure 1. The procedure starts with the online database and web interface that is maintained by the Data Manager (DM) in our group. The DM extracts a snapshot of the database

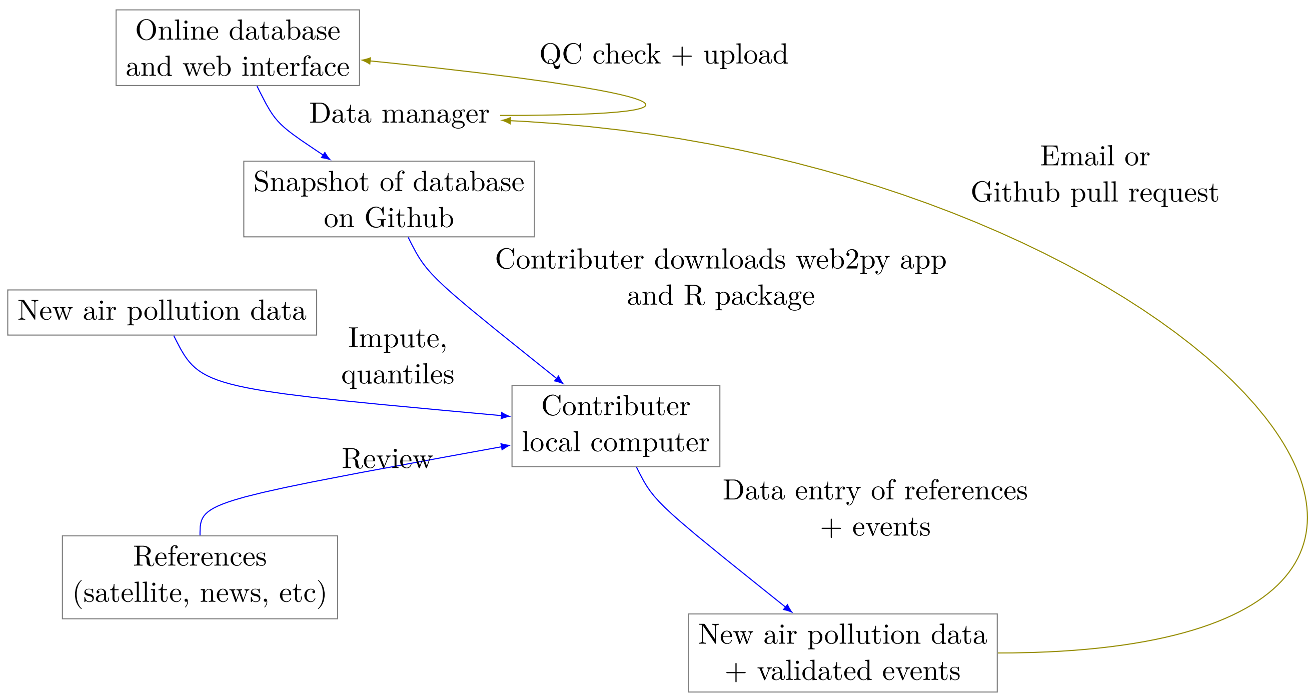

Figure 1: Schematic diagram of the online database and offline processes for extending the database

(with a specific version identifier from the Git version control system) and makes a 'standalone' version available on Github. This standalone version uses web2py so that it is capable of being downloaded and run on any operating system used by other computers. Contributers may download that version and use it as a local database. The R package is also available on Github, and contains functions that may be used to impute any missing data gaps using the APHEA procedure (Katsouyanni *et al.* 1996) as per the study protocol. The contributer needs to have new air pollution data available, and access to the required reference materials for validation. The R package is used to compute the quantiles of the new extended time-series of imputed pollution data, to identify events above the 95th percentile threshold that has been set to define 'extreme events'. The contributer uses the web2py data entry forms to add the information that is used to meet the validation criteria. Once they complete their review of all events they notify the DM either with email or by using the Github 'pull request' feature. The DM performs Quality Control (QC) checks and then uploads the new data to the online database. The procedure then starts again and a new version is loaded into the Github repository with descriptions of the additional changes that have been incorporated.

General overview of protocols

For each location in the original study there were up to 13 years (between 1994 and 2007) of daily air quality data measured as Particulate Matter (PM) less than 10um (PM_{10}) or less than 2.5 μm ($PM_{2.5}$) in aerodynamic diameter were examined. Air pollution data were provided by government agencies in the states of Western Australia, New South Wales, and Tasmania. Daily averages for each site were calculated excluding days with less than 75% of hourly measurements. In Sydney and Perth, where data were collected from several monitoring stations, the missing daily site-specific PM concentrations were imputed using available data from other proximate monitoring sites in the network. The daily city-wide PM concentrations were then estimated following the protocol of the Air Pollution and Health: a European Approach studies (Atkinson *et al.* 2001).

First a 'filling-in' procedure was used to improve data completeness. It entailed the substitution of the missing daily values with a weighted average, using the weights of the missing sites 3-month average proportional to the network average. The weights are calculated against the values from the rest of the monitoring stations. The pollutant measures from all stations providing data were then averaged to provide single, city-wide estimates of the daily levels of the pollutants

For each city, all days in which PM10 or PM2.5 exceeded the 95th percentile were identified over the entire time series. These extreme values were termed 'events'. A range of sources was ex- amined to identify the cause of particulate air pollution events, including electronic news archives, Internet searches for other reports, government and research agencies, satellite imagery and a Dust Storms database. Also examined were remotely sensed aerosol optical thickness (AOT) data to provide further information about days for which the other methods did not.

Detailed data preparation and validation methods

Step 1: Source air pollution data

Step 1.0 Source air pollution data. Both time series observations and spatial data regarding site locations.

Step 1.1. NSW data downloaded from an online data server. Site locations (Lat and Long) obtained from website.

Step 1.2. WA data sent on CD from contacts at the WA Government Department, these were hourly data as provided. Cleaned so as only days with >75% of hours are used. Licence puts restricions on our right to provide to a third party. Therefore those observed and imputed data are not included, only the events.

Step 1.3. Tasmanian data sent via email from contact at the Department, these were daily data.

Step 1.4. All data combined and Quality Control checked in the PostGIS database.

Step 2. Define spatial extent for cities

The cities and towns were selected based on the aims of the health study to investigate Cardio-respiratory disease and air pollution from biomass smoke events. These were Albany, Albury, Armidale, Bathurst, Bunbury, Busselton, Geraldton, Gosford-Wyong, Hobart, Illawarra, Launceston, Newcastle, Perth, Sydney, Tamworth and Wagga Wagga.

The spatial extent of each city and town was devised by intersecting Australian Bureau of Statistics Statistical Local Areas (SLAs) from the various Census editions. These boundaries were set so give the best possible representation of hospital admissions from the population.

Air pollution monitoring sites were then selected on the basis of their proximity to these populations.

Step 3. Imputation to fill in gaps in the time-series and calculate a network average

In cities where data were collected from several monitoring stations, the missing daily site-specific PM concentrations were imputed using available data from other proximate monitoring sites in the network. The daily city-wide PM concentrations were then estimated following the protocol of the Air Pollution and Health: a European Approach studies (Katsouyanni *et al.* 1996).

Step 3.1. Prepare Data. First it was necessary to find the minimum date that the series of continuous observations can be considered to start. In the Australian datasets the initial observations could not be used because the were sometimes only one day per week, only during a particular season or of poor quality due to teething problems with equipment and procedures. Then it was necessary to identify missing dates. Get a list of the sites to include – that is with more than 70% observed over the time period (as defined after assessing min and max dates of period).

Step 3.2. Loop over each station individually and calculate a daily network average of all the other non-missing sites (ie an average of all stations except the focal station of that iteration in the loop).

Step 3.3. Calculate three monthly seasonal mean of these non-missing stations. Calculate a three-month seasonal mean for MISSING site. Estimate missing days at missing sites. The missing value was replaced by the mean level of the remaining stations, multiplied by a factor equal to the ratio of the seasonal (centred three month) mean for the missing station, over the corresponding mean from the stations available on that particular day.

Step 3.4. Join all sites for city wide averages and fill any missing days at the site-level with average of the days immediately before and after the missing days (only when this is below a threshold).

Step 3.5 Take the average of all sites per day for city wide averages.

Step 3.6. Fill any missing days at the city-wide level with the average of before and after (if this is less than 5% of days).

Step 4. Validate events and identify the causes

Select any events with PM10 or PM2.5 greater than 95 percentile. Manually validate events using online newspaper archives, government and research agency records, satellite imagery and other sources (such as a Dust Storm database). Enter the information for each event into the custom built data entry forms. For any events with references for multiple types of source, assess the liklihood of any single source being the dominant source. Double check any remaining 99th percentile dates with no references.

Step 5. Insert contributed pollution and validated events, and downstream dissemination

- To close the loop the data are then inserted back into the DB.

Availability and requirements

Lists the following:

- Project name: BiosmokeValidatedEvents
- Project home page: http://swish-climate-impact-assessment.github.io/BiosmokeValidatedEvents/
- Operating system(s): R package is platform independent. Data Entry forms are Microsoft Windows.
- Programming language: R and SQL
- Other requirements: PostgreSQL (PostGIS is desirable)
- License: CC BY 4.0
- Any restrictions to use: amendments of errors of ommision or commission are invited but will be vetted before insertion into the master database.

Availability of supporting data

Air pollution data provided

The NSW Air pollution data are available to download from http://www.environment.nsw.gov.au/AQMS/search.htm

Data derived

The data set supporting the results of this article are available in the repository from the website http://swish-climate-impact-assessment.github.io/biomass_smoke_events_db

We have applied the license under Creative Commons - Attribution 4.0. This allows others to copy, distribute and create derivative works provided that they credit the original source.

Users should cite the Johnston 2011 Journal of the Air & Waste Management Association as the validation protocol and the Database itself as: Hanigan, IC., Johnston, FH., Morgan, GG., and contributers[*]. (2015). The Validated Bushfire Smoke Events Database. https://gislibrary-extreme-weather.anu.edu.au/biomass_smoke_events

References

Atkinson, R.W., Anderson, R.H., Sunyer, J., Ayres, J., Baccini, M., Vonk, J.M., Boumghar, A., Forastiere, F., Forsberg, B., Touloumi, G., Schwartz, J. & Katsouyanni, K. (2001). Acute Effects of Particulate Air Pollution on Respiratory Admissions. *American Journal of Respiratory and Critical Care Medicine*, 164(10), 1860–1866.

Johnston, F.H., Hanigan, I.C., Henderson, S.B., Morgan, G.G., Portner, T., Williamson, G.J. & Bowman, D.M.J.S. (2011). Creating an integrated historical record of extreme particulate air pollution events in Australian cities from 1994 to 2007. *Journal of the Air & Waste Management Association*, 61(4), 390–398.

Katsouyanni, K., Schwartz, J., Spix, C., Touloumi, G., Zmirou, D., Zanobetti, A., Wojtyniak, B., Vonk, J.M., Tobias, A., Ponka, A., Medina, S., Bacharova, L. & Anderson, H.R. (1996). Short term effects of air pollution on health: a European approach using epidemiologic time series data: the APHEA protocol. *Journal of Epidemiology & Community Health*, 50(Suppl 1), S12–S18.

Naeher, L.P., Brauer, M., Lipsett, M., Zelikoff, J.T., Simpson, C.D., Koenig, J.Q. & Smith, K.R. (2007). Woodsmoke health effects: A review. *Inhalation Toxicology*, 19(1), 67–106.

Peng, R.D. & Dominici, F. (2008). *Statistical Methods for Environmental Epidemiology with R. A Case Study in Air Pollution and Health.* Springer Science & Business Media, New York, USA.

Pope, C.A. & Dockery, D.W. (2006). Health Effects of Fine Particulate Air Pollution: Lines that Connect. *Journal of the Air & Waste Management Association*, 56(6), 709–742.

Smith, K. (1993). Fuel combustion, air pollution exposure, and health: The situation in developing countries. *Annual Review of Energy and the Environment*, 18: 529–66.