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*

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 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.

Description

The background and purpose of the database or data collection should be presented for readers without specialist knowledge in that area. For this database we should cite the original paper by Johnston et al. (2011a) as well as the two health analyses of Hospitalisation (Martin *et al.* 2013) and Mortality (Johnston *et al.* 2011b).

This will be followed by a brief description of the protocol for data collection, data curation and quality control, and what is being reported in the article.

The user interface should be described and a discussion of the intended uses of the database, and the benefits that are envisioned, should be included, together with data on how it compares with similar existing databases. A case study of the use of the database may be presented. The planned future development of new features, if any, should be mentioned.

The findings section can be broken into subsections with short informative headings. There is no maximum length for this section but we encourage authors to be concise.

General Protocols

For each location, up to 13 yr (between 1994 and 2007) of daily air quality data measured asPMless than 10um (PM10) or less than 2.5 um (PM2.5) in aerodynamic diameter were examined. Air pollution data were pro- vided 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 sta- tions, the missing daily site-specific PM10 and PM2.5 con- centrations were imputed using available data from other proximate monitoring sites in the network. The daily city-wide PM10 and PM2.5 concentrations were then estimated following the protocol of the Air Pollution and Health: a European Approach studies (Atkinson *et al.* 2001).

Detailed Data Collation and Validation Methods

Step 1: Imputation to fill in gaps in the time-series

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.

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. Spatial data for cities.

Step 3. 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. Atkinson et al. (2001).

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.

Step 3.4. Join all sites for city wide averages and fill any missing days with avg of before and after.

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

Step 3.6. Fill any missing days with avg 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.

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

BMC Research Notes encourages authors to deposit the data set(s) supporting the results reported in submitted manuscripts in a publicly-accessible data repository, when it is not possible to publish them as additional files. This section should only be included when supporting data are available and must include the name of the repository and the permanent identifier or accession number and persistent hyperlink(s) for the data set(s). The following format is required:

"The data set(s) supporting the results of this article is(are) available in the [repository name] repository, [unique persistent identifier and hyperlink to dataset(s) in http:// format]."

Where all supporting data are included in the article or additional files the following format is required:

"The data set(s) supporting the results of this article is(are) included within the article (and its additional file(s))"

We also recommend that the data set(s) be cited, where appropriate in the manuscript, and included in the reference list.

A list of available scientific research data repositories can be found here. A list of all BioMed Central journals that require or encourage this section to be included in research articles can be found here.

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., Hanigan, I., Henderson, S., Morgan, G. & Bowman, D. (2011a). Extreme air pollution events from bushfires and dust storms and their association with mortality in Sydney, Australia 1994-2007. *Environmental Research*, 111(6), 811–816.

Johnston, F.H., Hanigan, I.C., Henderson, S.B., Morgan, G.G., Portner, T., Williamson, G.J. & Bowman, D.M.J.S. (2011b). 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.

Martin, K.L., Hanigan, I.C., Morgan, G.G., Henderson, S.B. & Johnston, F.H. (2013). Air pollution from bushfires and their association with hospital admissions in Sydney, Newcastle and Wollongong, Australia 1994-2007. *Australian and New Zealand Journal of Public Health*, 37(3), 238–243.