Method Article

Improving long-term air pollution estimates with incomplete data: A method-fusion approach

Karl Chastko, Matthew Adams*

Department of Geography, University of Toronto Mississauga, Ontario, Canada

ABSTRACT

Mobile air pollution monitoring is an effective means of collecting spatially and temporally diverse air pollution samples. These observations are often used to predict long-term air pollution concentrations using temporal adjustments based on the time-series of a fixed location monitor. Temporal adjustments are required because the time-series is often incomplete at each spatial location. We describe a method-fusion temporal adjustment that has been demonstrated to improve the accuracy of long-term estimates from incomplete time-series data. Our adjustment approach combines the techniques of using a log transformation to modify the air pollution samples to a near normal distribution and incorporates the long-term median of a reference monitor to mediate the effects of estimate inflation created by outliers in the data. We demonstrate the approach with hourly Nitrogen Dioxide observations from Paris, France in 2016.

Method-Fusion Benefits:

- Log transformations control for estimate inflation created by log normally distributed data.
- Adjusting data with the long-term median, rather than the mean, controls for estimate inflation.
- Produces more accurate long-term estimates than other adjustments independent of the pollutant being estimated.

© 2019 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

ARTICLE INFO

Method name: Log median-scaled adjustment

Keywords: Temporal adjustments, Air pollution, Long-term estimates

Article history: Received 19 April 2019; Accepted 9 June 2019; Available online 12 June 2019

* Corresponding author.
E-mail address: md.adams@mail.utoronto.ca (M. Adams).

https://doi.org/10.1016/j.mex.2019.06.005
2215-0161/© 2019 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
Air pollution monitoring requires high-precision instruments, which are expensive and often limit the number of locations observed [1]. In addition to financial constraints, time constraints presented by monitoring multiple sites with a limited number of monitors often prevent long-term continuous monitoring, which is required to accurately estimate long-term air pollution at a given location [2]. To address these challenges, researchers often deploy portable monitors for short periods to observe air pollution at many different locations. The goal of the portable monitor use is to obtain spatially and temporally diverse data. The data collected from the short-term monitoring campaigns have been used to estimate long-term concentrations and develop predictive models of air pollution. Long-term estimates are necessary because these are the values most often used in epidemiological studies of health effects from air pollution [3,4]. Aside from calculating the raw average of short-term samples and treating these values as the long-term concentration estimate, multiplicative temporal adjustments are often employed to predict long-term exposure estimates [5–7].

**Multiplicative temporal adjustments**

Multiplicative temporal adjustments are used to correct for temporal trends present in air pollution data collected through short term monitoring campaigns [8]. When data are collected at short intervals at different times throughout the year, or at different times throughout the day, seasonal and diurnal variation in air pollution becomes exaggerated. Multiplicative temporal adjustments account for these trends by adjusting the short-term observations against continuous data observed at a fixed-location monitoring station. For example, if data were collected during a period of above average air pollution, the adjustment would correct these observations downwards when estimating the long-term value. Applying temporal adjustments can improve accuracy of long-term pollution concentration estimates. Eq. (1) describes the basic form of a multiplicative temporal adjustment.

\[
O_A = \frac{O_t}{O_{Central \ Tendency \ FM}}
\]  

Long-term estimates of air pollution are calculated by dividing each short-term observation \(O_t\) by the ratio of its corresponding fixed monitor observation \(O_{FMt}\) to the long-term central tendency of a fixed location reference monitor \(O_{Central \ Tendency \ FM}\) [9].

**A method-fusion approach**

Air pollutants are often log-normally distributed, with many low concentration observations and fewer high concentration observations, see Fig. 1.
The log-normal distribution of air pollution presents challenges when estimating long-term concentrations, due to the influence of extreme values. Two main approaches have been employed to help account for the log-normal distribution commonly present in air pollution. Approach one, see Eq. (2), utilizes the median as a measure of central tendency for the fixed location monitor [7]. By calculating the median rather than the mean, the estimates produced with this approach are less inflated by extreme values and are a better representation of the observed long-term concentration.

\[ O_A = \frac{O_t}{\frac{O_{FMt}}{O_{FMmedian}}} \]  

Approach two, see Eq. (3), applies a log transformation to the data, which transforms the data to an approximately normal distribution. This transformation minimizes the potential exaggeration of estimates produced from the presence of extreme values. This approach utilizes the mean as its measure of central tendency.

\[ O_A = \frac{O_t}{\frac{\text{log}(O_{FMt} + e)}{\text{log}(O_{FMmedian} + e)}} \]  

By incorporating elements from both approaches into one temporal adjustment, more accurate results may be obtained. Eq. (3) displays the combined temporal adjustment. Unlike other temporal adjustments, Eq. (3) utilizes both the log adjustment and median value of the fixed monitor to control for the log-normal distribution of the data being adjusted. We refer to this method-fusion approach as the log median-scaled adjustment.

\[ O_A = \frac{O_t}{\frac{\text{log}(O_{FMt} + e)}{\text{log}(O_{FMmedian} + e)}} \]  

**Model validation**

This method has been demonstrated to improve estimates compared to other multiplicative adjustments [10]. Chastko and Adams [10], simulated mobile monitoring campaigns with air pollution observations from three different cities and eight pollutants. Mobile pollution samples were adjusted using multiple temporal adjustment approaches to predict long-term concentrations for each pollutant. This analysis revealed that the log median-scaled adjustment was more accurate the all other temporal adjustments included in the study. For full details on model validation, see Chastko and Adams [10].
Conclusion

The method fusion approach can be applied to any mobile air pollution monitoring dataset and can produce more accurate long-term estimates compared to existing temporal adjustments. These estimates are more accurate because the method fusion approach controls for inflation of the central tendency produced by log-normal distributions, which are often present in air pollution data.

Sample workflow

The following example demonstrates how to apply the log median-scaled adjustment to a sample of air pollution data. This workflow is presented in the programming language R. For this example, mobile data will be represented as a subset of data from a stationary air pollution monitor in Paris France.

R libraries

To access various functions used in this demonstration, the following R libraries are loaded.

```
library(tidyverse)
library(dplyr)
library(readr)
```

Loading the data

To access the sample data for this demonstration, a CSV file containing hourly Nitrogen Dioxide observations from 2016 in Paris France is loaded into R. This data is hosted on GitHub and was originally obtained from AirParif [11]. The following code block loads the data into R, assigning it to a variable `air.pollution.data`.

```
air.pollution.data <- read_csv("https://raw.githubusercontent.com/kchastko/ParisAirPollutionData/master/NO2_2016_Example.csv")
```

A sample of the air pollution data is provided in Table 1, it is a collection of air pollution observations from two air pollution monitors with time signatures for each observation.

Sampling and adjustment function

With the data loaded into R, a sample of air pollution data can be taken from one of the monitors to represent the mobile data. In this example, 24 h of air pollution observations will be used. Table 2 displays the sample data and the temporally corresponding reference data that will be used to adjust each sample from the mobile data. Station 1 is used as the reference station and Station 2 is used as the sample station. Additionally, the annual median NO2 concentration is calculated for the reference monitor. Fig. 2 displays the time slice of mobile data in relation to the entire time series of NO2 values observed at Station 2.

Table 1
Sample of air pollution observation obtained from AirParif. Values represent NO2 concentrations measured in μg/m3.

| Date       | Hour | DateTime             | Station1 | Station2 |
|------------|------|----------------------|----------|----------|
| 1/1/2016   | 1    | 1/1/2016 0:00        | 35       | 21       |
| 1/1/2016   | 2    | 1/1/2016 1:00        | 42       | 34       |
| 1/1/2016   | 3    | 1/1/2016 2:00        | 55       | 32       |
| 1/1/2016   | 4    | 1/1/2016 3:00        | 54       | 45       |
| 1/1/2016   | 5    | 1/1/2016 4:00        | 50       | 53       |
| 1/1/2016   | 6    | 1/1/2016 5:00        | 47       | 51       |
| 1/1/2016   | 7    | 1/1/2016 6:00        | 47       | 58       |
| 1/1/2016   | 8    | 1/1/2016 7:00        | 46       | 56       |
| 1/1/2016   | 9    | 1/1/2016 8:00        | 40       | 50       |
| 1/1/2016   | 10   | 1/1/2016 9:00        | 31       | 47       |
Table 2
Mobile sample taken from complete time series obtained from Paris air pollution monitors.

| Date       | Hour | Date Time   | Reference Data | Mobile Sample |
|------------|------|-------------|----------------|---------------|
| 1/4/2016   | 14   | 4/1/2016 14:00 | 19              | 12            |
| 1/4/2016   | 15   | 4/1/2016 15:00 | 20              | 11            |
| 1/4/2016   | 16   | 4/1/2016 16:00 | 27              | 17            |
| 1/4/2016   | 17   | 4/1/2016 17:00 | 25              | 26            |
| 1/4/2016   | 18   | 4/1/2016 18:00 | 30              | 38            |
| 1/4/2016   | 19   | 4/1/2016 19:00 | 40              | 53            |
| 1/4/2016   | 20   | 4/1/2016 20:00 | 60              | 58            |
| 1/4/2016   | 21   | 4/1/2016 21:00 | 62              | 56            |
| 1/4/2016   | 22   | 4/1/2016 22:00 | 61              | 56            |
| 1/4/2016   | 23   | 4/1/2016 23:00 | 40              | 56            |
| 1/4/2016   | 24   | 4/2/2016 0:00  | 46              | 37            |
| 2/4/2016   | 1    | 4/2/2016 1:00  | 58              | 38            |
| 2/4/2016   | 2    | 4/2/2016 2:00  | 63              | 43            |
| 2/4/2016   | 3    | 4/2/2016 3:00  | 60              | 58            |
| 2/4/2016   | 4    | 4/2/2016 4:00  | 60              | 64            |
| 2/4/2016   | 5    | 4/2/2016 5:00  | 62              | 62            |
| 2/4/2016   | 6    | 4/2/2016 6:00  | 60              | 59            |
| 2/4/2016   | 7    | 4/2/2016 7:00  | 50              | 57            |
| 2/4/2016   | 8    | 4/2/2016 8:00  | 62              | 56            |
| 2/4/2016   | 9    | 4/2/2016 9:00  | 48              | 52            |
| 2/4/2016   | 10   | 4/2/2016 10:00 | 30              | 47            |
| 2/4/2016   | 11   | 4/2/2016 11:00 | 22              | 26            |
| 2/4/2016   | 12   | 4/2/2016 12:00 | 15              | 23            |
| 2/4/2016   | 13   | 4/2/2016 13:00 | 16              | 18            |

Fig. 2. 72 h moving average of NO₂ observations in 2016 (black) and the 24 -h sample period (red) from April 1 st 2 pm to April 2nd 1 pm. All observations were recorded at the Argenteuil (Station2) air pollution monitoring station in Paris France.
Now that all the required data have been selected, we can proceed to apply the temporal adjustment. The R implementation of the adjustment is shown below as the LogMedianScaled function. This function requires a vector containing the reference data, a vector containing the mobile data and the annual median pollutant concentration calculated from the entire reference dataset.

```
LogMedianScaledAdjustment <- function(ReferenceData, SampleData, ReferenceMedian){
  ratio <- log(ReferenceData + exp(1)) / log(ReferenceMedian + exp(1))
  adjusted.samples <- SampleData / ratio
  estimate <- mean(adjusted.samples[,1], na.rm = T)
  return(estimate)
}
```

The LogMedianScaledAdjustment function returns a single value representing the annual average air pollution concentration estimated from the mobile sample by averaging each adjusted value in the mobile sample. Table 3 shows the adjusted values calculated by the temporal adjustment and the raw input values used to adjust the data. In this example, the raw data produce a long-term estimate of 42.62 ppb, the LogMedianScaledAdjustment produces an annual NO₂ estimate of 37.15 ppb, and the actual long-term value was 30.6 ppb. By applying the log median-scaled adjustment, estimation error was reduced from 12.02 ppb to 6.55 ppb using only a 24 h sample to estimate the annual average.

To visualize the accuracy of the temporally adjusted estimate we can plot the observed average annual NO₂ value, calculated from the stationary data, alongside the raw sample average, the temporally adjusted average and the sample data. Fig. 3 shows that the temporally adjusted average is more accurate than the raw sample’s average.

| Reference Data | Reference Median | Mobile Data | Adjusted Values |
|----------------|------------------|-------------|----------------|
| 19             | 26               | 12          | 13.09          |
| 20             | 26               | 11          | 11.83          |
| 27             | 26               | 17          | 16.83          |
| 25             | 26               | 26          | 26.28          |
| 30             | 26               | 38          | 36.58          |
| 40             | 26               | 53          | 47.39          |
| 60             | 26               | 58          | 47.05          |
| 62             | 26               | 56          | 45.09          |
| 61             | 26               | 56          | 45.26          |
| 40             | 26               | 56          | 50.08          |
| 46             | 26               | 37          | 31.97          |
| 58             | 26               | 38          | 31.07          |
| 63             | 26               | 43          | 34.49          |
| 60             | 26               | 58          | 47.05          |
| 60             | 26               | 64          | 51.92          |
| 62             | 26               | 62          | 49.92          |
| 60             | 26               | 59          | 47.86          |
| 50             | 26               | 57          | 48.27          |
| 62             | 26               | 56          | 45.09          |
| 48             | 26               | 52          | 44.47          |
| 30             | 26               | 47          | 45.24          |
| 22             | 26               | 26          | 27.22          |
| 15             | 26               | 23          | 26.86          |
| 16             | 26               | 18          | 20.63          |
Fig. 3. Time series of mobile sample and adjusted sample values with estimates and observed annual average. Red: Raw sample mean, Green: Adjusted mean, Black: Observed annual mean calculated from complete time series.

Acknowledgement

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), [RGPIN-2018-04845].

References

[1] E.G. Snyder, T.H. Watkins, P.A. Solomon, E.D. Thoma, R.W. Williams, G.S.W. Hagler, D. Shelow, D.A. Hindin, V.J. Kilaru, P.W. Preuss, The changing paradigm of air pollution monitoring, Environ. Sci. Technol. 47 (2013) 11369–11377, doi:http://dx.doi.org/10.1021/es4022602.

[2] J.C. Chow, J.P. Engelbrecht, J.G. Watson, W.E. Wilson, N.H. Frank, T. Zhu, Designing monitoring networks to represent outdoor human exposure, Chemosphere 49 (2002) 961–978.

[3] B. Brunekreef, S.T. Holgate, Air pollution and health, Lancet 360 (2002) 1233–1242, doi:http://dx.doi.org/10.1016/S0140-6736(02)11274-8.

[4] M. Kampa, E. Castanas, Human health effects of air pollution, Environ. Pollut. 151 (2008) 362–367, doi:http://dx.doi.org/10.1016/j.envpol.2007.06.012.

[5] R. Beelen, G. Hoek, D. Vienneau, M. Eeftens, K. Dimakopoulou, X. Pedeli, M.Y. Tsai, N. Künzli, T. Schikowski, A. Marcon, K.T. Eriksen, O. Raaschou-Nielsen, E. Stephanou, E. Patelarou, T. Lanki, T. Yli-Tuomi, C. Declercq, G. Falq, M. Stempflet, M. Birk, J. Cyrys, S. von Klot, G. Nádó, M.J. Varró, A. Dedele, R. Gražulevičiene, A. Mölter, S. Lindley, C. Madsen, G. Cesaroni, A. Ranzi, C. Badaloni, B. Hoffmann, M. Nonnenmacher, U. Krämer, T. Kuhlbusch, M. Cirach, A. de Nazelle, M. Nieuwenhuijsen, T. Bellander, M. Korek, D. Olsson, M. Strömberg, E. Dons, M. Jerrett, P. Fischer, M. Wang, B. Brunekreef, K. de Hoogh, Development of NO2 and NOx land use regression models for estimating air pollution exposure in 36 study areas in Europe—the ESCAPE project, Atmos. Environ. 72 (2013) 10–23, doi:http://dx.doi.org/10.1016/j.atmosenv.2013.02.037.

[6] E. Dons, L. Int Panis, M. Van Poppel, J. Theunis, G. Wets, Personal exposure to Black Carbon in transport microenvironments, Atmos. Environ. 55 (2012) 392–398, doi:http://dx.doi.org/10.1016/j.atmosenv.2012.03.020.

[7] T. Larson, S.B. Henderson, M. Brauer, Mobile monitoring of particle light absorption coefficient in an urban area as a basis for land use regression, Environ. Sci. Technol. 43 (2009) 4672–4678, doi:http://dx.doi.org/10.1021/es900368e.

[8] J.S. Apte, K.P. Messier, S. Gani, M. Brauer, T.W. Kirchstetter, M.M. Lunden, J.D. Marshall, C.J. Portier, R.C.H. Vermeulen, S.P. Hamburg, High-resolution air pollution mapping with google street view cars: exploiting big data, Environ. Sci. Technol. 51 (2017) 6999–7008, doi:http://dx.doi.org/10.1021/acs.est.7b00891.

[9] G. Hoek, K. Meuliefe, J. Cyrys, M. Lewné, T. Bellander, M. Brauer, P. Fischer, U. Gehring, J. Heinrich, P. van Vliet, B. Brunekreef, Spatial variability of fine particle concentrations in three European areas, Atmos. Environ. 36 (2002) 4077–4088, doi:http://dx.doi.org/10.1016/S1352-2310(02)00297-2.

[10] K. Chastko, A. Matthew, Assessing the accuracy of long-term air pollution estimates produced with temporally adjusted short-term observations from unstructured sampling, J. Environ. Manage. 240 (2019) 249–258.

[11] Airparif, Air Pollution Monitoring Network [URL https://www.airparif.asso.fr/en/telechargement/telechargement-station (Accessed 13 July 2018), (2016).]