Air Pollution Data Management and Compression

David Fitzgerald, Luke Knibbs August 2020

Compressing the Air Pollution data

An air pollution and road distance dataset was produced by Luke Knibbs, called *AP_roads*. This dataset had levels of nitrogen dioxide (NO_2), fine particulate matter ($PM_{2.5}$) air pollutants and also road distances and road density variables, as proxies of traffic pollution. A description of Luke Knibbs' method of producing the dataset is in the document *AirPollutionExposureVariables*.

The air pollution part of the initial dataset, AP_roads, was both long and wide; it contained more information than necessary. It contained air pollution data from times and places when the ALSWH respondent was not present, and therefore the data had to be compressed so that it only had air pollution that the respondent was exposed to. Note that the road distance data was not wide and did not need any such management.

This document describes the data structure, the compressing method, and the resulting dataset for analysis.

A description of the resulting dataset and its variables is in the document *Air Pollution Datasets.*

The Road Distance data was differently structured and did not need any compression. The Road Distance data is explained and displayed at the bottom of this document.

Data Structure

The initial data, AP_roads, had air pollution levels for every wave the ALSWH women responded in. These spanned the years 1996 to 2017. The Air Pollution levels were based on the addresses we had for each women at survey response time. Table 1 shows some NO_2 data for one respondent, ID 9, in the 1973-78 cohort.

Note that the last three columns were not on the AP_roads data but were added. The response date, resp date, were from the ALSWH participant status files. The estimated date moved and year moved were derived and are explained below.

Table 1 Example NO2 data, Years 2003 to 2009 only, from the 1973-78 cohort.

id	row	wave	moved	no2_2003	no2_2004	no2_2005	no2_2006	no2_2007	no2_2008	no2_2009	Resp Date	Est Date moved	Est Year Moved
9	1	1	.	3.7	3.6	3.3	3.3	3.2	2.9	3.0	16SEP1996	.	.
9	2	2	0	3.7	3.6	3.3	3.3	3.2	2.9	3.0	19APR2000	.	.
9	3	4	1	5.0	4.8	4.6	4.6	4.5	4.2	4.3	31OCT2006	26JUL2003	2004
9	4	5	1	5.3	5.2	5.0	4.9	4.9	4.6	4.7	11JUN2009	20FEB2008	2008
9	5	6	0	5.3	5.2	5.0	4.9	4.9	4.6	4.7	10AUG2012	.	.
9	6	7	0	5.3	5.2	5.0	4.9	4.9	4.6	4.7	21MAR2016	.	.
9	7	8	0	5.3	5.2	5.0	4.9	4.9	4.6	4.7	30DEC2018	.	

Data Description

The first row shows the NO_2 levels for years 2003 to 2009 at the address at wave 1 for ID 9. The second row shows ID 9 next responded at wave 2 and the MOVED column shows that she did not change address from wave 1. Row 2 also shows the NO_2 levels for that address for years 2003 to 2009 and since this is the same address as row 1 the NO_2 levels are exactly the same. Row 3 shows she next responded at wave 4, she missed wave 3, and that she is in a new address at wave 4 because MOVED is indicated. The NO_2 levels for row 3 are for the new address she reported in wave 4. Similarly, row 4 shows she is now in a new address again in wave 5 and the NO_2 levels are for that new address. Rows 5, 6 and 7 show that she responded to waves 6, 7 and 8 and she did not change address. Therefore, the NO_2 levels for rows 4 to 7 are the same because they are from the same address.

It is clear that a lot of the data in Table 1 is not needed because the exposure NO_2 levels are repeated and also they are not always from the respondent's address at the time they were living there. For example, in the table the respondent had moved address by wave 4 in 2006 but the first two rows show NO_2 levels for the earlier address for 2006 and later. Therefore a lot of the data needed to be removed and the data needed to be compressed into a single row with yearly exposure NO_2 levels. This is complicated by having yearly NO_2 levels in the columns but response year levels in the rows. Also, we do not know in what year they changed address; we only know the when they reported a new address at the response time.

Determining Change of Address

We knew the respondent's longitude and latitude at each survey wave (i.e., geocoded addresses). If these changed from one wave to the next then the respondent was deemed to have moved.

By comparing the co-ordinates at each wave with the most recent previous wave for each woman, we were able to determine if they had moved in that time period. Suspected moves were flagged, and the straight-line distance between old and new address was determined using ArcGIS. This allowed a more detailed check of the plausibility of suspected moves. In total, there were 73,382 suspected moves. Of those moves, 1,647 (2.2%) were re-coded to non-moves because the distance was: a) implausibly small and possibly due to rounding during geocoding (e.g., 10 m), or b) smaller than the spatial resolution of the highest resolution pollution model (NO_2, 100 m). This left 71,735 moves among 35,658 individuals. Both the mover flag and the associated distance were recorded, as this will be important for sensitivity analyses where differences between movers and non-movers, and among short- and long-distance movers, are compared in terms of the associations between pollutants and health outcomes.

Compression method

We needed to have an estimate for the year they moved. We took the mid-point date between the survey response date where the new address was reported and the date of the most recent previous response. In Table 1 the first movement was noted in Oct 2006 and the previous response date was April 2000. The mid-point between these was 26 July 2003 and this date was rounded to 2004. Similarly, the second movement, noted in wave 5, was estimated to be 20 Feb 2008 and rounded to 2008. The mid-points were rounded up to the next year for months July to December, otherwise they were rounded down. These rounded years were the estimated Year Moved values.

The estimated year moved was used to determine which NO_2 levels should be used in the compressed data. The method firstly assigned the NO_2 levels from the first address, or first row, to all years. If there was no change of address then these will be the final NO_2 levels. However, if there was a change of address then the estimated Year Moved values were used to assign the NO_2 levels from that year onwards from the wave/row when the move was reported. For example, in the highlighted data in Table 2, 2004 was the first estimated year moved, so NO_2 values for 2004 onwards taken from wave 4, the wave the movement was noted, were applied. These values would only be replaced if there was another move. This method was repeated for all moves. In

the example 2008 was the next move in wave 5, and so NO_2 levels for 2008 onwards were used from wave 5's row. Because there were no more changes of addresses in the example these wave 5 values were used from 2008 onwards.

Table 2 shows in highlight which NO2 levels were assigned to the compressed single row.

Table 2 Highlighting values used in final compressed data

id	wave	moved	no2_2003	no2_2004	no2_2005	no2_2006	no2_2007	no2_2008	no2_2009	Resp date	Date Moved	Year Moved
9	1	0	3.7	3.6	3.3	3.3	3.2	2.9	3.0	16SEP1996	.	.
9	2	0	3.7	3.6	3.3	3.3	3.2	2.9	3.0	19APR2000	.	.
9	4	1	5.0	4.8	4.6	4.6	4.5	4.2	4.3	31OCT2006	26JUL2003	2004
9	5	1	5.3	5.2	5.0	4.9	4.9	4.6	4.7	11JUN2009	20FEB2008	2008
9	6	0	5.3	5.2	5.0	4.9	4.9	4.6	4.7	10AUG2012	.	.
9	7	0	5.3	5.2	5.0	4.9	4.9	4.6	4.7	21MAR2016	.	.
9	8	0	5.3	5.2	5.0	4.9	4.9	4.6	4.7	30DEC2018	.	

The final compress data has one row per respondent, as shown in Table 3.

Table 3 Compressed data

id	no2_2003	no2_2004	no2_2005	no2_2006	no2_2007	no2_2008	no2_2009
9	3.7	4.8	4.6	4.6	4.5	4.6	4.7

Table 3 is a single row from the resulting Air Pollution dataset released to researchers. It also includes the fine particulate matter ($PM_{2.5}$) air pollutants for years 1998 to 2016. The exact same compression method was repeated for the fine Particulate Matter levels.

The final compressed data has 57,346 records, one for each woman in the four ALSWH cohorts. The variables are NO_2 levels from 1996 to 2017, and two versions of the fine particular matter levels from 1998 to 2016. It is called AirPollution

Roads Data are different

The roads data were not given for each year because they were not time-varying, unlike the pollution estimates which were, therefore the roads data did not need to be compressed. However, there was a record for each person and each response wave so the data is long not wide. The 1989-95 cohort only has 5 waves, the 1973-78 and 1946-51 cohorts have 8 waves, and the 1921-26 cohort has up to 19 waves. There are 304,851 records. Table 5 shows an example of the Roads Data for an ID from each cohort.

Table 5 Example of Roads data with one respondent per cohort

ID	cohort	wave	SUM_ALLROADS_100M	SUM_ALLROADS_200M	SUM_ALLROADS_500M	ROAD_DIST_ALL
6	YNG	1	0.338	1.051	5.379	14.5
6	YNG	2	0.338	1.051	5.379	14.5
6	YNG	4	0.147	1.533	7.134	16.0
6	YNG	5	0.256	1.156	8.039	9.2
6	YNG	6	0.256	1.156	8.039	9.2
6	YNG	7	0.256	1.156	8.039	9.2
6	YNG	8	0.256	1.156	8.039	9.2
7	MID	1	0.428	1.236	8.195	13.6
7	MID	2	0.396	1.497	4.213	29.9
7	MID	3	0.332	1.278	2.815	22.2
7	MID	4	0.523	1.871	7.814	27.0
7	MID	5	0.523	1.871	7.814	27.0
7	MID	6	0.523	1.871	7.814	27.0
7	MID	7	0.523	1.871	7.814	27.0
7	MID	8	0.523	1.871	7.814	27.0

ID	cohort	wave	SUM_ALLROADS_100M	SUM_ALLROADS_200M	SUM_ALLROADS_500M	ROAD_DIST_ALL
8	OLD	1	0.476	1.587	9.798	18.9
8	OLD	2	0.476	1.587	9.798	18.9
8	OLD	3	0.399	1.429	9.416	1.8
8	OLD	4	0.399	1.429	9.416	1.8
8	OLD	5	0.399	1.429	9.416	1.8
8	OLD	6	0.183	1.442	10.344	58.8
8	OLD	7	0.183	1.442	10.344	58.8
8	OLD	8	0.183	1.442	10.344	58.8
8	OLD	9	0.183	1.442	10.344	58.8
8	OLD	10	0.183	1.442	10.344	58.8
8	OLD	11	0.183	1.442	10.344	58.8
8	OLD	12	0.183	1.442	10.344	58.8
8	OLD	13	0.183	1.442	10.344	58.8
8	OLD	14	0.183	1.442	10.344	58.8
8	OLD	15	0.183	1.442	10.344	58.8
8	OLD	16	0.183	1.442	10.344	58.8
8	OLD	17	0.183	1.442	10.344	58.8
8	OLD	18	0.183	1.442	10.344	58.8
8	OLD	19	0.183	1.442	10.344	58.8
9	NYC	1	0.430	1.566	9.666	9.4
9	NYC	2	0.430	1.566	9.666	9.4
9	NYC	3	0.587	1.497	8.052	21.1
9	NYC	4	0.200	0.630	2.658	1.6
9	NYC	5	0.000	0.250	3.237	134.5