# Air Pollution and Roads Distance Data Management, Compression, Change of Address dataset and Recodes

David Fitzgerald, Luke Knibbs August 2020

Contents:



# Compressing the Air Pollution data

The AP_roads data was produced by Luke Knibbs. The data has levels of nitrogen dioxide ($NO_2$) and 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 is in the document *AirPollutionExposureVariables*.

A description of the three datasets and their variables is in the document *Air Pollution Datasets*

**Data Structure**

The initial data 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 woman, ID 9, in the 1973-78 cohort.

**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 | Date moved | 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 year 2003 to 2009 and since this is the same address as in row 1 the $NO_2$ levels are exactly the same. Row 3 shows she next responded at wave 4, missing wave 3, and that she is in a new address at wave 4 since MOVED is indicated. The $NO_2$ levels for row 3 are for the 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 for 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 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 2007 but the first two rows show $NO_2$ levels for the earlier address for 2007 and later. Therefore a lot of the data needed to be removed and the data could 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 had the respondent's address'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 (or changing address) 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 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. 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 |

This will be the 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.

**Change of Address Dataset**

The researchers may want to know if the respondents had changed address, 'moved' in the above example, and if so, when and how far they moved. A ChangeAddress dataset will be made available with this information, including distance moved and estimated year moved. Table 4 shows the ChangeAddress dataset rows the ID 9 in the above example.

**Table 4 Mover dataset example**

| Obs | ID | wave | Cohort | moved | dist_km | EstYearMoved |
|---|---|---|---|---|---|---|
| **1** | 9 | 1 | YNG | . | . | . |
| **2** | 9 | 2 | YNG | 0 | . | . |
| **3** | 9 | 4 | YNG | 1 | 157.65 | 2004 |
| **4** | 9 | 5 | YNG | 1 | 1.82 | 2008 |
| **5** | 9 | 6 | YNG | 0 | . | . |
| **6** | 9 | 7 | YNG | 0 | . | . |
| **7** | 9 | 8 | YNG | 0 | . | . |

**Cohort differences**

The example data used above was from the 1973-78 cohort, referred to as YNG, that is, 'young', in the dataset. The 1946-51 cohort is similar in that collection started in 1996 and continued roughly every 3 years over 8 waves at the time of this work. The 1989-95 cohort only started collection in 2012/13 so we did not know about any address changes before this time and there were only 5 waves at the time this work was done. The 1921-26 cohort began in 1996 but then also began a six-month follow up survey in 2011 and so there were many waves to record any changes of address. Nevertheless, the method used to compress was the same in each cohort but with different dates and number of waves.

**Air Pollution Compressed data**

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.

**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 |
| 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 |

| ID | cohort | wave | SUM_ALLROADS_100M | SUM_ALLROADS_200M | SUM_ALLROADS_500M | ROAD_DIST_ALL |
|---|---|---|---|---|---|---|
| 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 |

# Recodes of AP after compressing the AP data

ALSWH does not allow its data to identify any respondent. The air pollution and road distance data is based on addresses and so we were extra careful not to allow any identification from these data. Therefore we decided that the air pollution data would have any value with fewer than 5 respondents having this value. Any values that had a frequency of fewer than 5 were re-coded by the following method.

**Recode for Air Pollution**

These recodes were performed after the compression of the data described above.

There recodes were done for each cohort and for each value from 1996 to 2017 separately.

The air pollution data are two-tailed and the low frequency values are all found in the lower and upper tails. The lower tail often did not have low frequencies since many had the lowest value of zero. For the upper tail any ‘fewer than 5 frequency’ value was recoded to highest value less than the ‘fewer than 5 frequency’ value that had at least 5 frequency. This is shown in the following example.

**Example Recode in Upper Tail**

The example below shows the recoding process. The data show frequencies for $NO_2$ 1996 values ranging from 18.3 to 19.4. These upper tail values are from the 1989-95 cohort. The highlighted values have frequencies for 5 or above while the other rows have fewer than 5. All values from 18.4 to 19.0 will be recoded to 18.3 because 18.3 is the highest value less than these values that has at least 5 frequency. Similarly, the values 19.2 and 19.4 get recoded to 19.1

| id | cohort | PRED_NO2 | freq96 |
|---|---|---|---|
| 22 | NYC | 18.3 | 11 |
| 22 | NYC | 18.4 | 2 |
| 22 | NYC | 18.5 | 4 |
| 22 | NYC | 18.6 | 1 |
| 22 | NYC | 18.7 | 2 |
| 22 | NYC | 18.8 | 1 |
| 22 | NYC | 18.9 | 1 |
| 22 | NYC | 19.0 | 2 |

| id | cohort | PRED_NO2 | freq96 |
|---|---|---|---|
| 22 | NYC | 19.1 | 7 |
| 22 | NYC | 19.2 | 4 |
| 22 | NYC | 19.3 | . |
| 22 | NYC | 19.4 | 3 |

A frequency analysis shows the frequencies after the recode.

| pred_no2_1996 | Frequency |
|---|---|
| **18.3** | 24 |
| **19.1** | 14 |

Lower Tail

Similarly any lower tail 'fewer than 5 frequency' value was recoded to lowest value less than the 'fewer than 5 frequency' value that had at least 5 frequency. As shown in the example.

**Example Recode in Lower Tail**

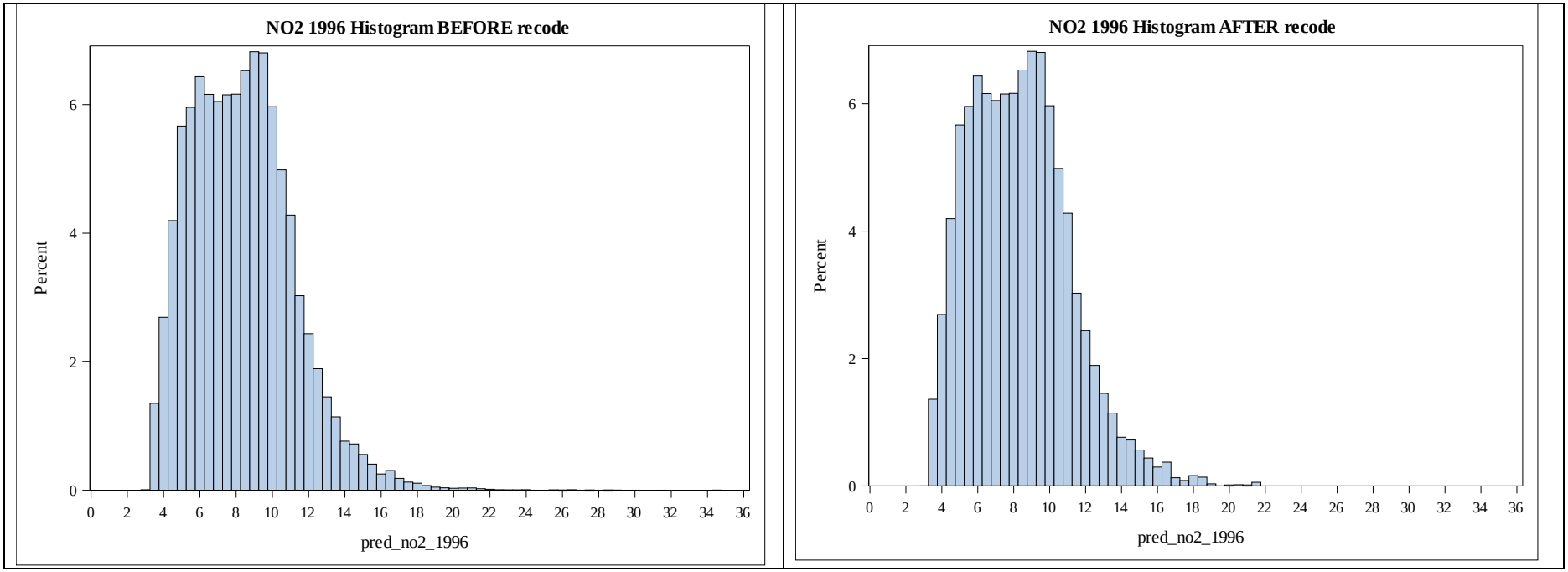
| id | cohort | PRED_NO2 | freq96 |
|---|---|---|---|
| 55 | MID | 3.1 | . |
| 55 | MID | 3.2 | 4 |
| 55 | MID | 3.3 | 11 |
| 55 | MID | 3.4 | 33 |

The value of 3.2 for $NO_2$ in 1996 in the 1946-51 cohort has only 4 instances therefore it was recoded to 3.3 because this was the lowest value less than 3.2 that had at least 5 frequency. The frequency analysis shows the numbers after the recode.

| pred_no2_1996 | Frequency |
|---|---|
| **3.3** | 15 |
| **3.4** | 33 |

**How much recoding went on?**

The numbers recoded were relatively low. There were 323 recoded NO2 1996 values in all the cohorts which was typical for other variables. This is out of a total of 57,346 records. The two histograms below show the distribution of the 1996 NO2 values before and after the recodes. This data is from all cohorts.



## Road Data Recodes

The roads data have six circular buffer variables. The total length of roads in circular buffers with radii of 100, 200 and 500 m. Estimates are presented for all road types ('SUM_ALLROADS_buffer') and major roads only ('SUM_MAJROADS_buffer'). The circular buffer variables were rounded to three decimal points.

The other two road variables were the two straight line distance from each address to the nearest road and major road. Both these were rounded to one decimal point.

The roads data have long upper tails with some low frequencies. These low values could be identifying so we capped each road distance at a maximum value.

**Capped Values / Maximum Values**

| Road Variable | Capped Value | Number of recoded |
|---|---|---|
| Sum_AllROADS_100M | 1 | 728 |
| Sum_MajROADS_100M | 1 | 46 |
| Sum_AllROADS_100M | 4 | 149 |
| Sum_MajROADS_100M | 2 | 253 |
| Sum_AllROADS_100M | 20 | 253 |
| Sum_MajROADS_100M | 7 | 535 |
| ROAD_DIST_ALL | 100000 | 7577 |
| ROAD_DIST_MAJ | 100000 | 82894 |

The number recoded is very large in Road_DIST_MAJ but most of these were rounding rather than capping.

Road Distance All and Road Distance Major had further recodes beyond the capping.

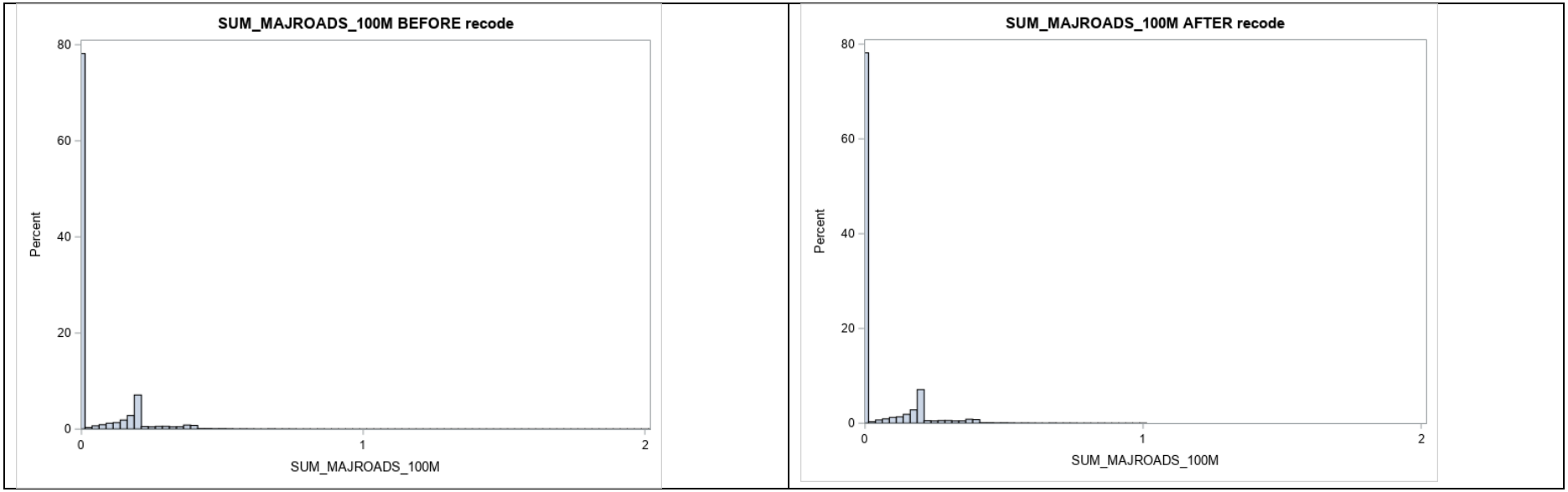
Road Distance All Values were recoded as followed:

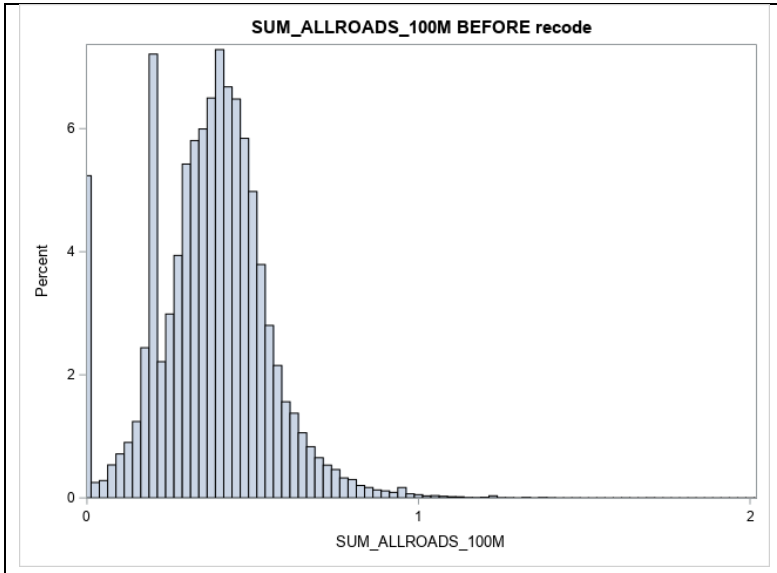
- If greater than 100,000 then capped to the nearest 100,000 (as the table above explained).
- Otherwise if greater than 30,000 then capped at 30,000.
- Otherwise if greater than 1000 then rounded to the nearest 100.
- Otherwise if greater than 200 then rounded to the nearest 1.

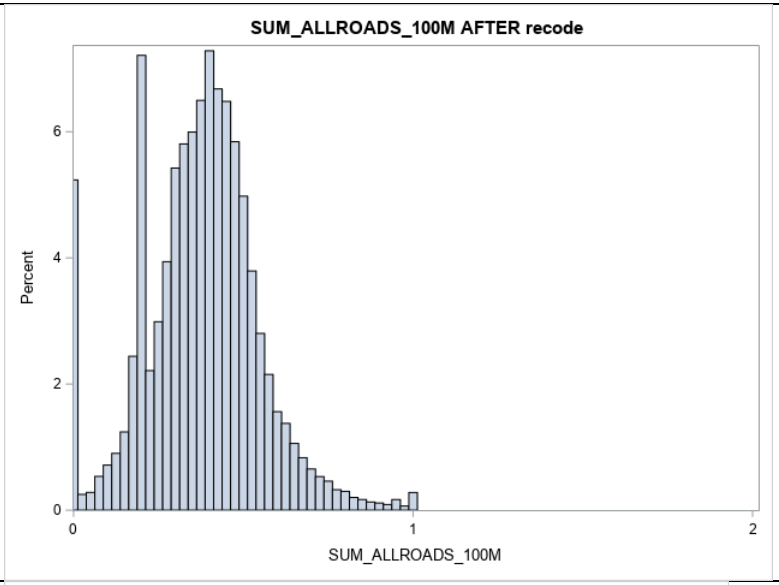
Road Distance Major were recoded as followed:

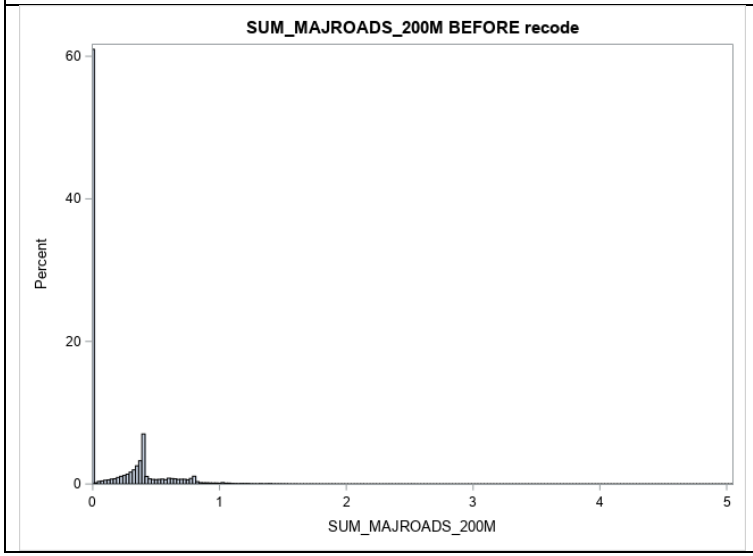
- If greater than 100,000 then capped to the nearest 100,000 (as the table above explained).
- Otherwise if greater than 10,000 then rounded to the nearest 10,000.
- Otherwise if greater than 1000 then rounded to the nearest 10.
- Otherwise if greater than 500 then rounded to the nearest 1.

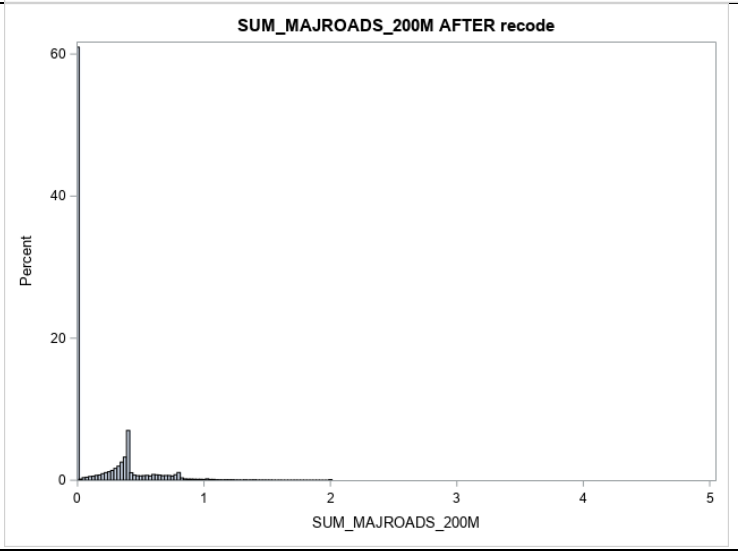
**Histograms of Roads data before and after recodes**

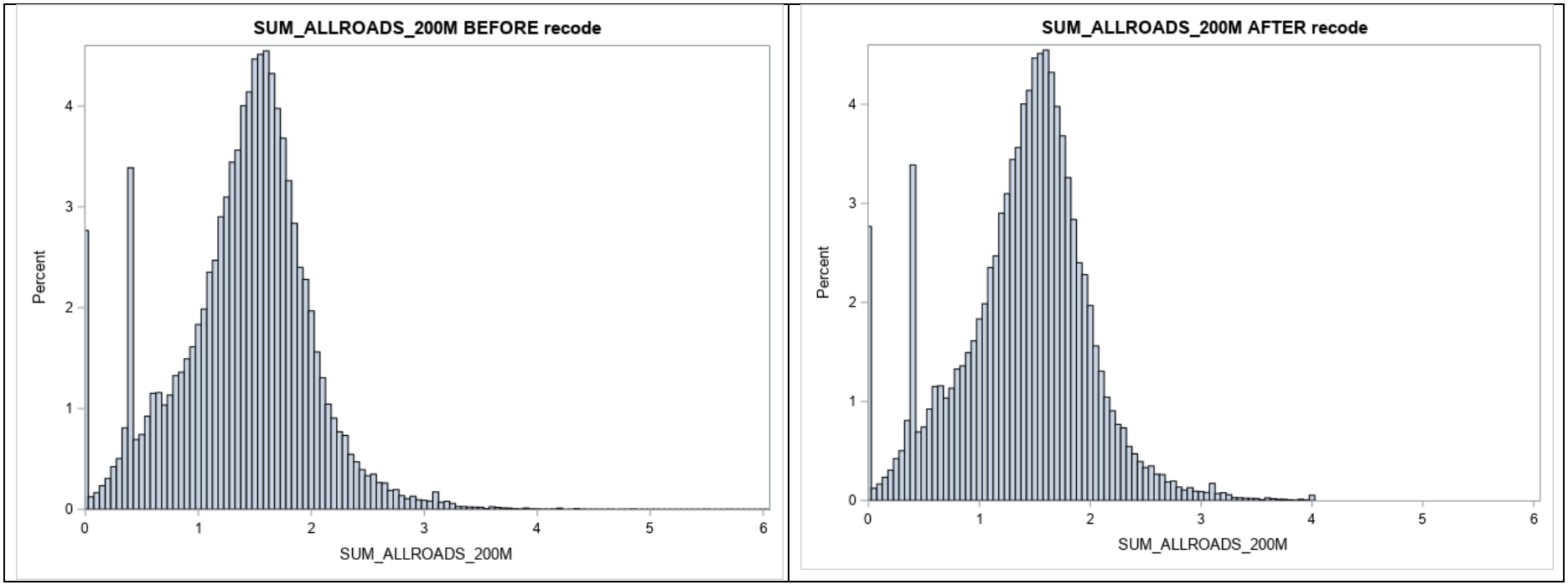


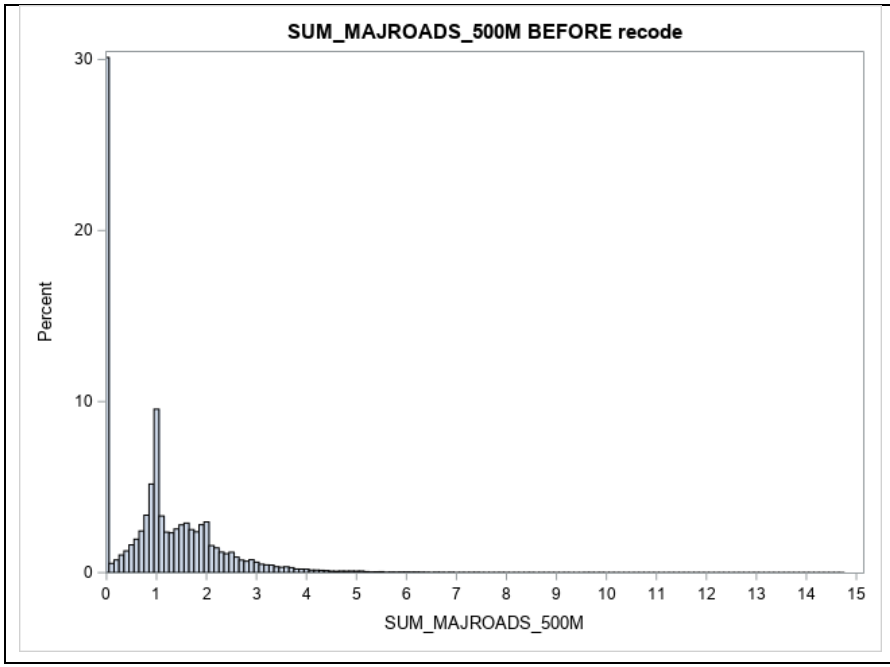

SUM_ALLROADS_100M BEFORE recode
Percent
0
2
4
6
0
1
2
SUM_ALLROADS_100M

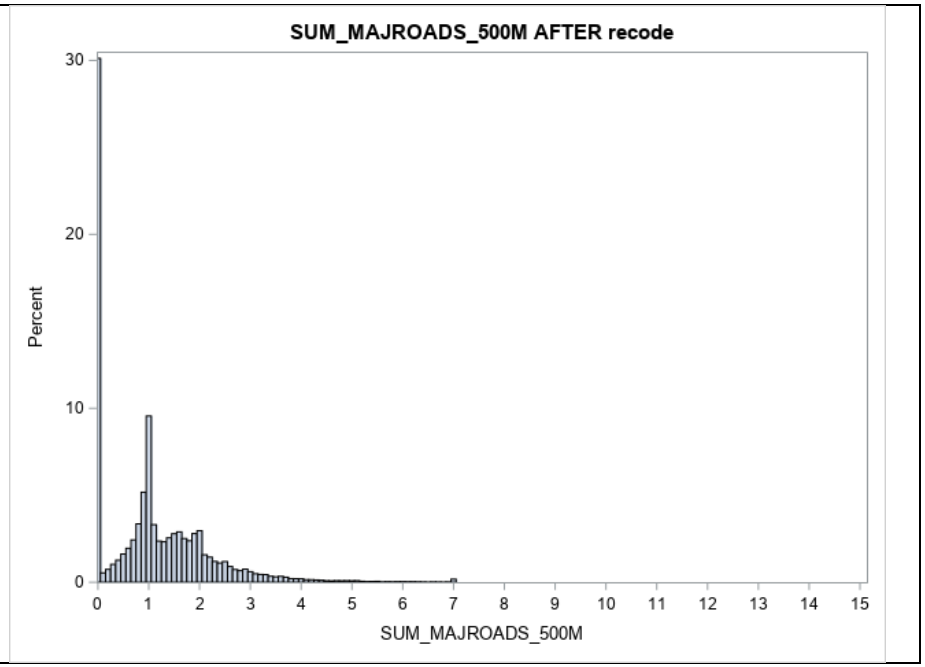

SUM_ALLROADS_100M AFTER recode
Percent
0
2
4
6
0
1
2
SUM_ALLROADS_100M

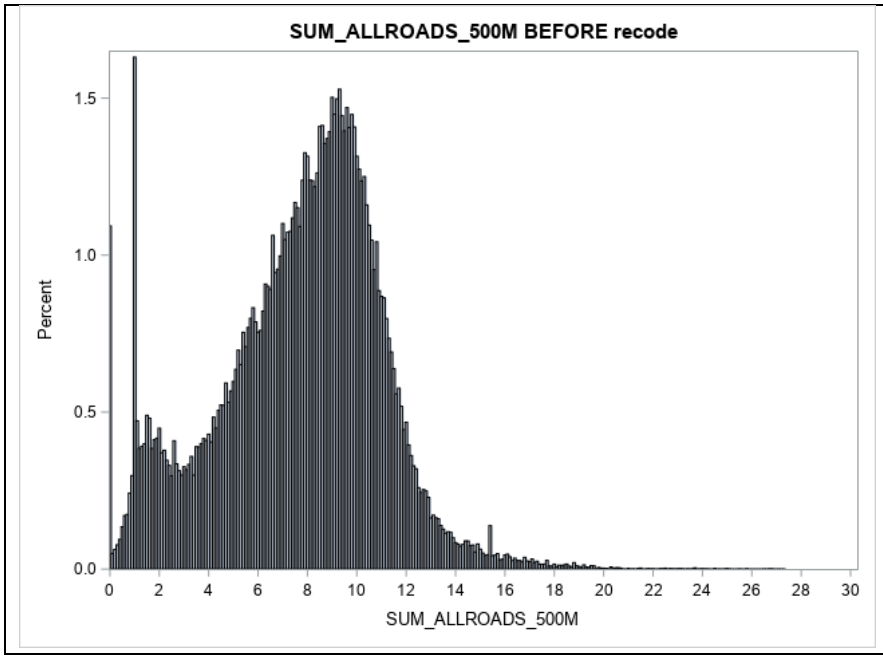

SUM_MAJROADS_200M BEFORE recode
Percent
0
20
40
60
0
1
2
3
4
5
SUM_MAJROADS_200M

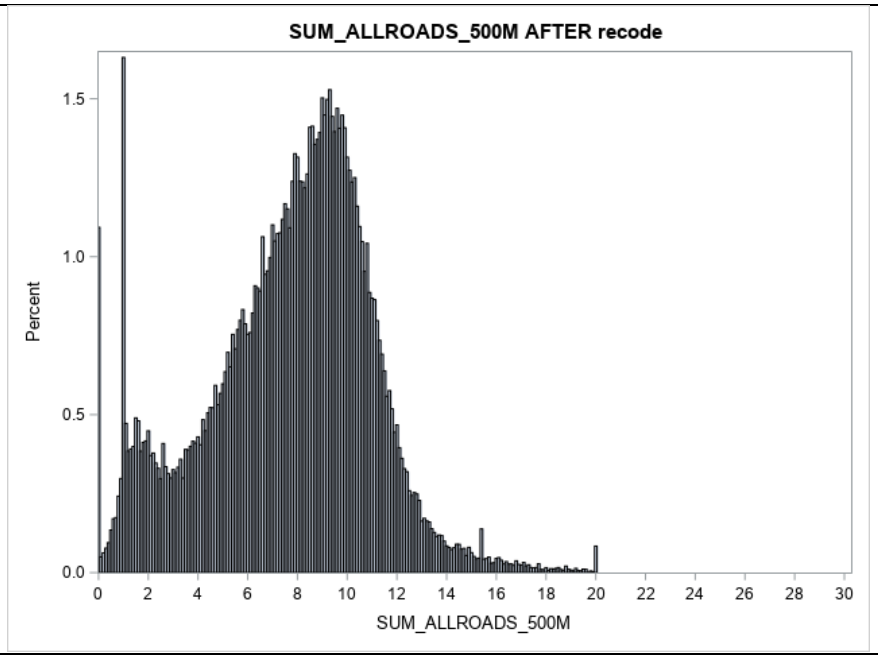

SUM_MAJROADS_200M AFTER recode
Percent
0
20
40
60
0
1
2
3
4
5
SUM_MAJROADS_200M

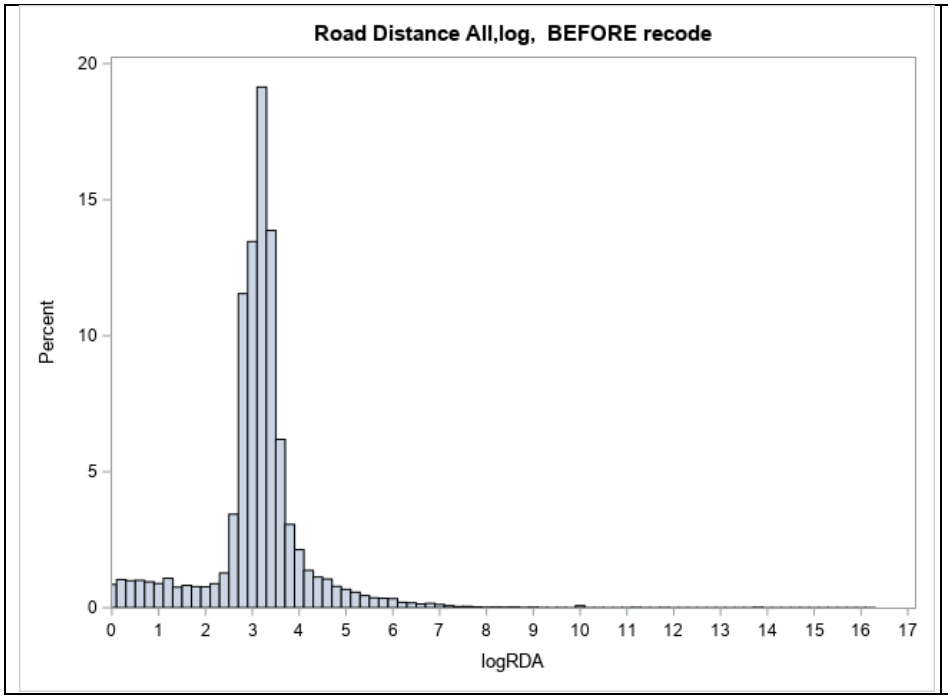

SUM_ALLROADS_200M BEFORE recode
Percent
0
1
2
3
4
0
1
2
3
4
5
6
SUM_ALLROADS_200M
SUM_ALLROADS_200M AFTER recode
Percent
0
1
2
3
4
0
1
2
3
4
5
6
SUM_ALLROADS_200M

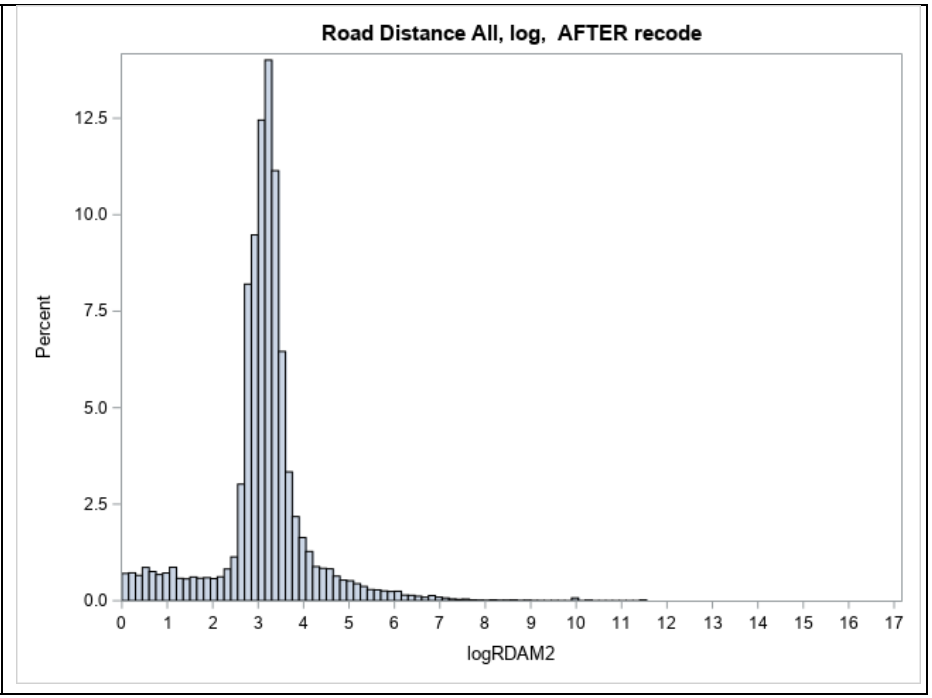

SUM_MAJROADS_500M BEFORE recode
Percent
30
20
10
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
SUM_MAJROADS_500M

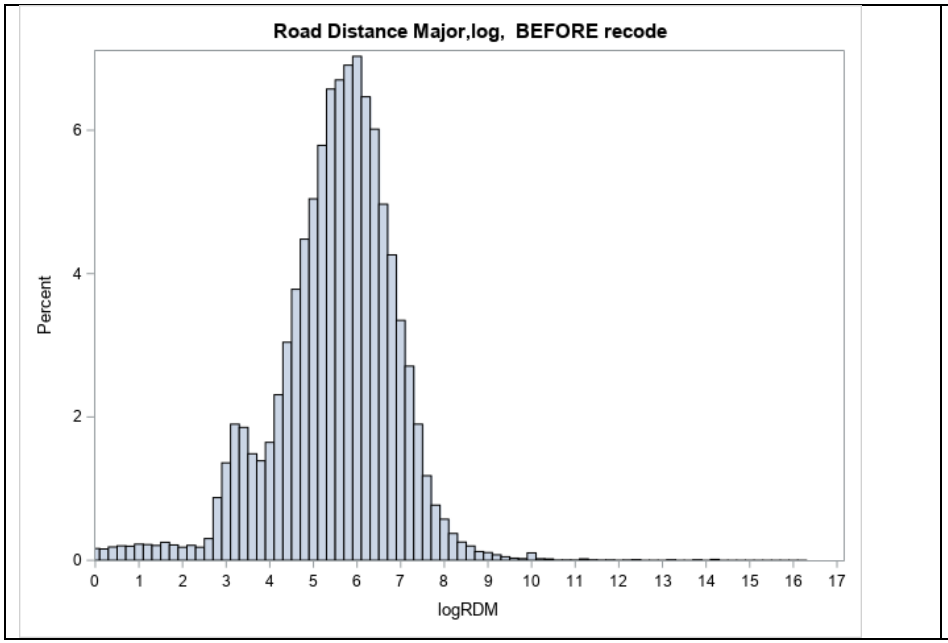

SUM_MAJROADS_500M AFTER recode
Percent
30
20
10
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
SUM_MAJROADS_500M

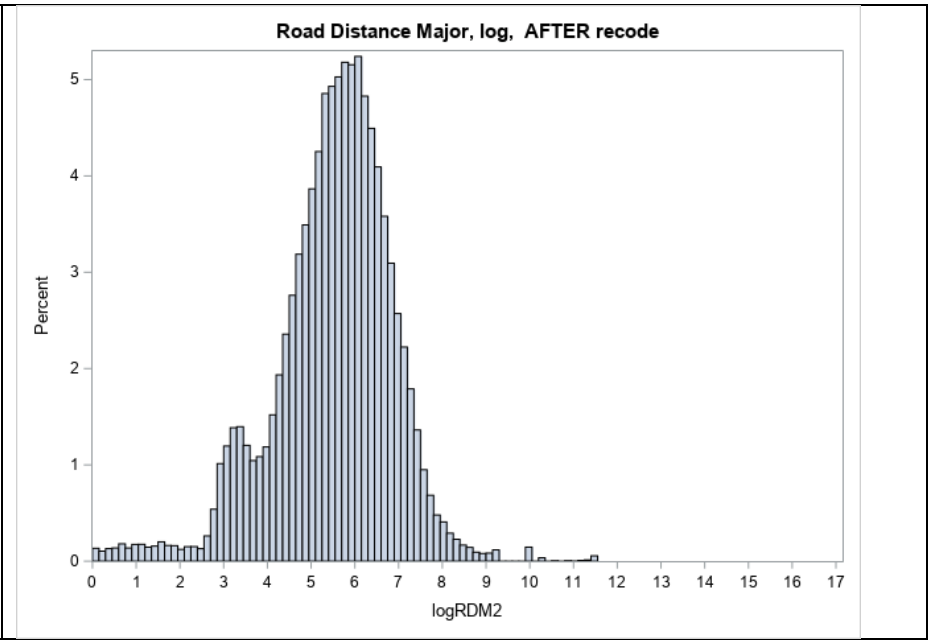

SUM_ALLROADS_500M BEFORE recode
Percent
0.0
0.5
1.0
1.5
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
SUM_ALLROADS_500M


SUM_ALLROADS_500M AFTER recode
Percent
0.0
0.5
1.0
1.5
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
SUM_ALLROADS_500M


Road Distance All,log, BEFORE recode
Percent
20
15
10
5
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
logRDA


Road Distance All, log, AFTER recode
Percent
12.5
10.0
7.5
5.0
2.5
0.0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
logRDAM2


Road Distance Major,log, BEFORE recode
Percent
0
2
4
6
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
logRDM


Road Distance Major, log, AFTER recode
Percent
0
1
2
3
4
5
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
logRDM2