Recodes in Air Pollution and Roads Distance Data

David Fitzgerald August 2020

This document describes how the air pollution and roads data were recoded to avoid any potential identification.

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 not fewer than 5 respondents having any one 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 elsewhere.

These recodes were done for each cohort and for each air pollutant 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 recoding method for low frequency values in both tails was the same but the direction they were recoded was reversed. 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. For the lower tail any 'fewer than 5 frequency' value was recoded to lowest value less than the 'fewer than 5 frequency' value that had at least 5 frequency. These are shown in the following examples.

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 of 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

id	cohort	PRED_NO2	freq96
22	NYC	18.8	1
22	NYC	18.9	1
22	NYC	19.0	2
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

Example Recode in 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

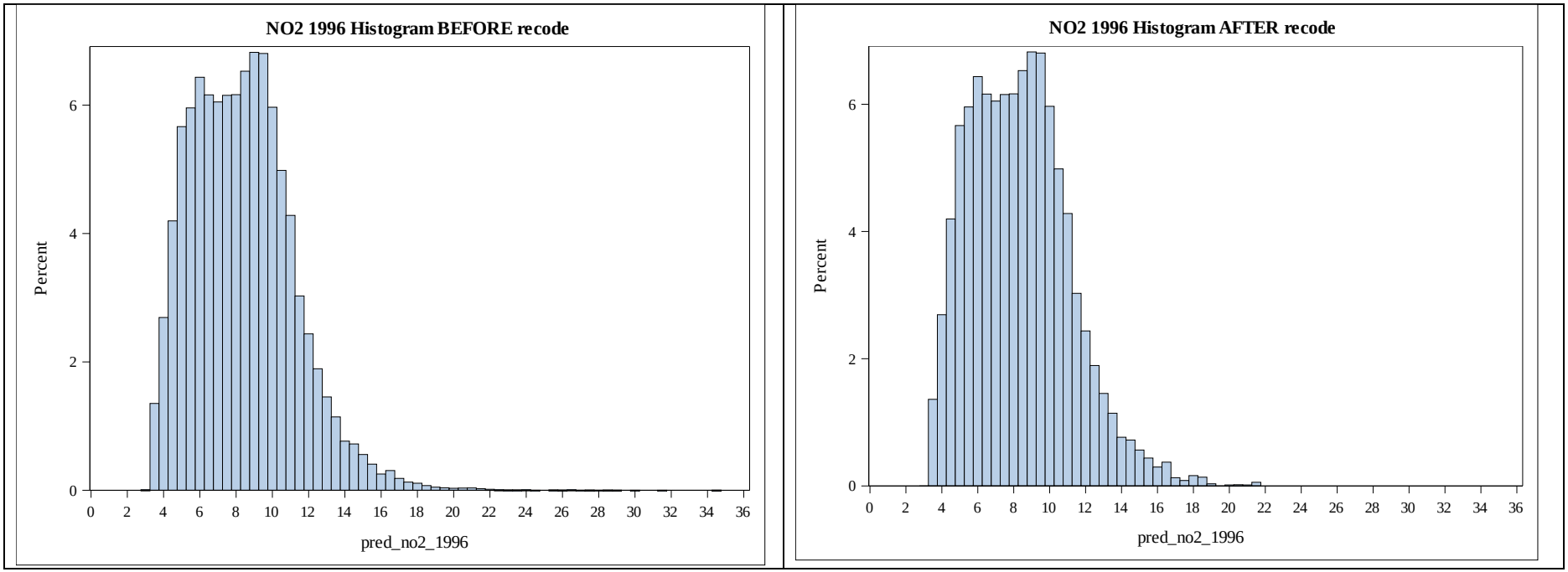
id	cohort	PRED_NO2	freq96
55	MID	3.1	.
55	MID	3.2	4
55	MID	3.3	11

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

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.



I have not shown the histograms for the other years and the other air pollutions but all the histograms show the same very minor changes.

Road Distance Data Recodes

The road distance 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 distance 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 road distance 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 from rounding rather than capping.

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

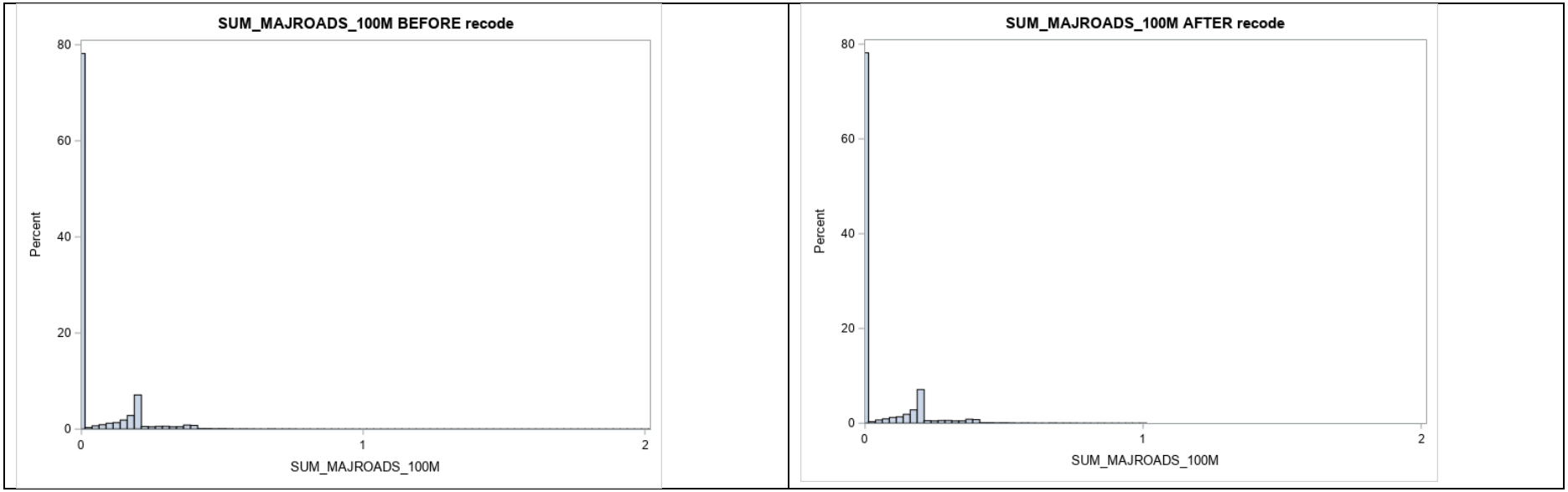
Road Distance All Values were recoded as followed:

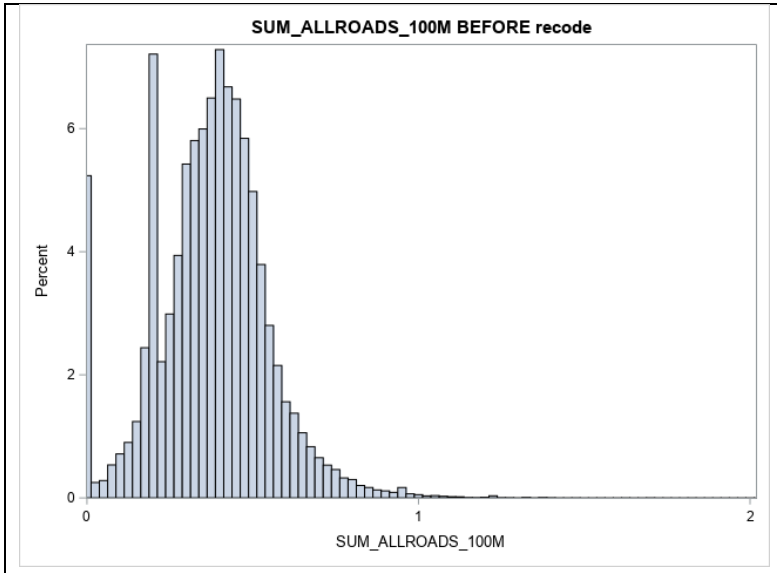
- If greater than 100,000 then capped to the nearest 100,000 (as the table above explained).
- Otherwise if greater than 30,000 then capped at 30,000.
- Otherwise if greater than 1000 then rounded to the nearest 100.
- Otherwise if greater than 200 then rounded to the nearest 1.

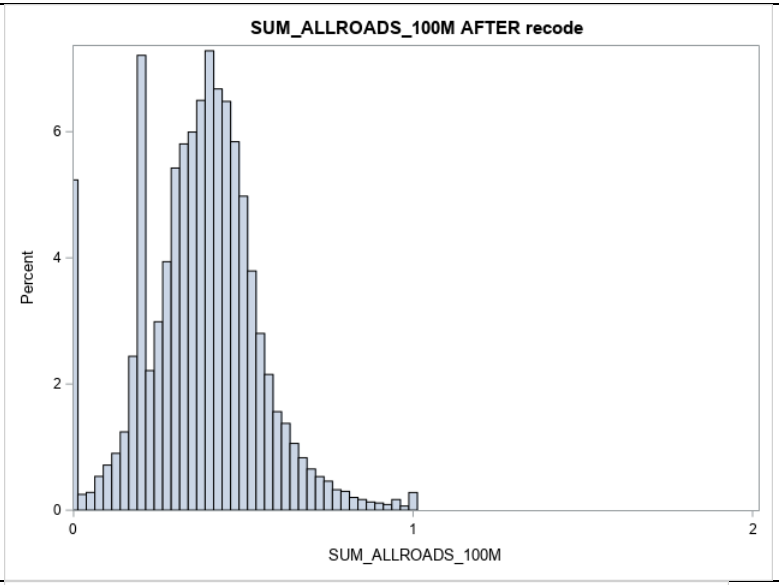
Road Distance Major were recoded as followed:

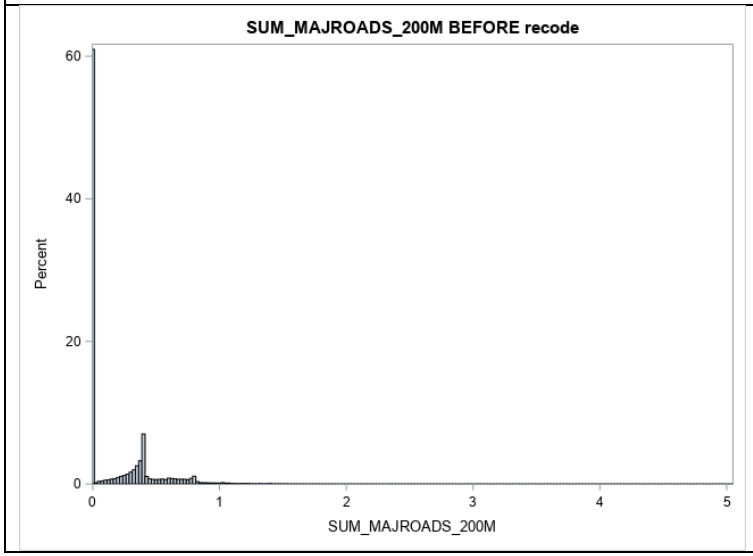
- If greater than 100,000 then capped to the nearest 100,000 (as the table above explained).
- Otherwise if greater than 10,000 then rounded to the nearest 10,000.
- Otherwise if greater than 1000 then rounded to the nearest 10.
- Otherwise if greater than 500 then rounded to the nearest 1.

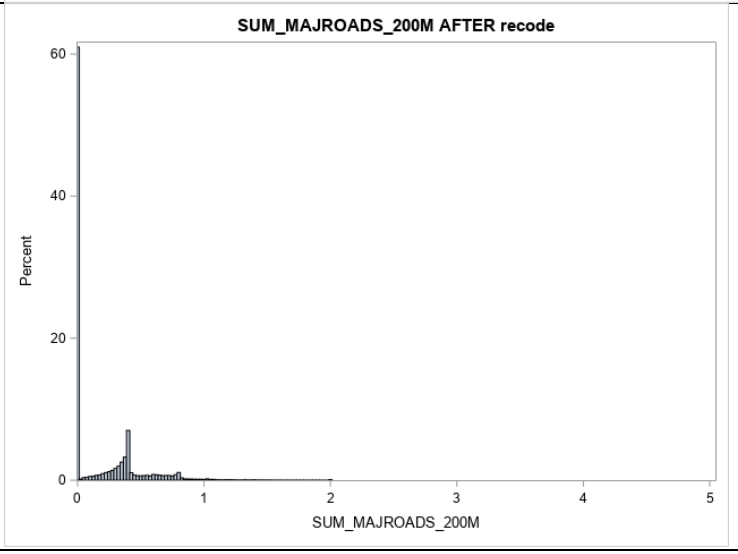
Histograms of Roads data before and after recodes

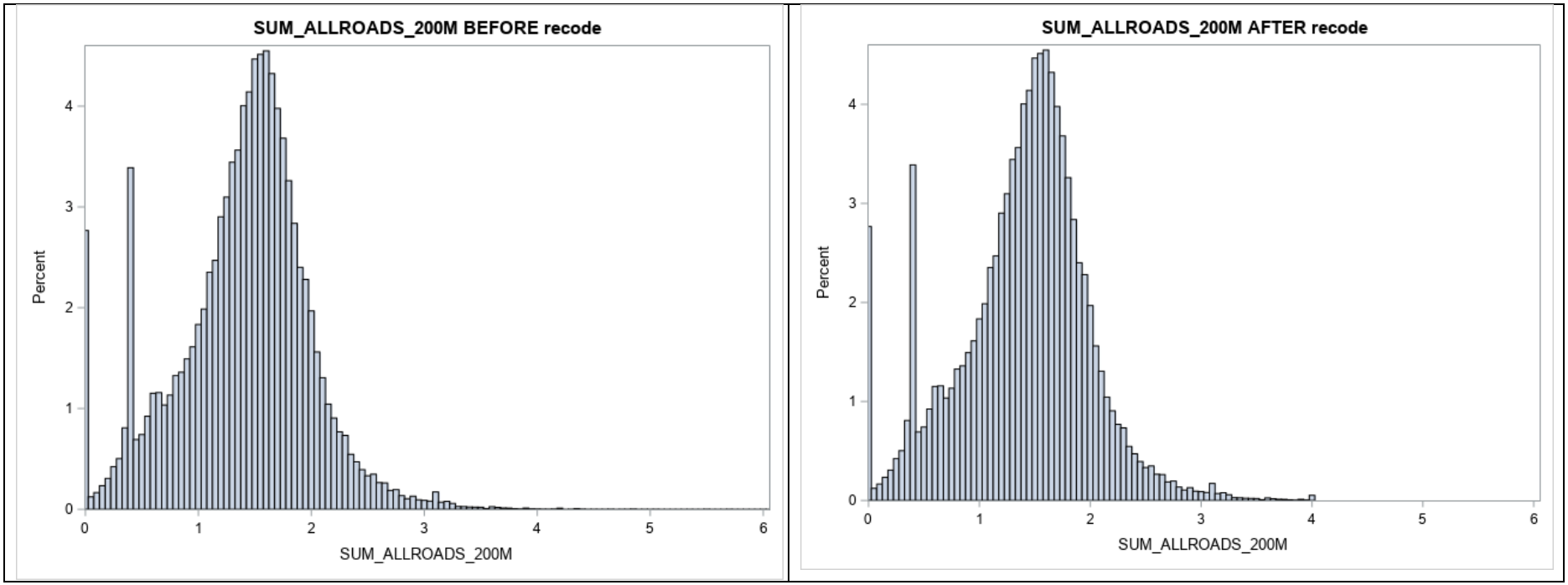


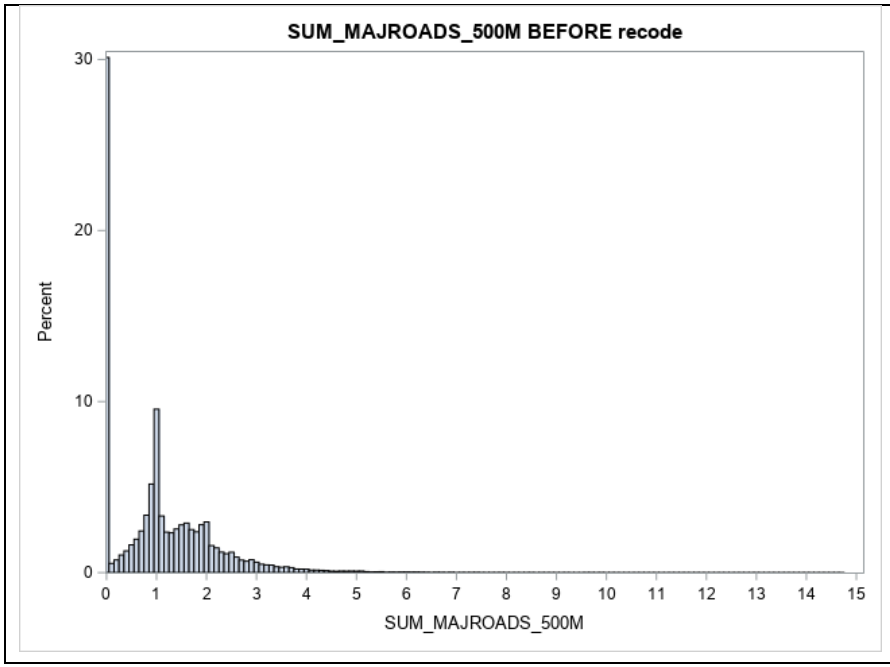

SUM_ALLROADS_100M BEFORE recode
Percent
0
2
4
6
0
1
2
SUM_ALLROADS_100M

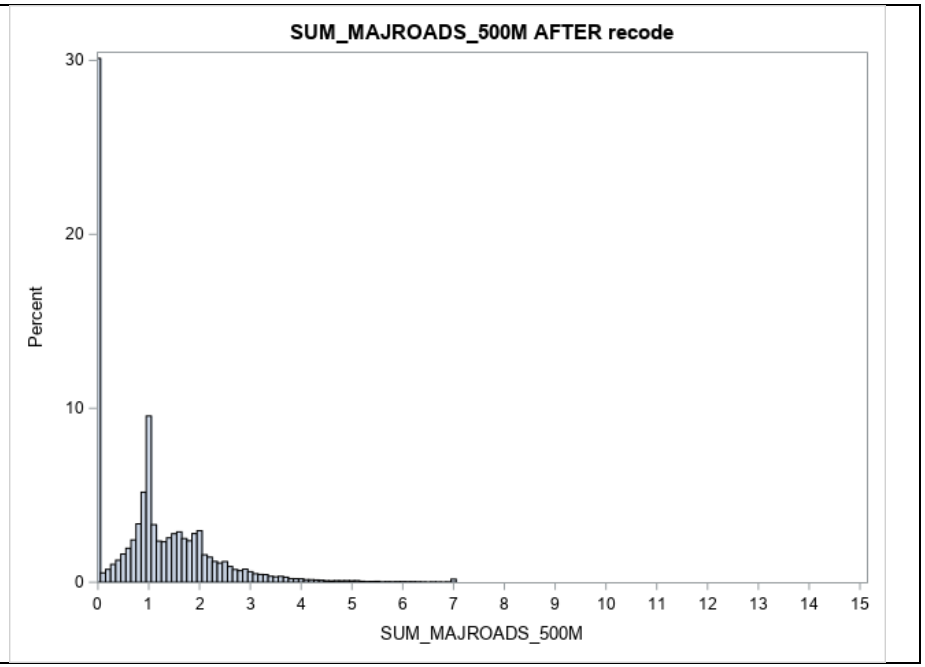

SUM_ALLROADS_100M AFTER recode
Percent
0
2
4
6
0
1
2
SUM_ALLROADS_100M

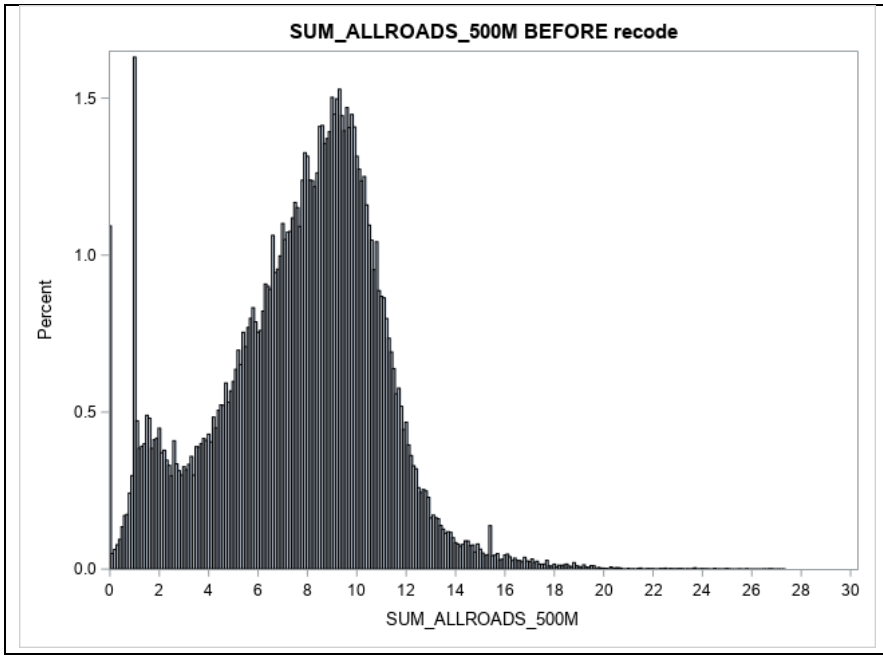

SUM_MAJROADS_200M BEFORE recode
Percent
0
20
40
60
0
1
2
3
4
5
SUM_MAJROADS_200M

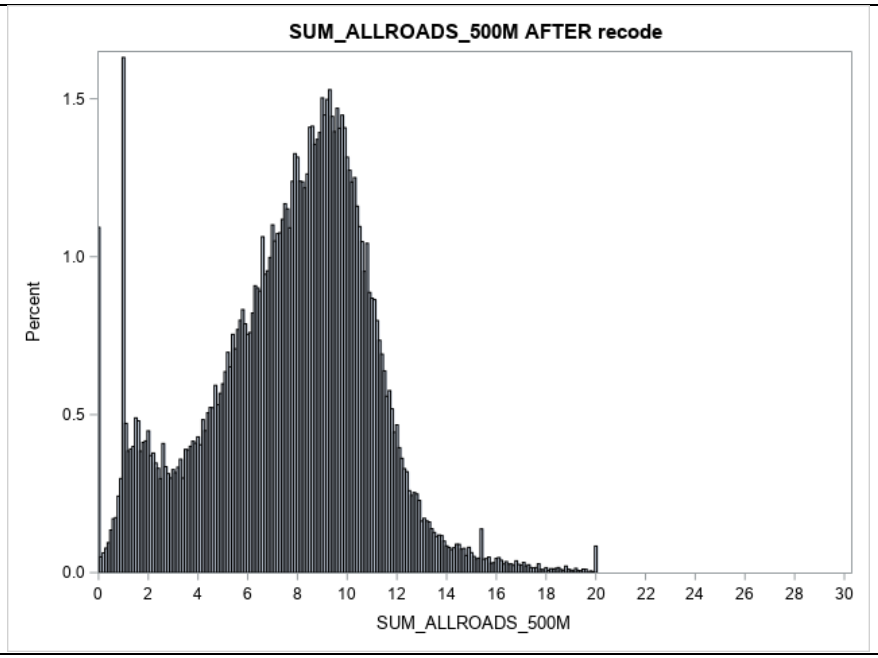

SUM_MAJROADS_200M AFTER recode
Percent
0
20
40
60
0
1
2
3
4
5
SUM_MAJROADS_200M

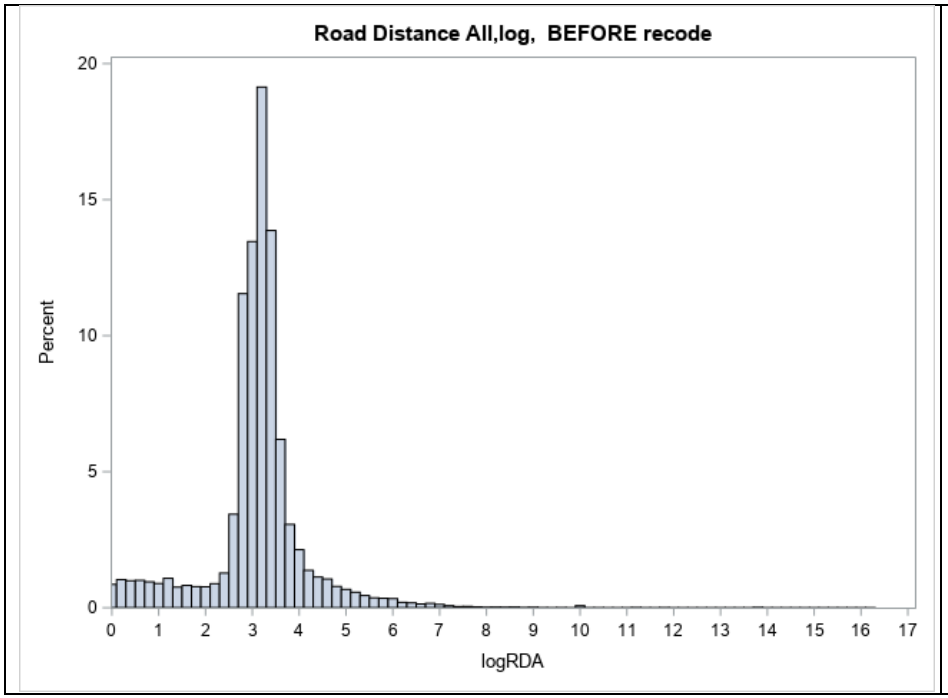

SUM_ALLROADS_200M BEFORE recode
Percent
0
1
2
3
4
0
1
2
3
4
5
6
SUM_ALLROADS_200M
SUM_ALLROADS_200M AFTER recode
Percent
0
1
2
3
4
0
1
2
3
4
5
6
SUM_ALLROADS_200M

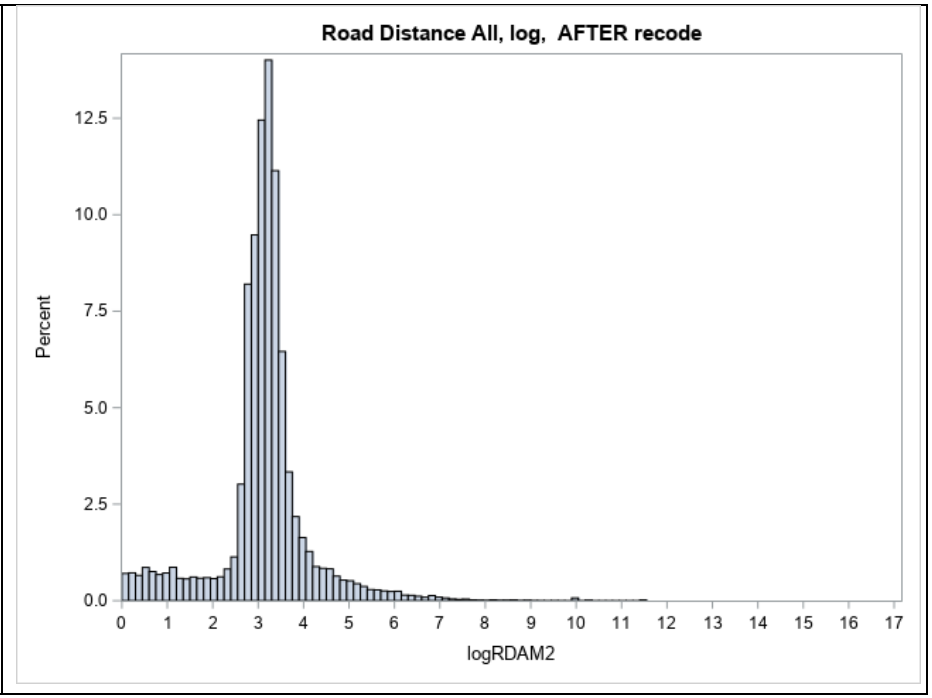

SUM_MAJROADS_500M BEFORE recode
Percent
30
20
10
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
SUM_MAJROADS_500M

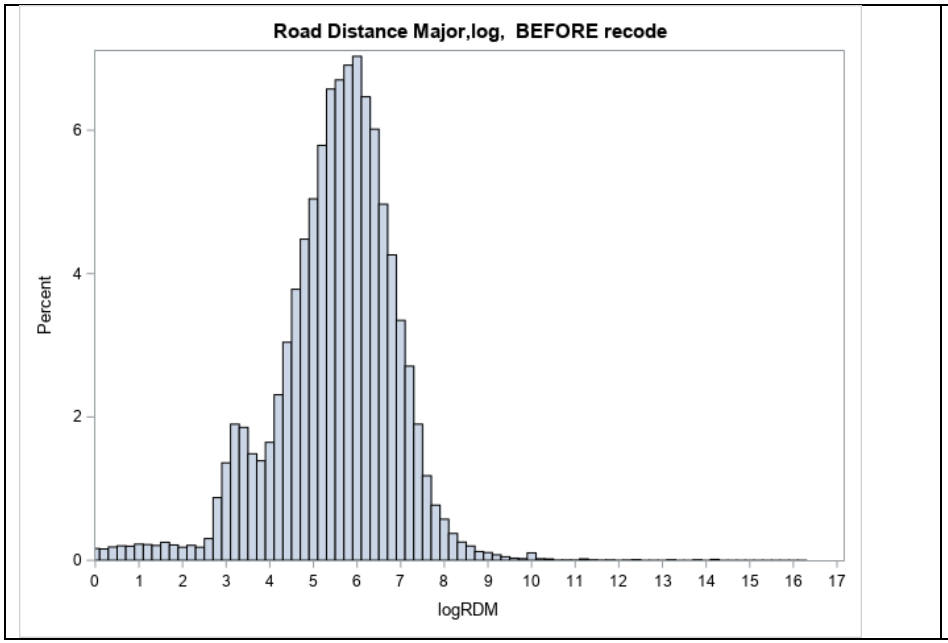

SUM_MAJROADS_500M AFTER recode
Percent
30
20
10
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
SUM_MAJROADS_500M

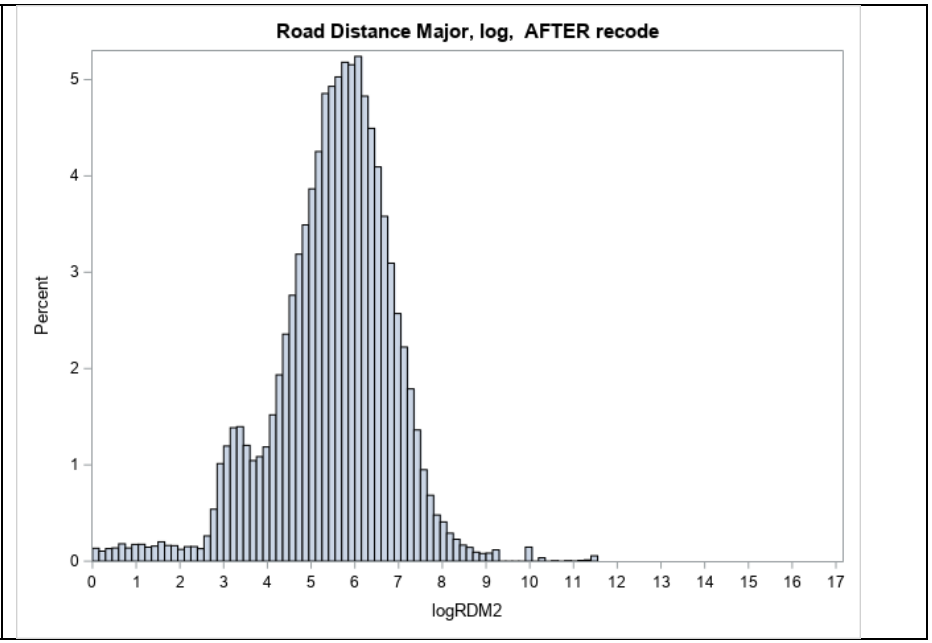

SUM_ALLROADS_500M BEFORE recode
Percent
0.0
0.5
1.0
1.5
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
SUM_ALLROADS_500M


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