Weekly Report 3 Census data and Comparing exposure data sets Raed K Alotaibi Feb/22/2018

What we did this week

Census Data

Downloaded census data tables from the NHGIS website for years 2000/2010 {https://data2.nhgis.org/main} available tables are as following:

Block level data:

- Total population
- Race + Sex + Age_group characteristics.
    - Age_group [<5, 5-9, 10-14, 15-17, 18-19]
    - Race [White alone, Black or African American alone, American Indian and Alaska Native alone, Asian alone, Native Hawaiian and Other Pacific Islander alone, Some other race alone, Two or more races] { How should we reduce Race categories}
- Urban vs Rural

Block group level:

- Household and family Income
- Personal Income
- Educational attainment >25 years of age by sex

A simple note on the data sets:

- When downloading the data we excluded the states of (Alaska, Hawaii) and Porto Rico.
- Each observation (row) for block level data is a census block
    - 8,164,719 observation in the 2000 data
        - Total population (279,583,437)
    - 11,007,989 observations in the 2010 data
        - Total population (306,675,006)
- Each observation (row) for block group data is a census block group
- With each variable the data set becomes increasingly large to handle, some suggestion to work around:
    - Break data sets by states
    - Work with one state, i.e. Texas. Write up the complete code for analysis
    - Apply the code to the remaining states one state at a time.

We also have total population for the year 2010 at block level provided by Mathew (Processed) data set with the following characteristics:

- Each observation (row) is a census block
    - 11,007,989 observation for both 2000/2010 years
        - Total population (306,675,006) for year 2010
        - No population counts for 2000
    - **No missing values** for 2010 population counts variable

NO2 Data

We have multiple data sets with average NO2 levels by {ppb?} at the census block centroid:

- NO2 levels for years 2000/2010 from the Processed files.
- NO2 levels for Annual averages from the website http://spatialmodel.com/concentrations/#no2 for the years 2000 ***through*** 2010

Description of the Data sets

Processed Data set:

The processed Data set provides estimated of average NO2 levels at the census block level (according to 2010) for the contiguous US, 48 states & D.C.. (Excluding Alaska, Hawaii and Porto Rico) for the years 2000 and 2010 and has the following characteristics:

- Each column is variable (“GISJOIN”, “Y2000”, “Y2010”)
- Each observation (row) is a census block
    - 11,007,989 observations for both 2000/2010
        - Average NO2 for 2000 = 9.69
        - Average NO2 for 2010 = 6.32
        - Sum of NO2 for 2000 = 105,547,776
        - Sum of NO2 for 2010 = 68,879,401
    - 114,332 missing values (NA’s)

Annual averages Data set:

The processed Data set provides estimated of average NO2 levels at the census block level (according to 2000) for the contiguous US, 48 states & D.C.. (Excluding Alaska, Hawaii and Porto Rico) for the years 2000 ***through*** 2010 and has the following characteristics:

- Each column is variable (“BlockID”, “Y2000”, “Y2001”, “Y2002”, ……., “Y2010”)
- Each observation (row) is a census block
    - 8,151,762 observations for all years
        - Average NO2 for 2000 = 9.942
        - Average NO2 for 2010 = 6.472
        - Sum of NO2 for 2000 = 81,143,555
        - Sum of NO2 for 2010 = 52,757,333
    - No missing values (NA’s)

Weekly Report 3 Census data and Comparing exposure data sets Raed K Alotaibi Feb/22/2018

Notes on Processed Data set

- Why are there missing values?
- There are 11,007,989 census blocks, equal to the 2010 number of census blocks but more than the 2000 number of census block which should equal to 8,164,719.
    - How did they convert the 2000 census blocks?
- I suggest to work on the 2010 data set from now forward to complete the analysis, since we have the proper reading for all the census blocks for that year.
    - Processed data has all 11,007,989 readings with proper population estimates at block level that can be joined with the NHGIS demographic data.
    - We don't have proper counts for the 2000 data (more than 8 million observations), thus until we fix this we cannot join the NO2 observation with the NHGIS demographic data.

Notes on *Annual averages* Data set

- Why are there only 8,151,762 observations (census blocks) while the total number of blocks in 2000 is 8,164,719. In essence we have 12,956 blocks not estimated?
- All years have the same number of blocks based on the 2000 block numbering?
    - However, this is not equivalent to the NHGIS number of blocks for year 2000, we cannot, therefore join demographic data for year 2000 with the NO2 data!

Comments and Questions:

- We have the proper data sets for census and demographic data for both years at block and block group level (check)
- We are still facing some problems with the exposure data sets:
    - Processed data:
        - We have proper population counts by block for 2010 but none for 2000.
        - We can join this data set with the NHGIS dataset even with some missing exposure estimates.
        - Why are there missing values for exposure estimates?
    - Annual averages data:
        - The number of observations does not match with the number of observations (rows) from NHGIS data bases? Where are the remaining 12,956 blocks?
    - When comparing average NO2 levels for years 2000/2010 from both data sets there is some discrepancy, why? i.e.
        - Processed:
            - Average NO2 for 2000 = 9.69
            - Average NO2 for 2010 = 6.32
        - Annual Average:
            - Average NO2 for 2000 = 9.942
            - Average NO2 for 2010 = 6.472
    - Finally, I think we are good to go using the 2010 Processed data set to complete our analysis (NO2, demographics …etc.) if the reading are the correct. however, for year 2000 we cannot join together NO2 with demographic data sets due to differences in the number of observations (unless we add missing values to the NO2 data set)