Data Report

Antonio Jurlina

Here, I will go through each variable that was created, report on how it was done, where the data came from, whether any observations are missing and what might have caused it. Finally, I will add recommendations on how to proceed.

Starting notes:

- Each variable is reported on the county level.
- You will need an API key to access the American Community Survey data through RStudio and you can apply for it here.
- I wrote some functions that help the process along, and they can be found at the beginning of the script. They help me clean the FIPS codes and deal with some missing data on the spot (and this was just NAs introduced by me, through many joins).
- FIPS codes are presented in the 5-digit format which consists of the State FIPS (first two digits) and the county FIPS (remaining 3 digits) and they serve as unique identifiers for each location.
- Each variable, in each county, is reported for both 2012 and 2017.

1. American Community Survey (ACS) Variables

a. Population

- These data were posted to the ERS website (May 2020). Contact: John Cromartie, ERS/USDA, john.cromartie@usda.gov, (816) 994-4302.
- I removed counties from Alaska, Hawaii, Puerto Rico and the totals for the entire U.S.
- This data set serves as the backbone to which all others are joined, given that it contains all counties in the contiguous United States.

b. Education

- To get this data, I used a government provided API key and the *acs* library in R.
- The variables I obtained (as percentage of county population) are:
  - High school diplomas
  - GED or alternative
  - Some college, under one year
  - Some college, over one year
  - Associate degrees
  - Bachelor's degrees
  - Master's degrees
  - Professional (law school, med school, etc.) degrees
  - Doctorates
- Out of those, I constructed three variables of interest:
  - High school or GED = High school diplomas + GED or alternative
  - Some college or associates = some college, under one year + some college, over one year + associate degrees
  - Bachelor's or higher = bachelor's degrees + master's degrees + professional degrees + doctorates

2. Distance of County to Land Grant Universities

- I started by creating a data set of all land grant universities in the United States by going through the government provided list and finding coordinates for each location using GeoHack (e.g. here is the page for UC Davis).
- The Coordinate Reference System for this data set was projected using EPSG 4326, popular in the United States and commonly used by Google Earth and the U.S. Department of Defense for all their mapping, tends to be used for global reference systems. GPS satellites broadcast the predicted WGS84 orbits which are used for this system. For more information, here is a useful pdf.
- For each university I also noted what type of land grant it received:
    - 1862 – initial Morrill Acts
    - 1890 – African American institution Morrill Acts
    - 1994 – tribal land grants
- Subsequently, I filtered only for the universities that received the 1862 grants and proceeded to do further calculations with them.
- In order to calculate the distance from each county to the nearest land grant university, I needed to calculate the centroids for each county, and before that, I needed the county shape files, which I obtained from the U.S. Census website.
- In order to calculate the centroids and the distances, especially using the *sf* library in R, I needed to convert all the geographic coordinate systems (like EPSG 4326 data mentioned above and the county shape files) to a projected coordinate system which would treat them as if located on a flat *xy* grid. For this, I used EPSG 26919, the commonly used North American projected coordinate system.
- Finally, for each county, I calculated the distance to the nearest overall land grant university and the distance to the nearest in-state land grant university.
- The distances are reported in miles.
- I also created a Boolean variable indicating the presence of a university for every county.

3. National Agricultural Statistics Service (NASS) Census

- I got two *.txt* files of data from the census, one for 2012 and one for 2017, and they include all the variables.

a) Normalization Variables (County Data)

- First, I extract three variables that, while not relevant for the final data set, are to be used in creating it. They are:
    - Number of operations
    - Number of acres
    - Median acres per operation
- Each is to be used in normalizing subsequent variables (i.e., dividing by the number of operations to get the percent of operations, dividing by the number of acres to get the percent of acres being used, etc.).
- These variables have been extracted from the census and left for the next person that takes over because it is important to decide what needs to be done with missing variables (more on this later) before any actual normalizations are performed.

b) Conservation and Federal Program Participation

- For each county, I get the data on the receipts ($ / operation) and the number of operations participating in Conservation and Federal programs.
- These need to be normalized using normalization variables.

c) Organic Operations

- I extracted the number of operations that were classified as organic by certification, and the number of operations that were classified as organic but exempt from certification, and I summed them into a single variable called organic, which indicates the total number of organic operations per county.
- Once the missing observations are dealt with, this variable needs to be divided by the total number of operations per county to get the percentages.

d) <u>Crop Insurance Participation</u>

- Here, I extracted the number of farms participating in crop insurance.
- Once the missing observations are dealt with, this variable needs to be divided by the total number of operations per county to get the percentages.

e) <u>Operating Expenses</u>

- Here, I extracted the average amount of $ / operation in each county.
- Once the missing observations are dealt with, this variable needs to be divided by the median acres per operation per county.

f) <u>Net Farm Income</u>

- Here, I extracted the net income, measured as the average amount of $ / operation in each county.
- Once the missing observations are dealt with, this variable needs to be divided by the median acres per operation per county.

g) <u>Labor Costs</u>

- Here, I extracted two variables for each county – the total expenses, measured in $, for labor that was hired and labor that was contracted.
- Once the missing observations are dealt with, these two variables need to be summed and the resulting variable needs to be divided by the median acres per operation per county.

h) <u>Fertilizer and Chemical Totals</u>

- Here, I extracted two variables for each county – the total expenses, measured in $, for fertilizer and chemical usage.
- Once the missing observations are dealt with, these two variables need to be divided by the median acres per operation per county, each.

i) <u>Grazing Land</u>

- Here I extracted two variables for each county, that represented the total grazing area in acres.
- Once the missing observations are dealt with, these two variables need to be summed and the resulting variable needs to be divided by the number of acres per county.

j) <u>Practices</u>

- Here, for each practice (no-till, cover crop, tile drainage) I extracted the number of operations, the total acres and acres / operation, for each county.
- 2017 data came directly from the census, but the 2012 data came from a separate excel file that had to be requested from the Census Bureau in order to get county level data

k) <u>Female Producers</u>

- Here I extracted the variable detailing how many operations per county had a female principal produced.
- Once the missing observations are dealt with, this variable needs to be divided by the number of operations per county.

l) <u>Non-White Producers</u>

- Here I extracted variables for the number of operations with principal producers identifying as Hispanic, African American / Black, American Indian or Alaska Native, Hawaiian or Pacific Islander or Multi Race and I summed them all into a single variable called Non-White.

- Once the missing observations are dealt with, this variable needs to be divided by the number of operations per county.

m) Beginning/New Producers

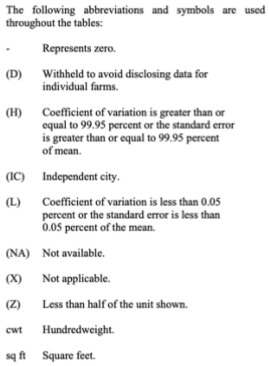
- This variable indicates all operations within a county that have principal operators with less than 10 years of experience.
- For 2017 data, I summed the variables Under 6 Years Exp and Between 6 and 10 Years Exp to create one indicating Beginning Producers.
- For 2012 data, I summed the variables Under 3 Years Exp, Between 3 and 4 and Between 5 and 9 to create one indicating Beginning Producers.
- Once the missing observations are dealt with, this variable needs to be divided by the number of operations per county.

n) Age Brackets

- This variable splits principal operators into age brackets.
- Since the brackets didn't match between the 2012 and the 2017 census, I had to reorganize them so that the resulting final brackets encompass all the available data.
- This resulted in the creation of the following brackets:
    - Under 25
    - 25 to 34
    - 35 to 44
    - 45 to 54
    - 55 to 64
    - 65 and Over

4. Missing Data

- The script I wrote in R will produce a "*missing_data*" data frame, which summarizes the situation.
- The data frame lists the total number of observations, which is 6216. This is the number of all contiguous counties for 2012 and 2017 (so double the number of counties).
- Next, it lists how many total missing values there are. These are further subdivided by origin.
- NAs that were introduced are the ones that did not exist in the census data but were introduced by me after joining them to the entire list of counties of interest. This means some of these counties were not reported for in the census. These can either be set to 0 or will need to be imputed. I am not sure. For example, the variable indicating the number of organic operations has over 2000 NAs introduced and this seems to be simply from the fact that many counties had no organic operations so were excluded from that portion of the census. I believe that missing values indicate 0 in this case. This is further confirmed by looking at Figure 1 which I copied from the census report and they seem to state that a missing value ("-") represents zero.



The following abbreviations and symbols are used throughout the tables:

-	Represents zero.
(D)	Withheld to avoid disclosing data for individual farms.
(H)	Coefficient of variation is greater than or equal to 99.95 percent or the standard error is greater than or equal to 99.95 percent of mean.
(IC)	Independent city.
(L)	Coefficient of variation is less than 0.05 percent or the standard error is less than 0.05 percent of the mean.
(NA)	Not available.
(X)	Not applicable.
(Z)	Less than half of the unit shown.
cwt	Hundredweight.
sq ft	Square feet.

Figure 1.

- The next variable in this data frame of missing values indicates how many non-numeric ones there are. In most cases these are all (D) (see Figure 1) except for Chemical Expenses which also includes some (Z)'s.
- These will most likely have to be imputed.

5. Other

- After I account for missing data, there is a short section that attempts to place all the zipcodes to their respective counties, but since we had issues with that data set and I was running out of time, I left it as is.