**Driving forces for agricultural changes: understanding the urban/rural development and climate contribution**

**Arash Modaresi Rad, Mostofa Najmus Sakib, Ahmad Hojatimalekshah**

**Project Goals:**

* Understanding the effect of urban development on agriculture in Idaho
* Understanding the effect of climate and weather on agriculture in Idaho

The project started with the goal of understanding the effect of urban development on agricultures in Idaho. However, the initial problem was to select the exact dataset/variables to continue the analysis. The USDA has agriculture data for in a 5-year interval from 2017 to 1987. We planned to use all the available years to check the temporal variation as much as possible for the study area. Through proper skimming and playing around the website we initially fixed some of the variables that can take us to the project goal. So, we looked for the dataset for our initial selected variables in the USDA website (<https://www.nass.usda.gov/Quick_Stats/CDQT/chapter/2/table/1/state/ID/year/2017>)

There are 2 formats of data available there. One of them is the pdf and the other is csv. We started with the csv file but unfortunately the csv files didn’t have the exact same data with the pdf files. The county wise division of the data was only seen on the pdf file.

A close up of text on a white background

Description automatically generated

Fig 1: USDA data in csv format

A close up of text on a white background

Description automatically generated

Fig 2: USDA data in pdf format

Both the above screenshot is for the same selection of data on the USDA website. Since the csv only has limited dataset, we had to look for the pdf files. We have faced a lot of difficulty in our analysis because of the pdf format of the data.

1. The first step that we took was to resolve the data extraction problem, so we contacted Randy Welk (the Idaho State Statistician with the USDA National Agricultural Statistics Service) asking for csv files with all the county-wise data which are available in the pdf file.

A screenshot of a social media post

Description automatically generated

This is the reply we received for our query, so we had no way left except for using all the pdf files that contain all the data. Unfortunately, The process of reading PDF files and converting them to CSV files was so difficult due to unorthodox format of the tables and that slowed our worked a lot.

1. We started our data extraction by doing the pdf to excel conversion systems in the python. We tried PyPDF2, tabula-py modules for the extraction works. None of them was able to convert the data in our required format. The tabula-py was able to separate the rows and columns but the column orientation didn’t match.

A screenshot of a cell phone

Description automatically generated

We tried to paste this data in an excel file but the tab/comma/space no special paste option worked. So, we had to skip the python option.

1. Later we tried matlab. The ***readPDFFormData*** function seemed to be the option for converting the pdf file into excel but it required Text Analytics Toolbox which was not available through the free matlab license under the Boise State University.

1. We also looked for the possible solutions in R. The following code was used to convert the pdf file to csv formats.

A screenshot of a social media post

Description automatically generated

This code initially solved our problem for the recent years (2017, 2012, 2007 and 2002). But for earlier years like (1987, 1992, 1997) we encountered more problems. The data was separated by spaces between the hundreds and thousands places unlike the commas in the recent years. So, while pasting the data in the excel again because of the space the thousand digits got separated from the other places. Also this resulted in removal of zeros that were right after the spaces, thus returning wrong numbers.

A screenshot of a cell phone

Description automatically generated

Also, the 0 values were not available while we pasted in the excel. So, for 5049 we got 5 in one cell and 49 in another cell with 0 being removed.

1. To extract the data for the previous years from the pdf files we then went for the simpler way. We tried to convert the pdf files through the freely available online pdf converter. We have tried multiple converters but none of them did the job. Couple of them couldn’t convert it, and some of them put the whole dataset as an image inside the excel file. The one that worked better was able to put the data in the excel file but the whole column got stuck in only one cell (only one row). So, while we were trying to copy the cell to split the rows, different pasting formats didn’t work. Finally, we selected each column one by one and pasted in notepad. From there we again copied and pasted in the excel file to finally separate the rows. This had to be done for each column one by one. All this work took a quite handful amount of time.
2. Another problem we got into was the subcategories in the variable names. This was a problem irrespective of the data was extracted from R or pdf converter. The subcategories are described in the figure below from the pdf file.

A screenshot of a cell phone

Description automatically generated

Since the subcategories had the same name under multiple variables (i.e. Number/ Average size of farm/average per farm/ average per acre) and it was not possible to distinguish between them we had to manually rename the variables. For example, we added the main variable before the “Number” subcategory to separate between the different poultry number and cattle number. We manually did for all the variables.

1. Once the variables were renamed, we started combining the data for all the years. Unfortunately, the variables weren’t available in the exact same sequence for all the years. We have accumulated all the variables in one excel file for different years for purposes of comparison.

A close up of text on a white background

Description automatically generated

In the above picture the red ones are the rows where the same variable names didn’t match exactly in each year, so we had to come with one unique name for that variable in each year to use/select in the analysis part. The red ones are the variables which are available for one year but not for the others. Here we had to go through each row to select the variable that is available for all years and can be used as an impactful variable for our project goal.

Finally, we have the datasets in the proper format and order for 7 time periods. Our next steps will be as follow:

1. Merging these datasets as one
2. Finding the correlation between Census data and USDA variables
3. Finding the correlation between meteorological data and USDA variables
4. Analysing the spatial and temporal variations in agriculture in Idaho