USDA

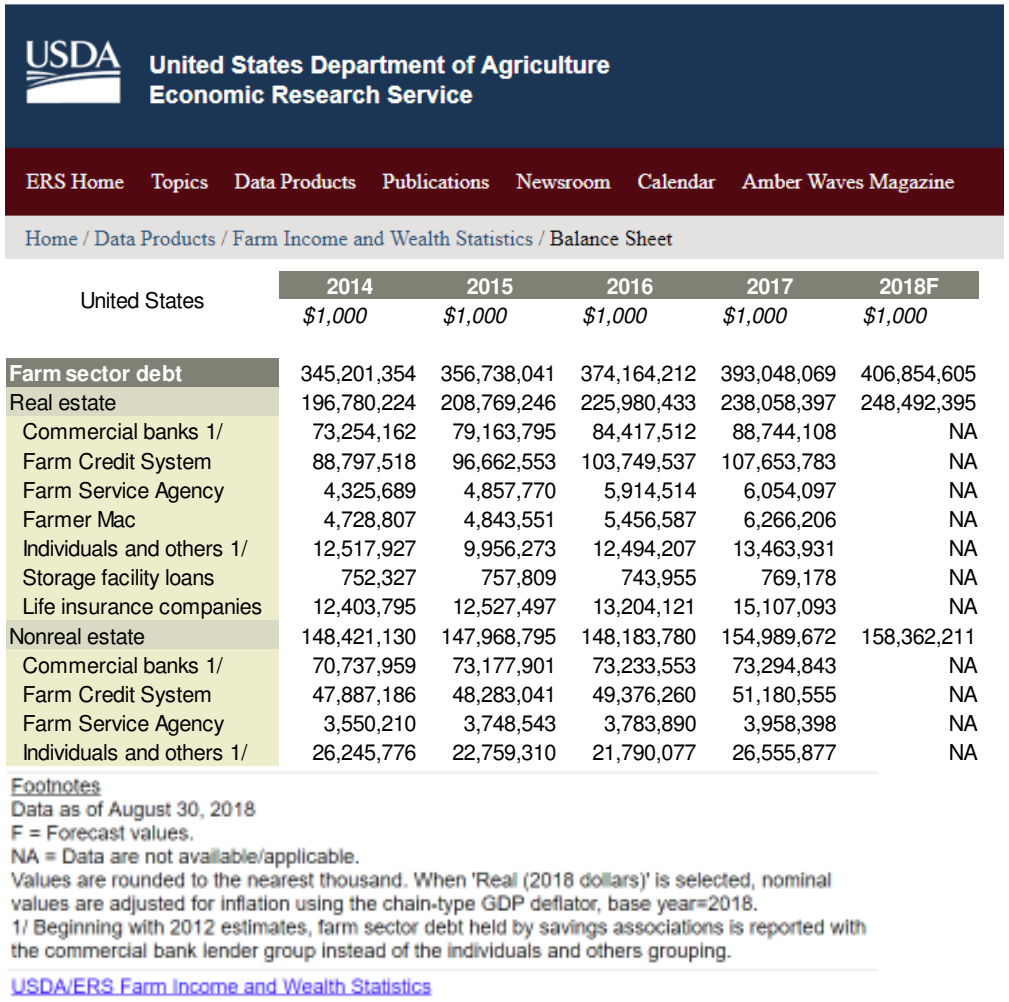
United States Department of Agriculture

Linking Public Data Sources to Create Localized Official Statistics

Greg Lyons & Dipak Subedi
Economic Research Service
October 24, 2018

The Findings and Conclusions in This Preliminary Presentation Have Not Been Formally Disseminated by the U. S. Department of Agriculture and Should Not Be Construed to Represent Any Agency Determination or Policy. This research was supported by the intramural research program of the U.S. Department of Agriculture, Economic Research Service.

Motivation



USDA United States Department of Agriculture
Economic Research Service

ERS Home | Topics | Data Products | Publications | Newsroom | Calendar | Amber Waves Magazine

Home / Data Products / Farm Income and Wealth Statistics / Balance Sheet

United States	2014	2015	2016	2017	2018F
	$1,000	*$1,000*	*$1,000*	*$1,000*	*$1,000*
Farm sector debt	345,201,354	356,738,041	374,164,212	393,048,069	406,854,605
Real estate	196,780,224	208,769,246	225,980,433	238,058,397	248,492,395
Commercial banks 1/	73,254,162	79,163,795	84,417,512	88,744,108	NA
Farm Credit System	88,797,518	96,662,553	103,749,537	107,653,783	NA
Farm Service Agency	4,325,689	4,857,770	5,914,514	6,054,097	NA
Farmer Mac	4,728,807	4,843,551	5,456,587	6,266,206	NA
Individuals and others 1/	12,517,927	9,956,273	12,494,207	13,463,931	NA
Storage facility loans	752,327	757,809	743,955	769,178	NA
Life insurance companies	12,403,795	12,527,497	13,204,121	15,107,093	NA
Nonreal estate	148,421,130	147,968,795	148,183,780	154,989,672	158,362,211
Commercial banks 1/	70,737,959	73,177,901	73,233,553	73,294,843	NA
Farm Credit System	47,887,186	48,283,041	49,376,260	51,180,555	NA
Farm Service Agency	3,550,210	3,748,543	3,783,890	3,958,398	NA
Individuals and others 1/	26,245,776	22,759,310	21,790,077	26,555,877	NA

Footnotes
Data as of August 30, 2018
F = Forecast values.
NA = Data are not available/applicable.
Values are rounded to the nearest thousand. When 'Real (2018 dollars)' is selected, nominal values are adjusted for inflation using the chain-type GDP deflator, base year=2018.
1/ Beginning with 2012 estimates, farm sector debt held by savings associations is reported with the commercial bank lender group instead of the individuals and others grouping.

USDA/ERS Farm Income and Wealth Statistics

The Economic Research Service produces national balance sheets as part of our Farm Income and Wealth Statistics data products

Objective: find a method to procure state-level estimates through better use of existing reports and new disaggregation methods of administrative data

Focus: 85% of loan volumes held by Commercial/Savings Banks, the Farm Service Agency and the Farm Credit System

Challenges for Top Institutional Lenders

Institution	Issue
Farm Service Agency	State-level data exists, but not in readily available format
Commercial/Savings Banks	Data is aggregated by bank, not by state. State-level data can be imputed with regulatory sources
Farm Credit System	Data is aggregated by bank, not by state. Limited regulatory information. State values must be estimated using other means (e.g. surveys)

Data Sources

Commercial/Savings Banks

Call Report Data: 1976 – 2018

- Federal Reserve Bank of Chicago; Federal Financial Institutions Examination Council

Summary of Deposits: 1994 - 2018

- Federal Deposit Insurance Corporation

Community Reinvestment Act: 1997 – 2018

- Federal Financial Institutions Examination Council

Home Mortgage Disclosure Act: 1999 – 2016

- Federal Financial Institutions Examination Council

Farm Service Agency

Monthly Management Summary Reports: 2003 – 2018

- Farm Service Agency

Farm Credit System

Call Report Data: 2005 – 2018

- Farm Credit Administration

Other Sources

Census of Agriculture: 1992 – 2012

- USDA National Agricultural Statistical Service

Accessing State-Level Information From PDFs

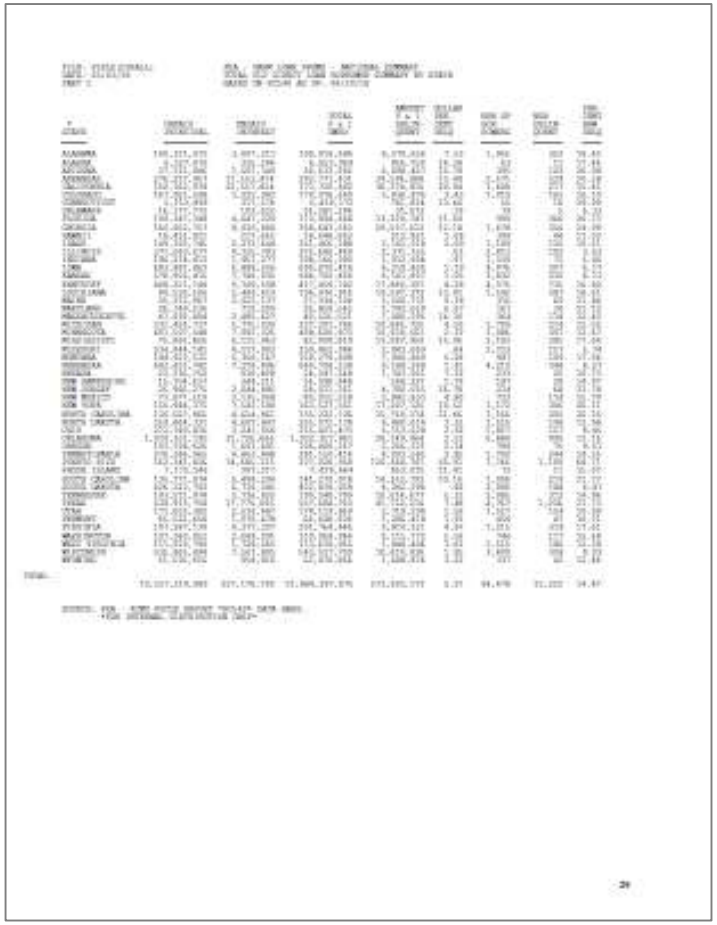
Using R to Scrape PDFs from the Farm Service Agency

Objective: Read in data frame of tabular data in PDF of loan data for states

Tabular data in FSA PDFs were accessed by

- Transforming the PDFs into a data frame containing lines of text
- Indexing start and end of table using regular expressions
- Coercing fixed width data into columns

Sample PDF



Using R to Scrape PDFs from the Farm Service Agency

Packages used:
Pdftools ← PDF to text
Stringr ← string manipulation

Full script had to account for quirks, such as

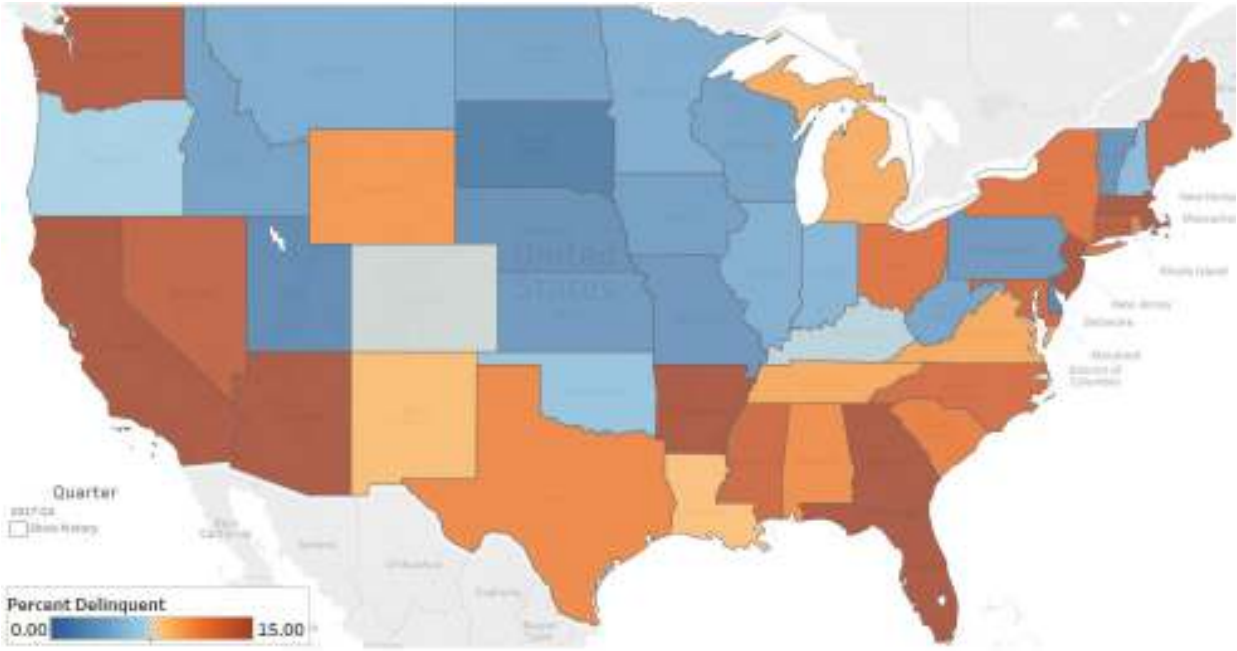
- Changes to table over time
- "West Virginia"

Example – PDF Scrape

```
#Example: Pulling table of state data from PDF
#Convert PDF into textfile
textfile <- pdf_text(filepath)

#Creating new row in dataframe for each space
flatfile <- strsplit(textfile,"\n")

#Searching for page containing table of interest TableName
for(i in 1:length(flatfile)){

  #Identifying start of table TableName
  if(grepl(sprintf(TableName),flatfile[[i]][3])){

    #Looping through rows of interest using regular expressions
    for(j in grep("ALABAMA",flatfile[[i]]):grep("WYOMING",flatfile[[i]])){

      #Converting each each row into a vector
      vector <- unlist(strsplit(str_replace(gsub("\\s+",
                " ", str_trim(tolower(flatfile[[i]][j]))), "B", "b")," ")))

      #Binding each  row to a new dataframe, outfile
      outfile <- rbind(outfile,vector,deparse.level = 0,stringsAsFactors=FALSE)
    }
  }
}
```


Using R to Scrape PDFs from the Farm Service Agency

Delinquency Rates for FSA Production Loans – Q2 2018



Benefits of this approach:

- Parameterized code allows for automatic quarterly updates
- Additional variables or tables can be extracted with minimal code changes
- No need for intermediate tables for data visualizations

Disaggregating Bank-Level Data with Regulatory Information

Using R to Disaggregate Commercial Bank Call Report Data

Call reports list information by headquarters, not where loans occur

Solution: disaggregate call reports information into counties using regulatory information that captures bank presence by county and re-aggregate at the state level

Overarching Process:

1) Read in data
2) Cleaning and imputation
3) Assigning county shares
4) Many-to-one merges
5) Calculation of county shares
6) Re-aggregation and upload

Original Data – Loan Volume by Institution

Disaggregated Data – Loan Volume by County

Accessing Call Report/Regulatory Information

Most sources used are contained in zip files that have URLs that can be used for direct access

Packages used:
RCurl ← url access
SASxport ← read xport files

Section Process
- Download zip file to temporary directory
- Identify index using regular expressions
- Merge pulled schedules
- Delete temporary files

Significant cleaning, but uses simple methods

Example – Zip File Extract

```
#Example - reading in call report data from multiple schedules
#Adding file.exists helps avoid failures
if(file.exists(call_report_url))){

  #Create a temporary directory and download the full zipfile
  td <- tempdir()
  tf = tempfile(tmpdir=td, fileext=".zip")
  download.file(call_report_url,tf,mode="wb")

  #Find the index number for the name of the first schedule and read in the data
  fname = unzip(tf, list=TRUE)$Name[grepl("SCHEDULE A",unzip(tf, list=TRUE)$Name)]
  unzip(tf, files=fname, exdir=td, overwrite=TRUE)
  fpath = file.path(td, fname)
  sched_a = read.xport(fpath)

  #Repeat the process for the second schedule
  fname = unzip(tf, list=TRUE)$Name[grepl("SCHEDULE B",unzip(tf, list=TRUE)$Name)]
  unzip(tf, files=fname, exdir=td, overwrite=TRUE)
  fpath = file.path(td, fname)
  sched_b = read.xport(fpath)

  #Combine schedules
  all_schedules <- Reduce(function(x, y) merge(x, y, all=TRUE), list(sched_a,sched_b))

  #Delete all files in temporary directory
  do.call(file.remove, list(list.files(td, full.names = TRUE)))
}
```


Assigning County Proportions

Merging call report data with regulatory data allows us to proxy for an institution's regional lending

One source will contain aggregated information you are attempting to disaggregate, and the other(s) contain disaggregated information that can be used to proxy for regional dispersion

Section Process

- Add column to regulatory table containing sum by group
- Merge in call report data (many to one)
- Multiply across for county share

Largest Lender: Call Report Information

Largest Lender: Regulatory Information

Assigning County Proportions

Packages used:
Rodbc ← SQL server connection

Challenge: often requires strong assumptions, in-depth knowledge and significant cleaning before merge

Note: many-to-one merges can result in tables that are too large for individual machines to hold in memory

Example – Assigning County Shares with Share of Total

```
#Example - using regulatory information to assign shares
#need to create sum of volume by unit_id and add as column
df$unit_sum <- ave(df$volume, df$unit_id, FUN=sum)
#Divide volume (county level) by this total for share
df$county_share <- df$volume / df$unit_sum
#checking to see if shares by unit add to 1
df$check <- ave(df$county_share, df$unit_id, FUN=sum)
#If correctly disaggregated, table should only include 1s
table(df$check)
#merge in call report information where reg info exists
df_with_cr <- merge(df,call_reports,by="unit_id",all.x=TRUE,all.y=FALSE)
#county-level volumes
df_with_cr$loan_volume_share <- df_with_cr$county_share * df_with_cr$loan_volume
#Saving database out to SQL - where dbhandle is database handle
sqlSave(dbhandle, df_with_cr, "disagg_call_reports", fast=TRUE,append=TRUE,
        rownames=FALSE)
```

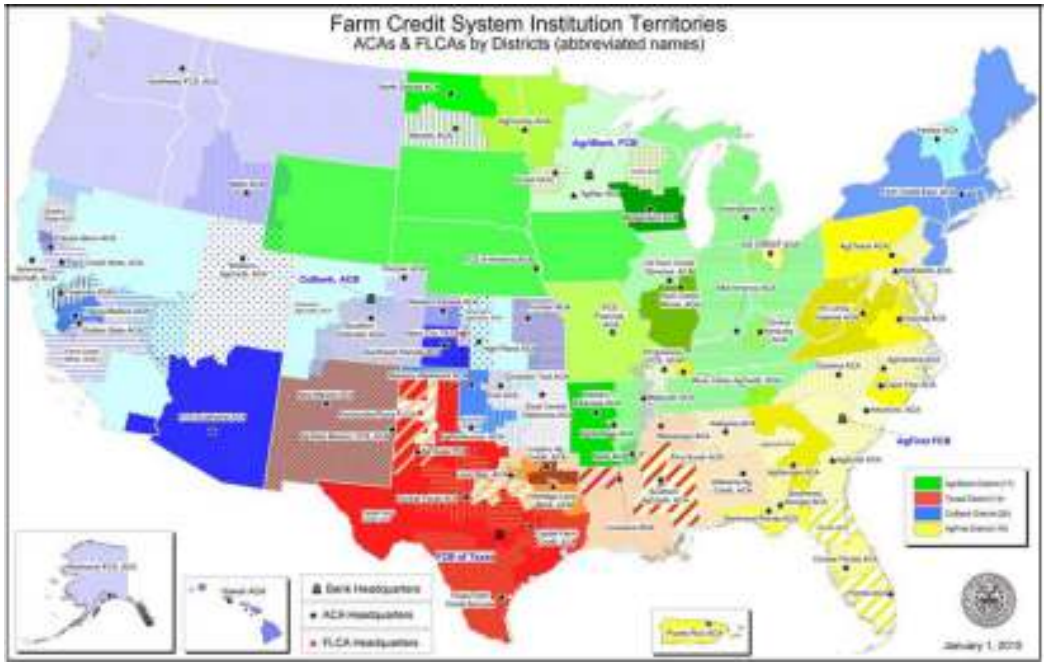

Refining Administrative Data with Surveys

Combining Administrative Data with Surveys

Similarities to commercial bank data:
1) Data reported at the institution level
2) Same importation method

Differences:
1) Limited regulatory information

Overarching Process:
1) Read in data
2) Survey analysis
3) Assign state shares



Using R to Disaggregate Commercial Bank Call Report Data

Packages used:
survey ← survey analysis

Useful specifically when survey total is less reliable than what is reported in administrative data, but produces valid proportions by group

Important to know survey limitations to know what mitigating factors to use (e.g. moving averages)

Example – Assigning State Share using Survey Share

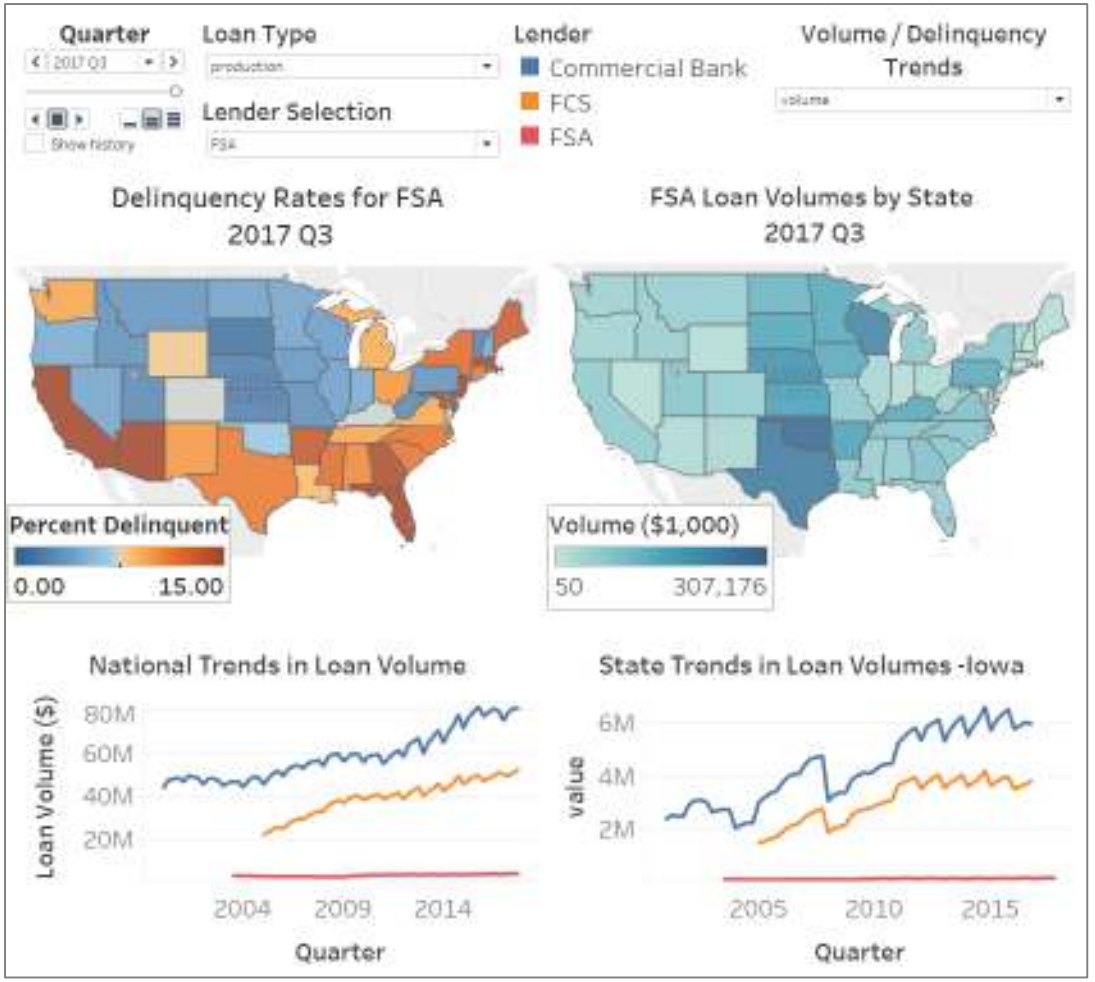
```
### Creating state shares based on national total
# read in survey data
data <- read.csv(surveypath,header=T)
# create survey design
survey.design <- svydesign(id=~id,data=data,weights = ~weights)
# summing debt by state
state_debt <- svyby(~debt,~state,survey.design,svytotal)
# finding share of debt by state
state_debt$debt_share <- state_debt$debt / sum(state_debt$debt)
# applying state share of debt to administrative total
state_debt$admin_share <- state_debt$debt_share * national_total
```


Combining All Methods

Potential Use Cases

- Extension of ERS data products
- Creation of new ERS visualizations
- Use for broader research purposes

Example: Dashboard Mockup



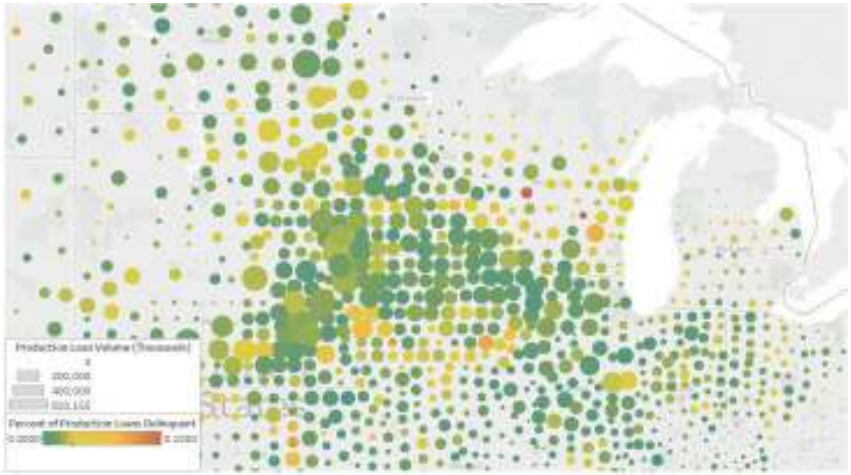
Contact Information

Greg Lyons, ERS/USDA
(202) 694-5147
greg.lyons@ers.usda.gov

Extensions: Financial Stress Factors

Real Estate Loan Delinquency Rate by County, Q2 2018

Production Loan Delinquency Rate by County, Q2 2018



Disaggregation methods can be extended to other schedules to look at regional financial stress, including:

- Delinquent loan volumes
- Loans in nonaccrual status
- Charge-offs

Extensions: Bank Branch Closures

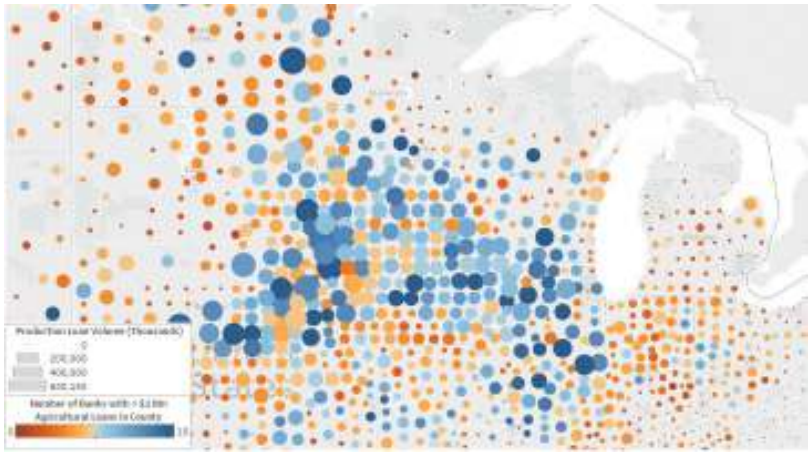
Number of Closed Branches by County, 2017

Subsequent research will pair county-level loan volumes and delinquencies with the bank branch closures from the FDIC's Reports of Structure Changes

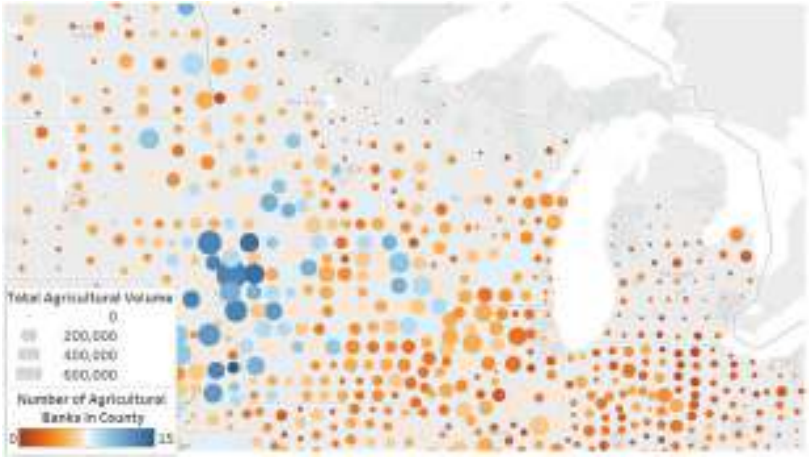
Primary aim is to understand the relationship between agricultural loan performance and bank branch closures

Extensions: Competition in Lending

Number of Banks with > $1Mn Loans in County



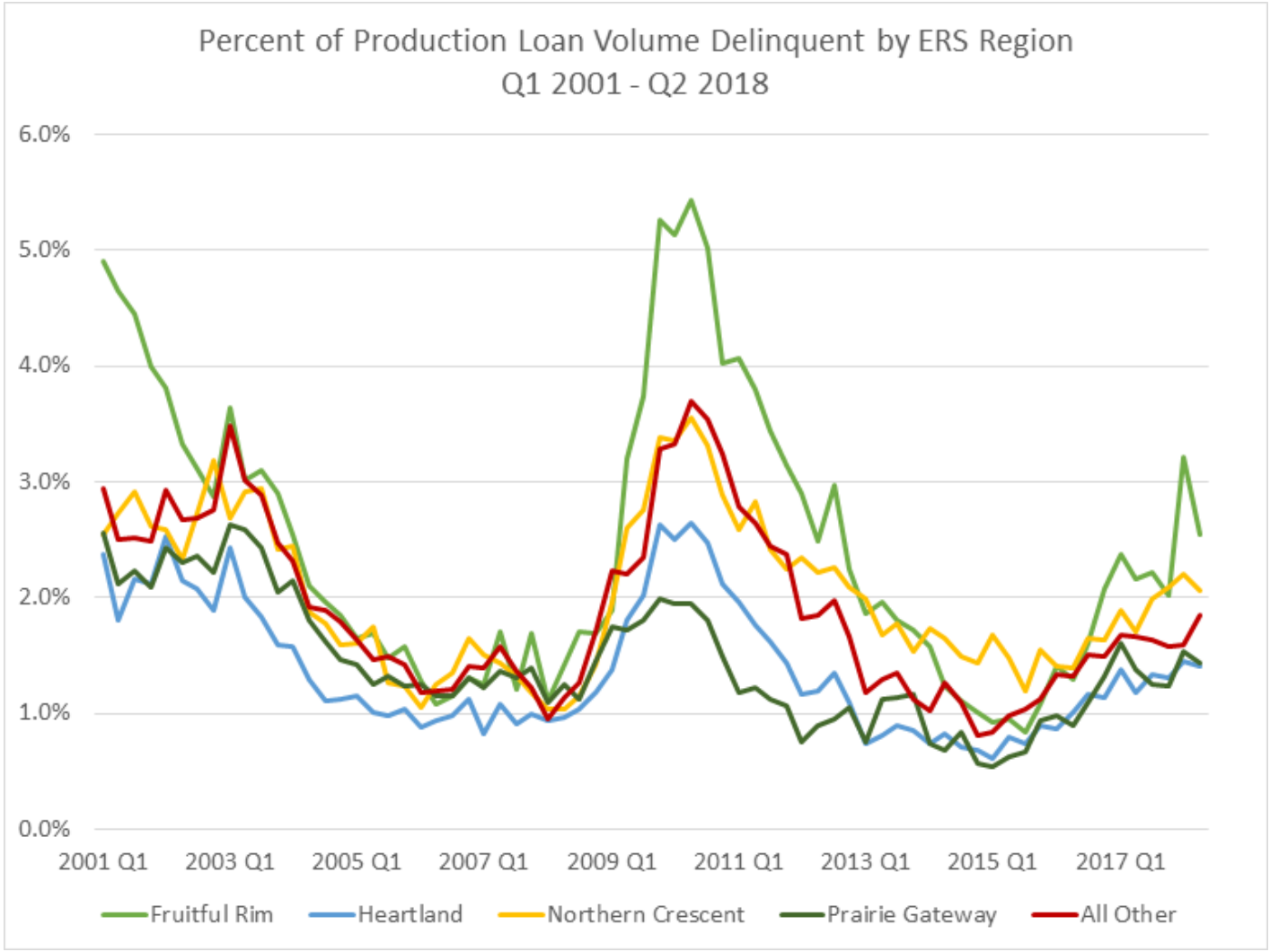
Number of FDIC-designated Ag. Banks in County



Can be used to check the robustness of agricultural credit markets across counties

- Number of institutions with x in agricultural lending
- Number of agricultural-focused lending institutions

Appendix: Production Loan Delinquencies by ERS Production Region



Appendix: Real Estate Loan Delinquencies by ERS Production Region