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

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

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

• 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
• 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
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[[1]])){
            #Converting 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

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

Delinquency Rates for FSA Production Loans – Q2 2018
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
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

```r
#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 zip file
  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
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

```
1  #Example - using regulatory information to assign shares
2  #need to create sum of volume by unit_id and add as column
3  df$unit_sum <- ave(df$volume, df$unit_id, FUN=sum)
4  #divide volume (county level) by this total for share
5  df$county_share <- df$volume / df$unit_sum
6  #checking to see if shares by unit add to 1
7  df$check <- ave(df$county_share, df$unit_id, FUN=sum)
8  #If correctly disaggregated, table should only include 1s
9  table(df$check)
10 #merge in call report information where reg info exists
11 df_with_cr <- merge(df, call_reports, by="unit_id", all.x=TRUE, all.y=FALSE)
12 #county-level volumes
13 df_with_cr$loan_volume_share <- df_with_cr$county_share * df_with_cr$loan_volume
14 #Saving database out to SQL - where dbhandle is database handle
15 sqlSave(dbhandle, df_with_cr, "disagg_call_reports", fast=TRUE, append=TRUE,
16   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

```r
# 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, svytot)
# 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

Example: Dashboard Mockup

Potential Use Cases

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

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

Number of FDIC-designated Ag. Banks in County
Appendix: Production Loan Delinquencies by ERS Production Region

Percent of Production Loan Volume Delinquent by ERS Region
Q1 2001 - Q2 2018

Fruitful Rim  Heartland  Northern Crescent  Prairie Gateway  All Other
Appendix: Real Estate Loan Delinquencies by ERS Production Region

Percent of Real Estate Loan Volume Delinquent by ERS Region
Q1 2001 - Q2 2018

- Fruitful Rim
- Heartland
- Northern Crescent
- Prairie Gateway
- Other

[Graph showing the percent of real estate loan volume delinquent by ERS region from Q1 2001 to Q2 2018.]