Computing HHI using CIT-IRP5 files

2025-02-27

Table of contents

1 Summary

Here are the summary of tasks:

- We have set up an account at the secure lab facility to access CIT-IRP5 merged employee-employer data sets of 2011-2018.
- We read all the files in `R`, converted into `data.table` format, and saved using `qs` package.
- We coutnted the number of employees per firm at the following levels of geographical areas:

  - Province
  - District municipality
  - Local municipality

- Within each geographical areas, we computed the total number of employees, shares of each firms, and HHI.
- We also summarised the data.
- In sum, as specified in the Specification, we have computed:
  - Number of total employees E_{mi} in the area m
  - Number of employees E_{mi} for each of firm i
  - Employment share E_{mi}/E_m of each firm in the area
  - HHI of the area h_m
  - Attached the above statistics to each administrative areas
  - Descriptive statistics at various administrative levels
- We have updated the analysis of the previous research team to incorporate more recent years and a wider definition of agricultural firms. We confirmed the results are qualitatively similar with the previous research.

2 Descriptive statistics

Using `LShare` data, we show the descriptive statistics at various aggregation levels.

Table 1: Descriptive statsitics at Province level

busprov_geo	taxrefno	Entity	Txrf	kerr_income
Length: 2122819	Length: 2122819	private: 2022554	Length: 2122819	Min.: 0.000e+00
Class: character	Class: character	gov: 100265	Class: character	1st Qu.: 2.017e+04
Mode: character	Mode: character		Mode: character	Median: 6.337e+04
				Mean: 1.829e+05
				3rd Qu.: 1.712e+05
				Max.: 5.941e+10

Table 2: Descriptive statsitics at Province level (continued)

AreaLevel	Total	Share	HHI	nHHI
Length: 2122819	Min.: 2686842	Min.: 3.700e-07	Min.: 1553	Min.: 1553
Class: character	1st Qu.: 2686842	1st Qu.: 2.643e-05	1st Qu.: 1553	1st Qu.: 1553
Mode: character	Median: 2686842	Median: 1.954e-04	Median: 1553	Median: 1553
	Mean: 2686842	Mean: 7.648e-03	Mean: 1553	Mean: 1553
	3rd Qu.: 2686842	3rd Qu.: 2.328e-03	3rd Qu.: 1553	3rd Qu.: 1553
	Max.: 2686842	Max.: 7.048e-02	Max.: 1553	Max.: 1553

Table 3: Descriptive statsitics at Province level (continued)

TotalG	ShareG	HHIG	nHHIG	busdistmuni_geo
Min.: 2497476	Min.: 0.000e+00	Min.: 709.3	Min.: 709.3	Length: 2122819
1st Qu.: 2497476	1st Qu.: 1.922e-05	1st Qu.: 709.3	1st Qu.: 709.3	Class: character
Median: 2497476	Median: 1.381e-04	Median: 709.3	Median: 709.3	Mode: character
Mean: 2497476	Mean: 4.646e-03	Mean: 709.3	Mean: 709.3	
3rd Qu.: 2497476	3rd Qu.: 1.435e-03	3rd Qu.: 709.3	3rd Qu.: 709.3	
Max.: 2497476	Max.: 6.148e-02	Max.: 709.3	Max.: 709.3	

Table 4: Descriptive statsitics at Province level (end)

buslocmuni_geo	busmainplc_geo
Length: 2122819	Length: 2122819
Class: character	Class: character
Mode: character	Mode: character

Table 5: Descriptive statsitics at district municipality level

busprov_geo	taxrefno	Entity	Txrf	kerr_income
Length: 2128635	Length: 2128635	private: 2023622	Length: 2128635	Min.: 0.000e+00
Class: character	Class: character	gov: 105013	Class: character	1st Qu.: 2.012e+04
Mode: character	Mode: character		Mode: character	Median: 6.335e+04
				Mean: 1.828e+05
				3rd Qu.: 1.715e+05
				Max.: 5.941e+10

Table 6: Descriptive statsitics at district municipality level (continued)

AreaLevel	Total	Share	HHI	nHHI
Length: 2128635	Min.: 450678	Min.: 4.500e-07	Min.: 246.3	Min.: 246.3
Class: character	1st Qu.: 2236164	1st Qu.: 3.994e-05	1st Qu.: 1639.8	1st Qu.: 1639.8
Mode: character	Median: 2236164	Median: 3.475e-04	Median: 1639.8	Median: 1639.8
	Mean: 1904555	Mean: 9.376e-03	Mean: 1381.0	Mean: 1381.0
	3rd Qu.: 2236164	3rd Qu.: 4.150e-03	3rd Qu.: 1639.8	3rd Qu.: 1639.8
	Max.: 2236164	Max.: 7.394e-02	Max.: 1639.8	Max.: 1639.8

Table 7: Descriptive statsitics at district municipality level (continued)

TotalG	ShareG	HHIG	nHHIG	busdistmuni_geo
Min.: 417357	Min.: 0.000e+00	Min.: 74.83	Min.: 74.83	Length: 2128635
1st Qu.: 2080119	1st Qu.: 2.788e-05	1st Qu.: 1016.93	1st Qu.: 1016.93	Class: character
Median: 2080119	Median: 2.360e-04	Median: 1016.93	Median: 1016.93	Mode: character
Mean: 1771303	Mean: 6.334e-03	Mean: 841.96	Mean: 841.96	
3rd Qu.: 2080119	3rd Qu.: 2.197e-03	3rd Qu.: 1016.93	3rd Qu.: 1016.93	
Max.: 2080119	Max.: 7.382e-02	Max.: 1016.93	Max.: 1016.93	

Table 8: Descriptive statsitics at district municipality level (end)

buslocmuni_geo	busmainplc_geo
Length: 2128635	Length: 2128635
Class: character	Class: character
Mode: character	Mode: character

Table 9: Descriptive statsitics at local municipality level

busprov_geo	taxrefno	Entity	Txrf	kerr_income
Length: 2130454	Length: 2130454	private: 2024228	Length: 2130454	Min.: 0.000e+00
Class: character	Class: character	gov: 106226	Class: character	1st Qu.: 2.007e+04
Mode: character	Mode: character		Mode: character	Median: 6.330e+04
				Mean: 1.827e+05
				3rd Qu.: 1.715e+05
				Max.: 5.941e+10

Table 10: Descriptive statsitics at local municipality level (continued)

AreaLevel	Total	Share	HHI	nHHI
Length: 2130454	Min.: 38318	Min.: 4.500e-07	Min.: 104	Min.: 104
Class: character	1st Qu.: 2236164	1st Qu.: 4.561e-05	1st Qu.: 1640	1st Qu.: 1640
Mode: character	Median: 2236164	Median: 4.642e-04	Median: 1640	Median: 1640
	Mean: 1838107	Mean: 1.244e-02	Mean: 1387	Mean: 1387
	3rd Qu.: 2236164	3rd Qu.: 8.500e-03	3rd Qu.: 1640	3rd Qu.: 1640
	Max.: 2236164	Max.: 1.523e-01	Max.: 1640	Max.: 1640

Table 11: Descriptive statsitics at local municipality level (continued)

TotalG	ShareG	HHIG	nHHIG	busdistmuni_geo
Min.: 36605	Min.: 0.000e+00	Min.: 24.1	Min.: 24.1	Length: 2130454
1st Qu.: 2080119	1st Qu.: 3.029e-05	1st Qu.: 1016.9	1st Qu.: 1016.9	Class: character
Median: 2080119	Median: 3.201e-04	Median: 1016.9	Median: 1016.9	Mode: character
Mean: 1709731	Mean: 9.428e-03	Mean: 876.1	Mean: 876.1	
3rd Qu.: 2080119	3rd Qu.: 4.359e-03	3rd Qu.: 1016.9	3rd Qu.: 1016.9	
Max.: 2080119	Max.: 1.634e-01	Max.: 1016.9	Max.: 1016.9	

Table 12: Descriptive statsitics at local municipality level (end)

buslocmuni_geo	busmainplc_geo
Length: 2130454	Length: 2130454
Class: character	Class: character
Mode: character	Mode: character

3 Codes

Read IRP5 v5 data.

```r
library(qs)
library(data.table)
library(readstata13)
for (yr in 9:22) {
  if (yr < 10) yr <- paste0("0", yr)
  ## Below code is run only once: Start
  irpyr <- read.dta13(paste0(pathdataIRP, "IRP5_20", yr, "_cleaned.dta"))
  ipyr <- data.table(irpyr)
  rm(irpyr)
  qsave(ipyr, paste0(pathdata, "irp", yr, ".qs"), nthreads = 8)
  ## Below code is run only once: End
  ipyr <- qread(paste0(pathdata, "irp", yr, ".qs"), nthreads = 8)
  ## Keep only nature of person is "an individual"
  ipyr <- ipyr[grepl("A", natureofperson), ]
  ## Drop obs with kerr_income == 0
  ipyr <- ipyr[kerr_income != 0, ]
  ## Note that there are header lines with taxrefno = "NULL"
  ## taxrefno = "NULL": no business geo info
  ## Num: Number of employees in a firm, after dropping taxrefno = "NULL"
  ipyr[, Num := as.integer(.N), by = .(taxrefno)]
  ipyr[is.na(Num) | grepl("NULL", taxrefno), Num := 0L]
  ## Num2: Number of employees in a firm, treat taxrefno = "NULL" as sole proprietor
  ## To do so, create and assign taxrefno to them
  ipyr[, Tref := taxrefno]
  ipyr[grepl("NULL", taxrefno), Tref := 1:.N]
  ipyr[, Num2 := as.integer(.N), by = .(Tref)]
```

```
  ipyr[is.na(Num) | grepl("NULL", taxrefno), Num := 0L]
  icount <- unique(ipyr[, .(busprov_geo, busdistmuni_geo, buslocmuni_geo,
    busmainplc_geo, taxrefno,
    #### need to run below if we want payereferenceno
    #### payereferenceno,
    TRef, natureofperson, kerr_income, Num, Num2)])
  qsave(icount, paste0(pathdata, "icount", yr, ".qs"), nthreads = 8)
}
```

Aggregate by Local Municipality, Districts.

```
library(qs)
library(data.table)
#### Read data
#### nthreds = 16 at NT-SDF, = 8 with my laptop (less 1, keep 1 for other computations)
ipyr <- qread(paste0(pathdata, "irp12.qs"), nthreads = 15)
ipyr <- ipyr[grepl("A", natureofperson), ]
#### Not sure if we need this. Test how many obs will be dropped.
nrow(ipyr[kerr_income != 0, ]); (nrow(ipyr[kerr_income != 0, ])/nrow(ipyr))*100
#### For 2012, no NA in payereferenceno
#### (n0 <- nrow(ipyr[is.na(payereferenceno) | payereferenceno == "", ]));
#### (n0)/nrow(ipyr)*100
#### 14% are gov employees [is.na(taxrefno) | taxrefno == ""]
#### So we cannot simply drop entries with NAs in taxrefno
(n0 <- nrow(ipyr[grepl("NULL", taxrefno) | is.na(taxrefno) | taxrefno == "", ]));
(n0/nrow(ipyr))*100
#### 1. Use payereferenceno to count the total
```

```
#### 2. Compute the shares of each firms
####   a. In doing so, create a hypothetical "gov entity" to
####      aggregate the entries with NAs in taxrefno
####   b. Compute the shares of each firms including "gov entity" thence HHI
ipyr[, Txrf := taxrefno]
#### Note: GovEntity has taxrefno == "" or NULL, so omit from unique operation below
ipyr[grepl("NULL", taxrefno) | taxrefno == "" | is.na(taxrefno), Txrf := "GovEntity"]
ip = copy(ipyr)
#### Drop 2nd GovEntity entries at each geo level
GeoLevel <- c("Prv", "Dis", "Loc", "Mai")[-4]
for (g in 1:length(GeoLevel)) {
  ip[, Gov2 := 0L]
  ip[grepl("Gov", Txrf), Gov2 := as.integer(1:.N), by=
    eval(c("busprov_geo", "busdistmuni_geo", "buslocmuni_geo", "busmainplc_geo")[1:g])]
  ip <- ip[Gov2 != 2, ]
}
ip[, Entity := "private"]
ip[grepl("NULL", taxrefno) | taxrefno == "" | is.na(taxrefno), Entity := "gov"]
ip[, Entity := factor(Entity, levels = c("private", "gov"))]
#### table(ipry[, Entity], exclude = NULL)
#### Count shares and HHI for the entire country
LShare <- NULL
for (g in 1:length(GeoLevel)) {
  ipGeo = copy(ip)
  ipGeo[, EachNum := as.integer(.N), by =
    eval(c(c("busprov_geo", "busdistmuni_geo",
```

```
    "buslocmuni_geo", "busmainplc_geo")[1:g], "Txrf"))]
ipGeo[, Total := as.integer(.N), by =
  eval(c("busprov_geo", "busdistmuni_geo",
    "buslocmuni_geo", "busmainplc_geo")[1:g])]
ipGeo[, Share := round(EachNum/Total, 8)]
#### ShareG, HHIG: Share and HHI after dropping GovEntity ####
#### We compute "total without GovEntity" by subtracting GTotal (nrow of Entity==gov)
ipGeo[grepl("gov", Entity), GTotal := as.integer(.N), by =
  eval(c("busprov_geo", "busdistmuni_geo", "buslocmuni_geo", "busmainplc_geo")[1:g])]
ipGeo[, GTotal := GTotal[!is.na(GTotal)][1], by =
  eval(c("busprov_geo", "busdistmuni_geo", "buslocmuni_geo", "busmainplc_geo")[1:g])]
#### We define EachNumG by replacing EachNum of GovEntity with 0L
ipGeo[, EachNumG := EachNum]
ipGeo[grepl("gov", Entity), EachNumG := 0L]
ipGeo[, TotalG := Total-GTotal]
ipGeo[, ShareG := round(EachNumG/TotalG, 8)]
ipGeo[, AreaLevel := GeoLevel[g]]
lshare <- unique(ipGeo[,
  c(c("busprov_geo", "busdistmuni_geo", "buslocmuni_geo", "busmainplc_geo")[1:g],
  "Entity", "Txrf",  #"taxrefno",
  "AreaLevel",
  "EachNum", "Total", "Share", "EachNumG", "TotalG", "ShareG"
  ), with = F])
lshare[, HHI := sum(Share^(2), na.rm = T), by =
  eval(c("busprov_geo", "busdistmuni_geo", "buslocmuni_geo", "busmainplc_geo")[1:g])]
lshare[, nHHI := (HHI-1/Total)/(1-1/Total)]
```

```
  lshare[, HHIG := sum(ShareG^(2), na.rm = T), by =
    eval(c("busprov_geo", "busdistmuni_geo", "buslocmuni_geo", "busmainplc_geo")[1:g])]
  lshare[, nHHIG := (HHIG-1/TotalG)/(1-1/TotalG)]
  print(GeoLevel[g])
  print(lshare[1:10, c(c("busprov_geo", "busdistmuni_geo", "buslocmuni_geo", "busmainplc_geo"
    "Txrf", "Total", "EachNum", "Share", "HHI", "nHHI", "TotalG"), with = F])
  LShare <- rbindlist(list(LShare, lshare), use.names = T, fill = T)
}
setkey(LShare, busdistmuni_geo, buslocmuni_geo, busmainplc_geo, Txrf)
qsave(LShare, paste0(pathdata, "ShareHHI12.qs"), nthreads = 16)
```