ds421 final project

xrw

5/3/2018

#Contents

This is a final project PDF document for DS421 stitched together from other experiments in this rpo.

Some major goals were: - Get satellite data/imagery for village and county names. - Poke around household income data - Poke around land use change for a few Taobao villages.

Section A

First we'll take a look at household income data from CHIP, an{rd geocode the counties based off a csv of "official administrative codes".

We'll also poke at the data a bit, looking at changes over time.

Section B

##CHIP

CHIP (China Household Income Project) is put out by the CIID Beijing as a longitudinal survey. It's been happening since 1988 and includes all kinds of juicy stuff including land use.

Load up necessary libraries. Some data is in `.dta` which is Stata file.

```
library(tidyr)
library(tidyverse)
```

```
## Warning: package 'tibble' was built under R version 3.4.3
```

```
## Warning: package 'stringr' was built under R version 3.4.3
```

```
library(dplyr)
library(foreign)
library(reticulate)
```

```
## Warning: package 'reticulate' was built under R version 3.4.4
```

```
library(haven)
```

```
chips_rur_1988 <- read_dta('data/1988/09836-0002-Data.dta')
chips_rur_1995 <- read_tsv('data/1995/DS0002/03012-0002-Data.tsv')
```

```
## Warning in rbind(names(probs), probs_f): number of columns of result is not
## a multiple of vector length (arg 1)
```

```
## Warning: 198 parsing failures.
## row # A tibble: 5 x 5 col     row col     expected               actual file
## ... ................. ... ..........................................................
## See problems(...) for more details.
```

```
chips_rur_2002<- read_tsv('data/2002/DS0006/21741-0006-Data.tsv')
```

```
## Warning in rbind(names(probs), probs_f): number of columns of result is not
## a multiple of vector length (arg 1)
```

```
## Warning: 244 parsing failures.
## row # A tibble: 5 x 5 col     row col   expected               actual           file
## ... ................. ... ..... ..... ..... ..... ..... ..... ..... ..... .....
## See problems(...) for more details.
```

```
chips_rur_2007abc <- read_dta('data/2007 (2008)/RHS_w1_abc.dta')
chips_rur_2007d <- read_dta('data/2007 (2008)/RHS_w1_d.dta')
chips_rur_2007e1 <- read_dta('data/2007 (2008)/RHS_w1_e1.dta')
chips_rur_2007e2 <- read_dta('data/2007 (2008)/RHS_w1_e2.dta')
chips_rur_2007e3 <- read_dta('data/2007 (2008)/RHS_w1_e3.dta')
chips_rur_2007e4 <- read_dta('data/2007 (2008)/RHS_w1_e4.dta')
chips_rur_2007hhiexp <- read_dta('data/2007 (2008)/CHIP2007_income_and_expenditure_20150408.dta')
chips_rur_2008abc <- read_dta('data/2008 (2009)/RHS_w2_abc.dta')
chips_rur_2008d <- read_dta('data/2008 (2009)/RHS_w2_d.dta')
chips_rur_2008e <- read_dta('data/2008 (2009)/RHS_w2_e.dta')
chips_rur_2008f <- read_dta('data/2008 (2009)/RHS_w2_f.dta')
chips_rur_2008hgsg <- read_dta('data/2008 (2009)/RHS_w2_hgsg.dta')
chips_rur_2008hijk <- read_dta('data/2008 (2009)/RHS_w2_hijk.dta')
chips_rur_2008vill <- read_dta('data/2008 (2009)/RHS_w2_vill.dta')
chips_rur_2013 <- read_dta('data/2013/CHIP2013_rural_household_f_income_asset.dta')
```

```
#Helper function for dta files


labelDataset <- function(data) {
  correctLabel <- function(x) {

    if(!is.null(attributes(x)$labels)) {
      class(attributes(x)$labels) <- typeof(x)
    }
    return(x)
  }

  for(i in colnames(data)) {
    data[, i] <- correctLabel(data[, i])
  }
  return(data)
}

labelDataset(chips_rur_2007hhiexp)
```

```
## # A tibble: 8,000 x 14
##       name_id  exp1   exp2   exp3   exp4  exp5   exp6   exp7    exp8
##         <dbl> <dbl>  <dbl>  <dbl>  <dbl> <dbl>  <dbl>  <dbl>   <dbl>
##  1 130181001. 2670.    0.   2239. 1778.   971.  520.   386.    12.0
##  2 130181002. 9633. 5364.   8978. 6396.   786. 1935.  1687.  1946.
##  3 130181003. 4012.  442.   1941.  180.   120.  134.    13.0    3.00
##  4 130181004. 9487. 1986.   5853. 1020.   884. 4192.   386.   357.
##  5 130181005. 3371. 1174.   2750.  712.   290.  270.    44.7   34.3
##  6 130181006. 6559.   47.0 20386.  147.   432.  120.   102.    14.0
##  7 130181007. 4449.  293.   1748.  201.  1254.  374.   118.    47.3
##  8 130181008. 6968.  808.   1072.  147.   300.   69.2   72.0   65.4
```

```
##  9 130181009. 3518.  742.   1344.  246.   567. 1162.   564.   137.
## 10 130181010. 3805.  414.   9059.   48.0  959.  400.   473.   114.
## # ... with 7,990 more rows, and 5 more variables: income_net <dbl>,
## #   income_net_1 <dbl>, income_net_2 <dbl>, income_net_3 <dbl>,
## #   income_net_4 <dbl>
```

Table of columns used:

Year	Net household income	Land cultivated	Number of rooms in House	Fixed production assets	Total household exp on production
1988	na	na	na	na	na

Let's use: - HNET88 Net household income - LAT Land cultivated - HHO Number of rooms in house - VHPFP fixed production assets, Value of family's fixed productive assets - EFP88 Total household expenditures on productive operations