Stat450 Project Report v1.0

Harry Xu

2018-03-01

Summary

This project aims to better understand the distribution and trends of employment for two of China's civil government branches: Taxation Administration Services (TAS) and Environmental Protection Agencies (EPA). This report focuses on visualization and panel regression analysis using TAS data with different age levels. The model suggests that the significant variables that explain the variation in number of staff per province over time are: Urban area, GDP and GDP per Capita.

Introduction

This report gives a graphical visualization and statistical analysis of employment levels for China's civil government branches: Taxation Administration Services (TAS) and Environmental Protection Agencies (EPA). The number of staff for each province were collected at each government, age, and education levels from years 1992 to 2013. The main objectives are to first give visualization to suggest patterns and trends for government staffing based on different social and economic factor levels, and secondly, to build a statistical model to explain the variation in staffing numbers at various levels of government. The graphs in this report are mostly generated by Tableau Desktop 10.5, and panel regression methods is used rigorously in this the analysis to model the data. (EPA's analysis will follow if time allows in the future). This report contains the following sections, and here is the layout:

-Data Description -Methods -Results -Conclusions -Summary Results and Interpretaion -Future Recommendations -Appendix -Figures

Data Description

Two collections of employment data for the two civil service braches (TAS and EPA) were obtained. These panel data include staff numbers for each of the 30 provinces, categorized by different bureau, age, education and department unit levels. TAS dataset spans 8 years (1996, 2000-2007) and the EPA dataset spans 18 years (from 1992 to 2011, missing year 1994). The third dataset has 10 different Chinese Provincial economic variables for each province spanning from 2000 to 2013. The layout of these datasets were cleaned and simplified moderately to ensure a balanced panel data in order to allow computer software to generate visualization and panel regression analysis. The table below illustrates the structure of the TAS dataset (the EPA dataset are similar, and will not be shown as the analysis is focused mainly on TAS dataset.)

Dataset TAS:

Variable	Levels
Age	6 levels
Edu	7 levels
Unit	5 levels
Bureau	2 levels (STB and LTB)
Years	8 years (1996, 2000-2007)
Response Variable	Staff Number
Number of Province	30

Dataset Econdemographic: contains 10 different economic variables per province per year, and the structure is as follows:

Variable	Levels
Years	14 years (2000-2013)
Response Variable	Staff Number
Number of Province	30

Methods

The main statistical methods discussed in this report for the TAS staffing data is the two-way fixed-effect panel regression. A panel regression is a regression method used to analyse two dimensional data. These data are usually collected over a time period for the same entities. In this report, the Provinces would be the entities. Such two-way fixed-effect panel regression model can be expressed mathematically as:

$Y_{it} = \beta_1 X_{it} + \alpha_i + \lambda_t + \epsilon_{it}$

where Y_{it} is the response variable representing the Number of staff for a specific province in a specific year. X_{it} are the possible economic explanatory variables that could explain the response variables, and $beta_1$ are the slopes for the explanatory variables of interest for each X_{it}. α_i is the entity (Province) fixed effect, and λ_t is the time fixed effect. Both terms have a specific effect on Y_{it}. It is assumed that the Province fixed effect does not vary over time but may be different for different provinces, and the time fixed effect vary across time, but stays the same for all provinces. In the model, time t takes value 1996, 2000 to 2013, and ϵ_{it} is the error term for Province i in year t. It is assumed that the errors are iid and $\epsilon_{it} \sim N(0, \sigma^2)$

Results

From the visualizations shown in Figure-1, where each age level are plotted against 3 different economic variables: GDP per Capita, number of above size industrial companies, and Public Expenditure. It is interesting to see that the trend over the years are different at each age level, the test below will show whether the significance of these variables. Similar plots are also generated for each Education level and different Unit which are not shown in this report.

Figure-3 shows the plot of staff numbers at each age level in High/Med/Low GDP per Capita groupings. The groupings are grouped manually based on the top plot of Figure-2. The highest 3 Provinces with GDP per Capita being the "High" group, medium 6 provinces being the "Med" group, and the rest being the "Low" group. This plot gives an over all trend within the province groupings which could provide some interesting story to tell.

Figure-4 shows the distribution of number of staff by age in percentage. This plot provides useful information for comparison between groups, because it depicts the change of staff numbers per level with a same percentage scale.

For simplification purposes, only the regression output for TAS dataset with Age levels < 30 is shown below. The rest of the results are shown in the Figures section. The two-way fixed-effect panel regression model incorporated the independent variables: urban area, number of above size industrial companies, public expenditure, GDP and GDP per Capita.

```
## Loading required package: Formula
```

```
## Twoways effects Within Model
##
## Call:
## plm(formula = age.30.or..below ~ urbanarea + No.of.above.size.industrial.companies +
##     public.expenditure + GDP + GDPperCapita, data = agemerged,
```

```
##     effect = "twoway", model = "within", index = c("Year", "Province"))
##
## Balanced Panel: n = 8, T = 28, N = 224
##
## Residuals:
##    Min. 1st Qu.  Median 3rd Qu.    Max.
## -2360.0  -420.0    -2.8   323.0  8770.0
##
## Coefficients:
##                                      Estimate Std. Error t-value
## urbanarea                            1.770623   0.879591  2.0130
## No.of.above.size.industrial.companies  0.013076   0.037628  0.3475
## public.expenditure                    0.406531   1.029129  0.3950
## GDP                                  -0.569750   0.165028 -3.4524
## GDPperCapita                          0.116126   0.023818  4.8756
##                                      Pr(>|t|)
## urbanarea                           0.0455714 *
## No.of.above.size.industrial.companies 0.7286130
## public.expenditure                   0.6932825
## GDP                                  0.0006892 ***
## GDPperCapita                         2.336e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    226870000
## Residual Sum of Squares: 162040000
## R-Squared:      0.28576
## Adj. R-Squared: 0.13437
## F-statistic: 14.723 on 5 and 184 DF, p-value: 3.8556e-12
```

The above output gives the slope estimates for each β with a p-value. The p-values indicated the significance of the slope for each of the explanatory variables. It shows that Urban area, GDP and GDP per Capita are significant covariates at 5% in terms of explaining the variations for the number of staff who 30 years old or less over the span of years. However, the adjusted R^2 value is really low, which suggests that this model is probably not a great fit for all the explanatory variables chosen as the input values for the model. Nonetheless, the results might provide a useful direction for variable selection in the future.

Figure-5 shows the summary result for all Age levels in the TAS dataset. It gives the significant variables at 5% at each age level separated by bureau that may help explain the variations for the number of staff in TAS.

Conclusions and Recommendations

In summary, we are able to conclude that for one of the TAS dataset regarding different employee age levels, Urban area, GDP and GDP per Capita are the significant variables for explaining the variation of staff numbers across provinces over the years at a 5% significant level. This method can also be applied to other datasets along with the visualization tools to select the important explanatory variables for inference. There are many ways we could possibly explore and analyze due to the complexity of the dataset. For future recommendations, a random effect could be introduced into the panel regression model. Also, due to the nature of the data provided, we are not able to test the dependency of the variables between levels, that is, for example, between age levels. Missing values in the dataset also has potential problems for ensure the right assumptions for the analysis. It is advised that we look future into the dimensions of the dependent variables itself and explore more about the variations within these 30 provinces.

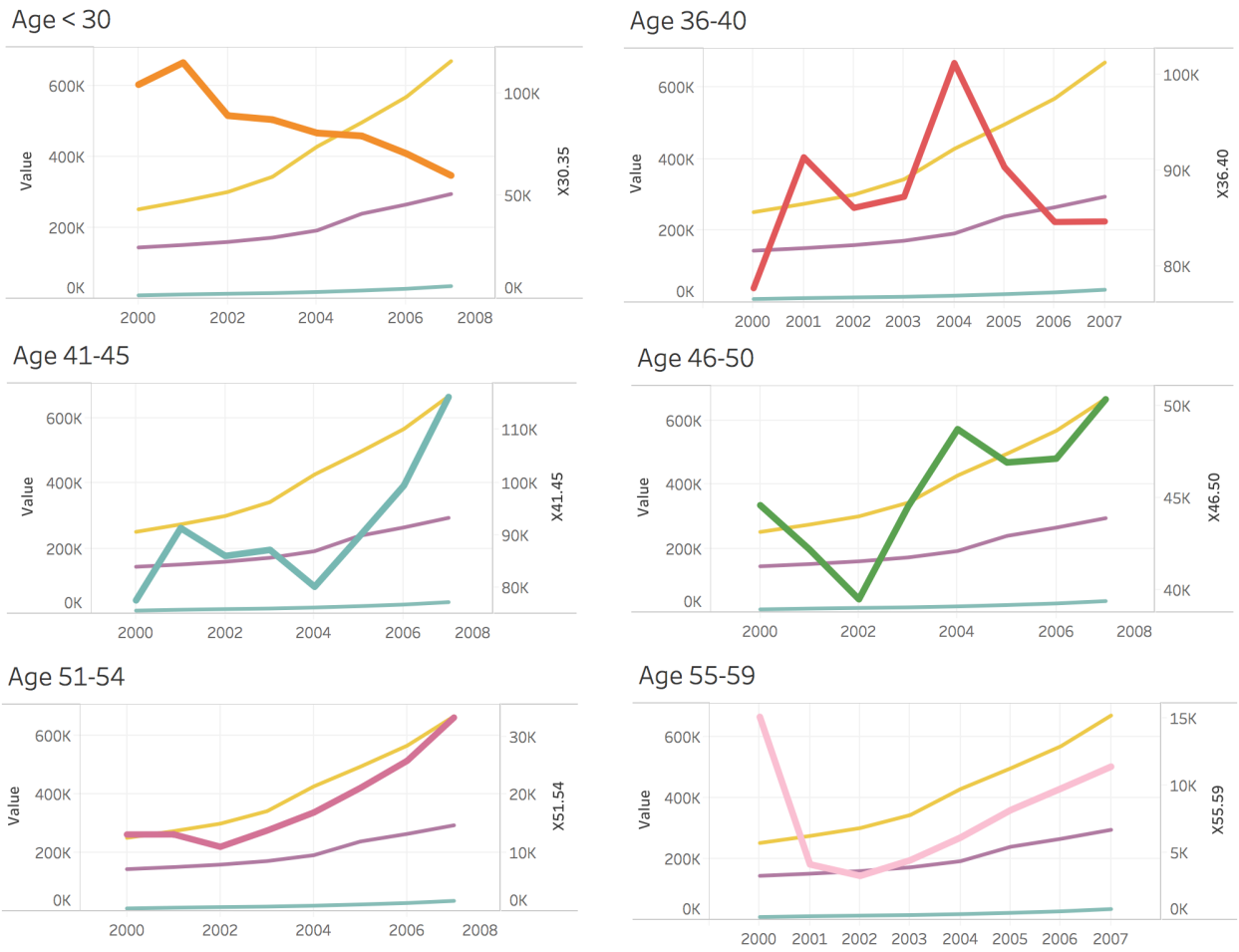
References

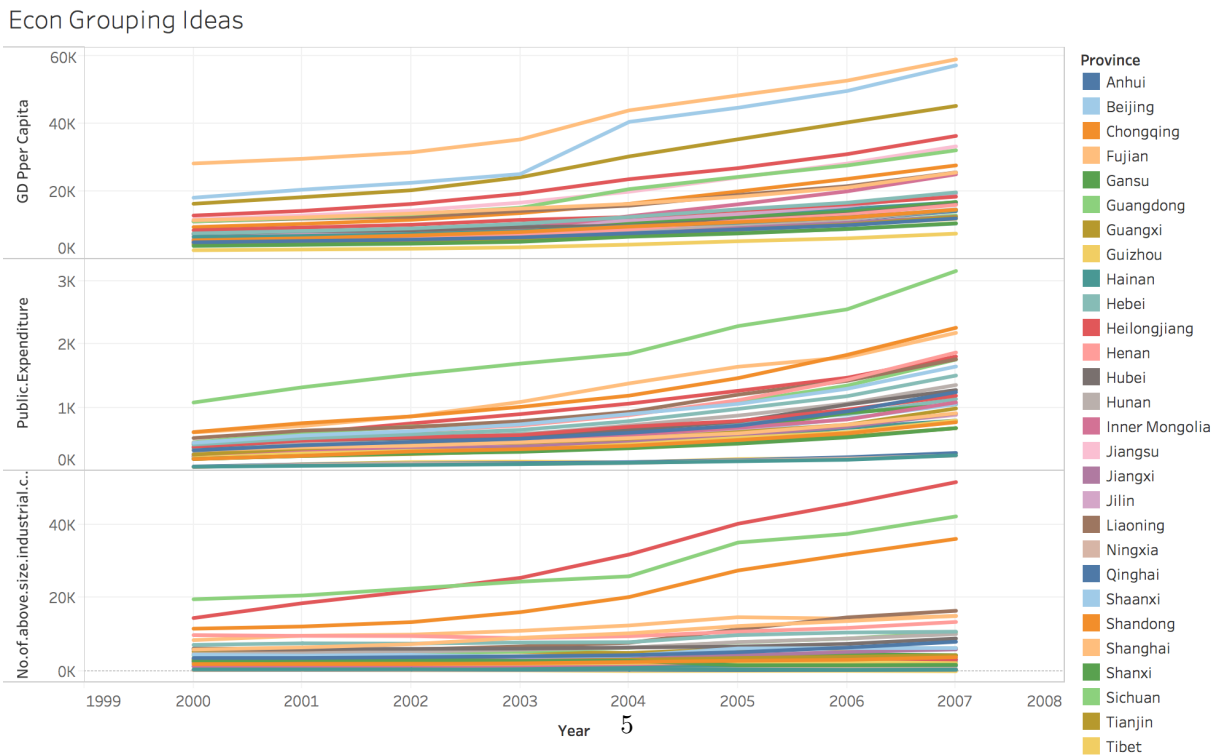
link to the reference for panel regression documentation: https://cran.r-project.org/web/packages/plm/plm.pdf

Appendix

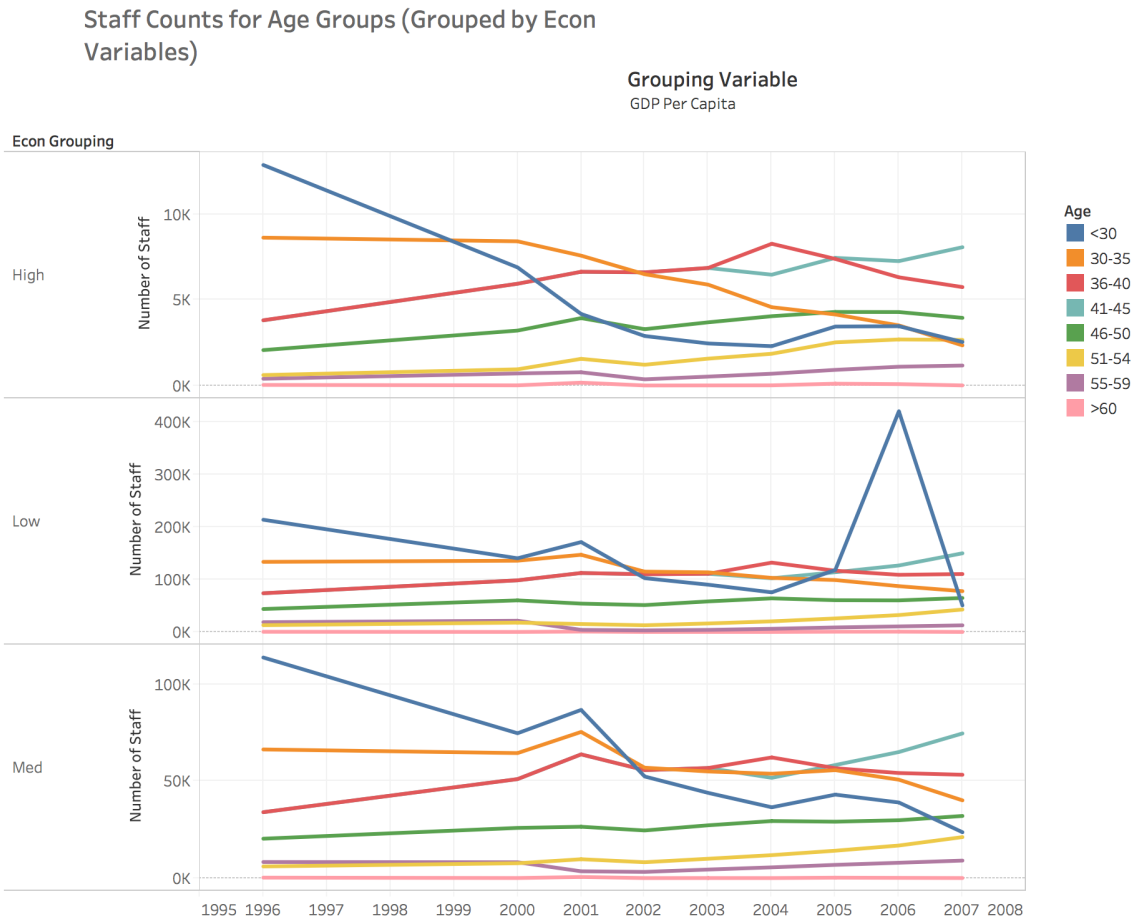
Git repository: https://github.com/wenzhengzzz/STAT450-550-Project-GovtAgencies/tree/master/stat450

Figures

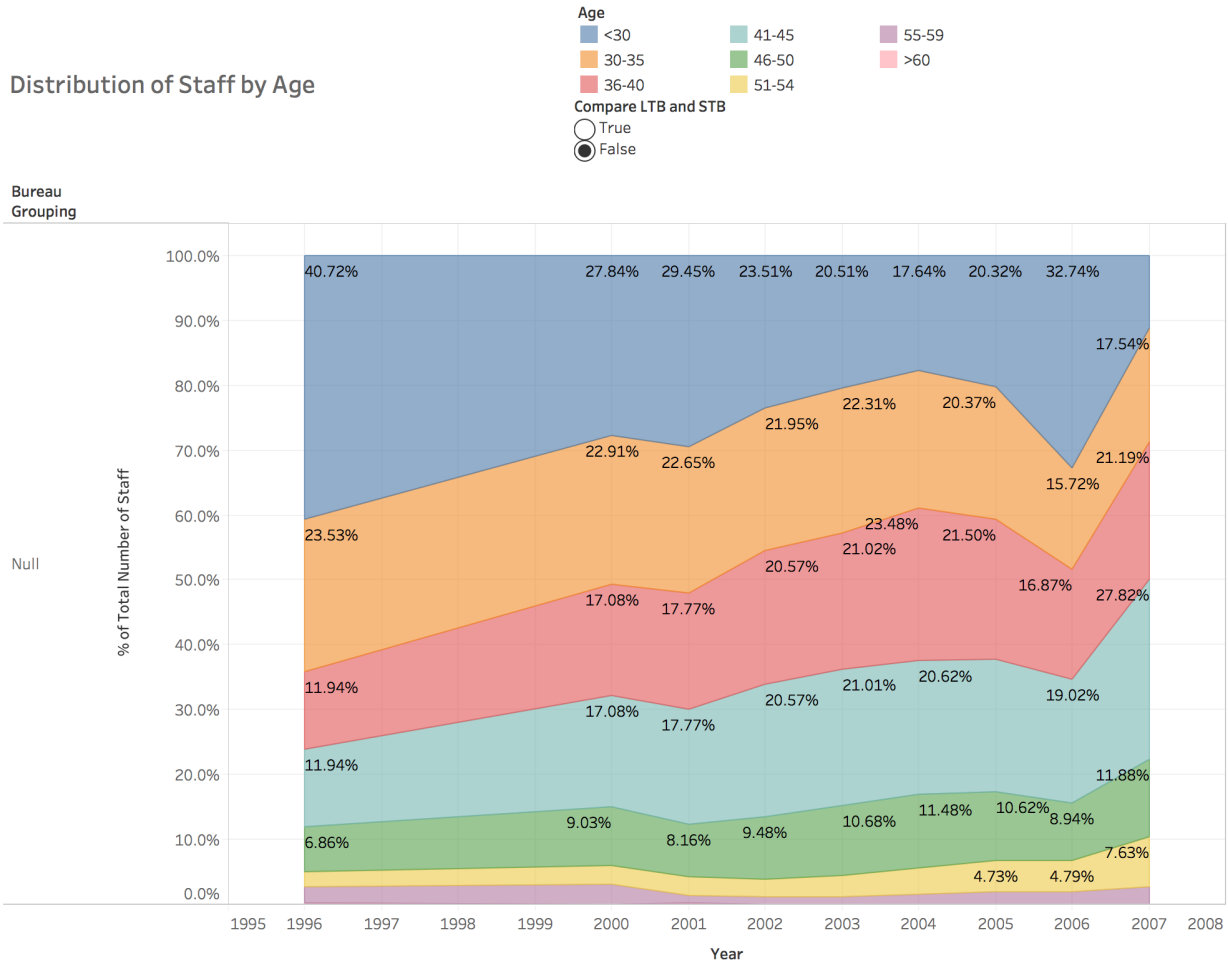




The trends of sum of GD Pper Capita, sum of Public.Expenditure and sum of No.of.above.size.industrial.companies for Year. Color shows details about Province.


Staff Counts for Age Groups (Grouped by Econ Variables)
Grouping Variable
GDP Per Capita
Econ Grouping
High
Low
Med
Number of Staff
10K
5K
0K
400K
300K
200K
100K
0K
100K
50K
0K
1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008
Age
<30
30-35
36-40
41-45
46-50
51-54
55-59
>60

Distribution of Staff by Age



Age - STB

- Under 30
  - GDP Per capital (6 x 10^-9)
  - GDP (.01431)
  - Public expenditure (0.036)
- 30-35
  - GDP (.044)
  - Urban Area (1.92 x 10^-6)
  - Public expenditure (0.0055)
  - GDP per capital (0.0038)
- 36-40
  - GDP (0.048)
  - Above size industrial (0.014)
  - GDP Per capita (0.01)
- 41-45
  - Public expenditure 0.003
  - GDP per capita (8.75 x 10^-6)
- 46-50 (none)
- 51-54
  - GDP (2.33x10^-7)
  - Urban Area (0.00013)
  - Above size industrial (0.0006)
  - GDP per capital (6.41x10^-6)
- 55-59 (none)
- 60 and above (none)

Age - LTB

- Under 30
  - GDP: 0.00069
  - Urban area: 0.0456
  - GDP per capita: 2.34 x 10^-6
- 30-35
  - GDP: 0.001244
  - Public expenditure: 7.16 x 10^-6
- 36-40
  - Urban area: 0.0417
  - Public expenditure: 0.0094
  - GDP per Capita: 0.0041
- 41-45
  - Urban area: 0.02
  - Public expenditure: 0.0055
  - GDP per Capita: 0.0099
- 46-50
  - Urban area: 0.00076
- 51-54
  - GDP (0.00027)
  - Urban Area (0.0003911)
  - Above size industrial (0.0048)
  - Public Expenditure (0.038)
  - GDP per capital (0.00057)
- 55-59 (none)
- 60 and above (none)