Stat 450 Project Report: Employment of the Chinese Government Agencies

Harry Xu

2018-03-06

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 the visualization and a panel regression analysis using the TAS data. The model suggests that the significant (5% level) variables that could possibly explain the variation in number of staff per province over time are: Public Expenditure and GDP per Capita. The EPA data analysis can be done in the similar manner.

Introduction

This report presents a graphical visualization and statistical analysis of employment levels for two of 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 generate visualizations to explore patterns and trends within 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 the government. The graphs in this report are generated by Tableau Desktop 10.5, and the panel regression method is used thoroughly in this analysis to model the data. This analysis only focuses on the TAS dataset, in particular, gives an example visualization and analysis for the number of staff at different age levels. The other factors, such as education and department unit levels, and the EPA analysis can follow a similar analysis structure, which are not shown in this report.

In addition, due to the limitation of the data provided, the possibility that the assumptions for the panel regression model may not hold, therefore the result of this analysis can be unsatisfactory. It is advised that the reader should take great attention to the conclusion and recommendation section of the report for clarification on these limitations and follows the guide to future steps.

The section layout for the report is as follows:

-Data Description

-Methods

-Results

-Conclusions and Recommendations (Future Steps)

-References

-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. The TAS dataset spans 9 years (1996, 2000-2007) and the EPA dataset spans 19 years (from 1992 to 2011, missing year 1994). The third dataset, namely Econ-demographic, has 10 different Chinese provincial economic variables for each province spanning from 2000 to 2013.

These datasets were cleaned and its structures simplified to ensure a balanced panel data, to allow computer software to generate visualization and panel regression analysis. The table below illustrates the structure of the TAS dataset (the EPA dataset is similar, and will not be shown as the analysis is focused mainly on the TAS dataset.)

Dataset TAS:

Variable	Levels
Age	8 levels (<30, 31-35, 36-40, 41-45, 46-50, 51-54, 55-59, >60)
Edu	7 levels (Junior, High, Technical, Other Post, Bachelor, Post-Grad)
Unit	4 levels (Admin, DA Admin, Tax, DA non-Admin)
Bureau	2 levels (STB and LTB)
Years	9 years (1996, 2000-2007)
Response Variable	Staff Number
Number of Provinces	30

Dataset Econ-demographic: contains the 10 different economic measures for each province per year, these measures include: Urban Area, Number of above size industries, public expenditure, GDP, Population, Ratio of Urban Area, GDP per Capita, and Area per Square Km. The structure of the dataset is as follows:

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

Methods

The main statistical method used for the analysis of TAS staffing data is the two-way fixed-effect panel regression. A panel regression is a regression method used to analyse two dimensional data. The data is usually collected over a time period for the same entities. The Provinces would be the entities. We use a fixed-effect model because we are interested in analyzing the impact of variables that vary over time. The Provincial effect is the effect at a specific year on the mean number of staff for that specific province. The time effect is the effect at a specific year. It is assumed that the Province effect is fixed, meaning that the mean number of staff for a province does not vary over time, but it might be different from other provinces. The time effect is fixed for provinces, and it indicates that the mean number of staff might be different in different years, but this staffing level difference is the same across all provinces.

Results

From the visualizations shown in Figure-1, each age level is 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 there is a significance of these variables to the number of staff at each age level. Note that the variable GDP is excluded

as it has strong correlation with GDP per Capita in order to avoid redundency. Similar plots can also be generated for the TAS dataset at different education level and office unit level, 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 are the "High" group, medium 6 provinces are the "Med" group, and the rest are the "Low" group. This plot gives an overall trend within the province groupings which can provide some interesting story to tell.

Figure-4 shows the distribution for the number of staff by age in percentage. This plot provides useful information for comparisons 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 level less than 30 is shown below. The rest of the results are shown in the Figures section. The two-way fixed-effect panel regression model incorporates the independent variables: urban area, number of above size industrial companies, public expenditure, and GDP per Capita.

The Appendix 2 R output gives the slope estimates for each β with a p-value. The p-values indicate the significance of the slope for each of the explanatory variables. It shows that public expenditure, 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 9 years. However, the adjusted R^2 value is quite low at 8%, which suggests that this linear model is probably not a great fit for all the explanatory variables chosen and the assumptions in the model are not satisfactory. Nonetheless, the results might provide a useful direction for variable selection or quadratic fit 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(Future Steps)

In summary, we are able to conclude from the TAS dataset regarding different employee age levels, Public Expenditure area 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, along with the visualization tools, can also be applied to other datasets with factors such as the education level and department unit levels, to select the important explanatory variables for inference. There are many ways we could possibly explore, visualize and analyze due to the complexity of the dataset.

The following section explain the inherited problems the analysis may have:

1. The abnormally low adjusted R^2 value indicates a poor fit of the model. (Other polynomial fit is encouraged)
2. Many parts of the given data contain imputed dataset in which the method of imputation was not proved to satisfactory, this can be identified from the visualization plots.
3. Due to the nature of the data provided, we are not able to test the dependency of the factor variables between levels (for example, between different ages). Although a correlation test between these factor levels can be done, but the conclusion is hard to interpret.
4. One of the interesting issue of this panel regression analysis pointed out by Professor Gabriela Cohen is that we could not examine the variation between provinces as there are no way of measuring the variability for number of staff per province. It is not possible to produce replicated measures for each province.

For future recommendations of panel regression analysis, a random effect for time could be introduced into the model to relax the fixed effect assumptions for time. We also run multiple diagnostic tests as shown in

the Appendix section for panel regression to ensure the proper type of panel regression model for a better fit with assumptions.

Although missing values in the dataset has potential problems for the analysis, we are advised by the client to look further into the dimensions of the dependent variables itself and explore more about the variations within these 30 provinces. Therefore, we would likely to take the data as is and perform unsupervised learning algorithms to gain more insights. One possible way to achieve the goal is to group the provinces based on their staff number's behaviour over time. For each province's behaviour, a spline can be fitted, and then clustering analysis could be applied to suggest groupings of these splines that possibly indicate similar behaviours. A link of this method is provided in the Reference section.

References

link to reference for panel regression documentation: https://cran.r-project.org/web/packages/plm/plm.pdf https://www.princeton.edu/~otorres/Panel101.pdf

link to reference for function clustering documentation: https://cran.r-project.org/web/packages/Funclustering/Funclustering.pdf

Appendix

Appendix 1: Two-way fixed-effect panel regression model

The two-way fixed-effect panel regression model used 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 i in a specific year t. The matrix X_{it} are the possible economic explanatory variables that could explain the response variables, and β_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}. In this 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 $\epsilon_{it} \sim N(0, \sigma^2)$.

Appendix 2: R output for Panel Regression Analysis

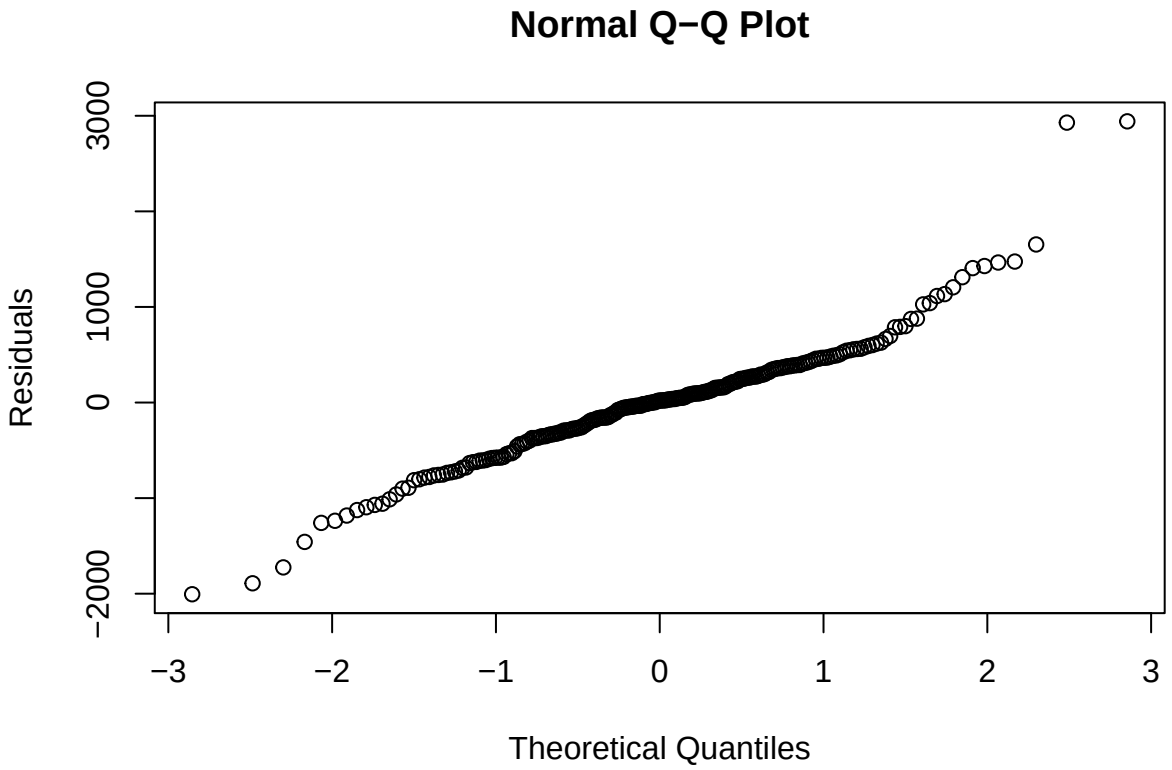
Two panel regression models are shown: one for number of staff at age <30, one for total number of staff. The following disgnostic tests checks for heteroskedasticity with residual plot, individual and time fixed-effect assumptions, and the Hausman test checks if each entity's error term and the constant are correlated with the others.

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

```
## Twoways effects Within Model
##
## Call:
## plm(formula = age.30.or..below ~ urbanarea + No.of.above.size.industrial.companies +
##     public.expenditure + GDPperCapita, data = agemerged, effect = "twoway",
##     model = "within", index = c("Year", "Province"))
##
## Balanced Panel: n = 8, T = 29, N = 232
##
## Residuals:
##    Min. 1st Qu.  Median 3rd Qu.    Max.
```

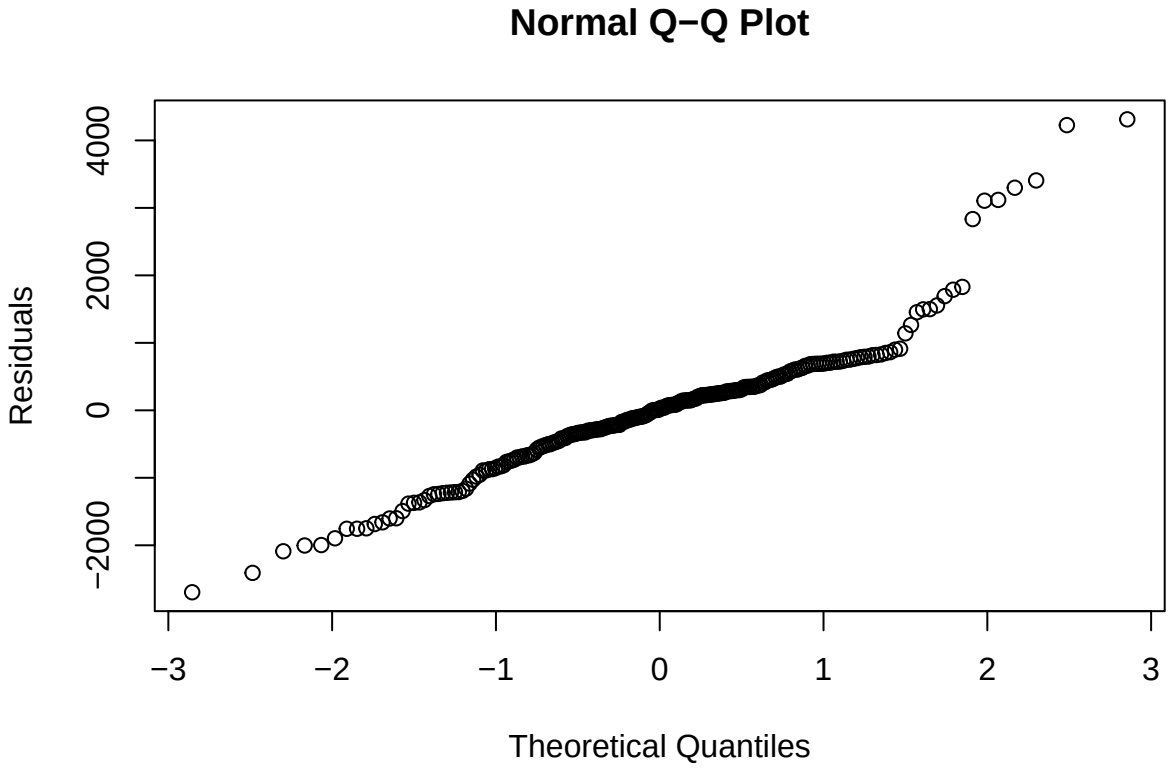
```
## -2010.0  -336.0    20.4   326.0  2940.0
##
## Coefficients:
##                                      Estimate Std. Error t-value
## urbanarea                           -0.229232   0.424973 -0.5394
## No.of.above.size.industrial.companies -0.023108   0.022936 -1.0075
## public.expenditure                  -2.979419   0.483227 -6.1657
## GDPperCapita                         0.104468   0.017639  5.9227
##                                      Pr(>|t|)
## urbanarea                              0.5902
## No.of.above.size.industrial.companies    0.3150
## public.expenditure                  4.057e-09 ***
## GDPperCapita                        1.438e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    142210000
## Residual Sum of Squares: 92917000
## R-Squared:      0.34662
## Adj. R-Squared: 0.21391
## F-statistic: 25.4644 on 4 and 192 DF, p-value: < 2.22e-16
```

```
## Twoways effects Within Model
##
## Call:
## plm(formula = Total..formal.employees ~ urbanarea + No.of.above.size.industrial.companies +
##     public.expenditure + GDPperCapita, data = agemerged, effect = "twoway",
##     model = "within", index = c("Year", "Province"))
##
## Balanced Panel: n = 8, T = 29, N = 232
##
## Residuals:
##    Min. 1st Qu.  Median 3rd Qu.    Max.
## -2700.0  -501.0    22.3   450.0  4310.0
##
## Coefficients:
##                                      Estimate Std. Error t-value
## urbanarea                            1.562387   0.664517  2.3512
## No.of.above.size.industrial.companies -0.076933   0.035864 -2.1451
## public.expenditure                  -2.404385   0.755607 -3.1821
## GDPperCapita                         0.076231   0.027581  2.7639
##                                     Pr(>|t|)
## urbanarea                           0.019729 *
## No.of.above.size.industrial.companies 0.033199 *
## public.expenditure                  0.001706 **
## GDPperCapita                        0.006268 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    254900000
## Residual Sum of Squares: 227190000
## R-Squared:      0.10871
## Adj. R-Squared: -0.072335
## F-statistic: 5.85446 on 4 and 192 DF, p-value: 0.00018203
```




```
##
##  Lagrange Multiplier Test - two-ways effects (Gourieroux, Holly
##  and Monfort) for balanced panels
##
## data:  age.30.or..below ~ urbanarea + No.of.above.size.industrial.companies +  ...
## chibarsq = 212.49, df0 = 0.00, df1 = 1.00, df2 = 2.00, w0 = 0.25,
## w1 = 0.50, w2 = 0.25, p-value < 2.2e-16
## alternative hypothesis: significant effects
```

```
##
##  Hausman Test
##
## data:  age.30.or..below ~ urbanarea + No.of.above.size.industrial.companies +  ...
## chisq = 193.69, df = 4, p-value < 2.2e-16
## alternative hypothesis: one model is inconsistent
```




```
##
##  Lagrange Multiplier Test - two-ways effects (Gourieroux, Holly
##  and Monfort) for balanced panels
##
## data:  Total..formal.employees ~ urbanarea + No.of.above.size.industrial.companies +  ...
## chibarsq = 331.82, df0 = 0.00, df1 = 1.00, df2 = 2.00, w0 = 0.25,
## w1 = 0.50, w2 = 0.25, p-value < 2.2e-16
## alternative hypothesis: significant effects
```

```
##
##  Hausman Test
##
## data:  Total..formal.employees ~ urbanarea + No.of.above.size.industrial.companies +  ...
## chisq = 62.982, df = 4, p-value = 6.846e-13
## alternative hypothesis: one model is inconsistent
```

Appendix 3:

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

Figures

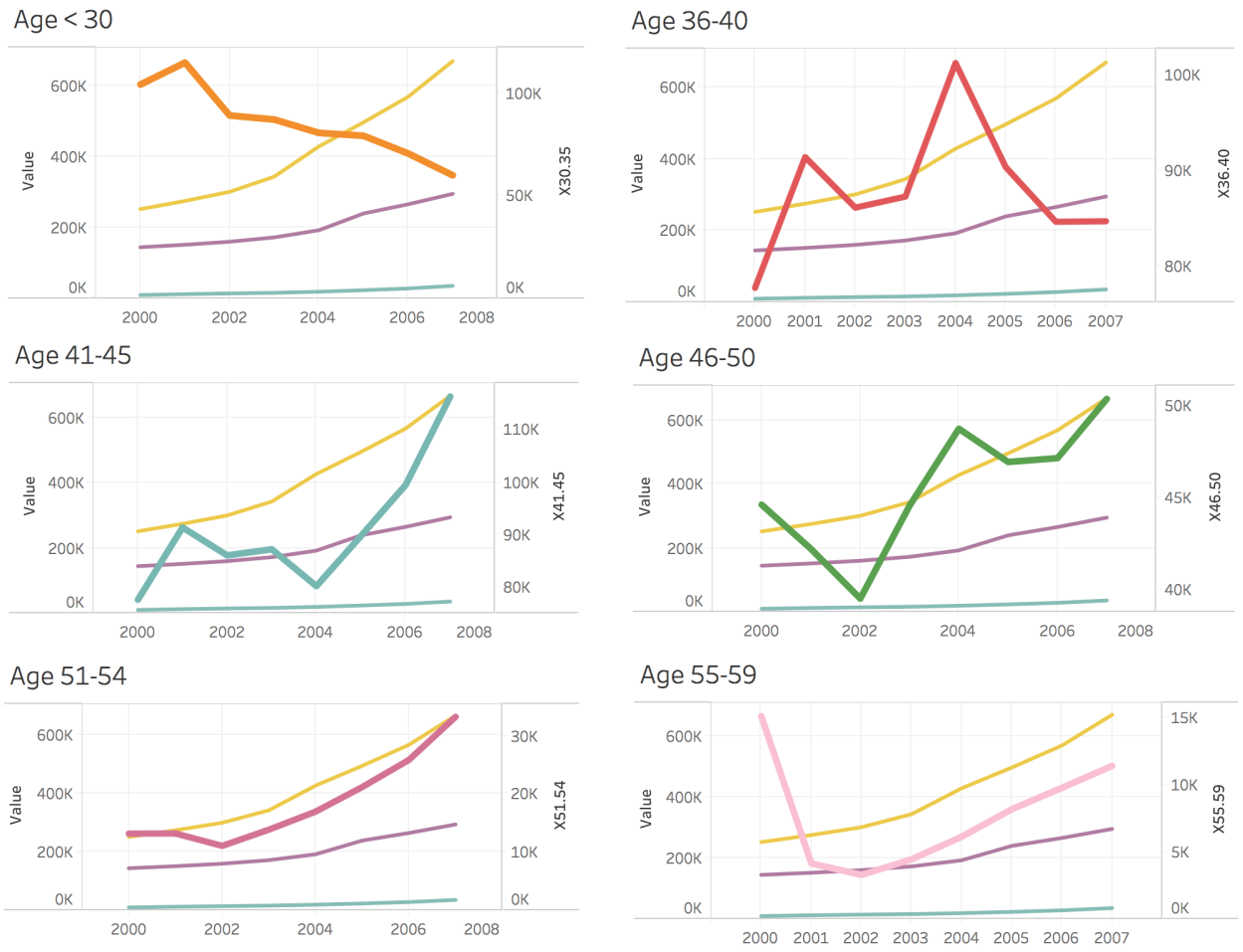


Figure 1: Age with Econ Variables

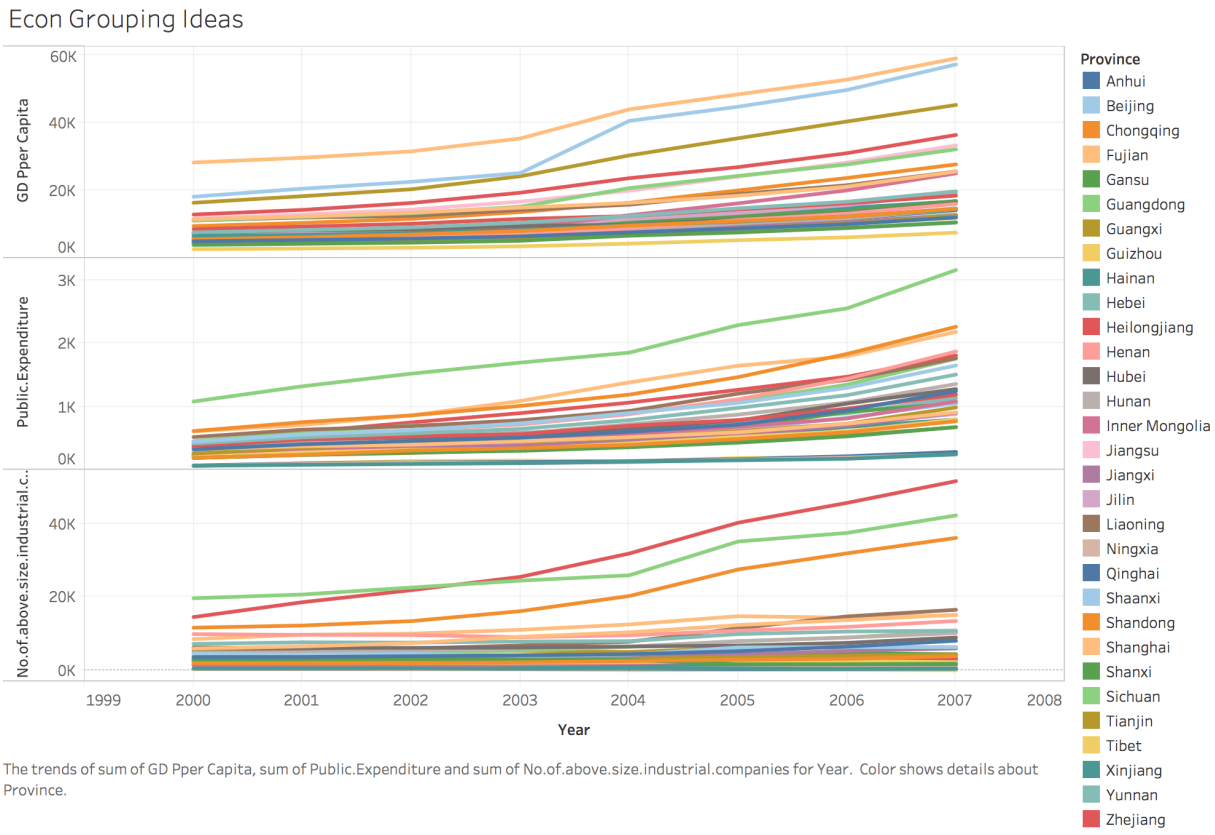


Figure 2: Econ Grouping Ideas

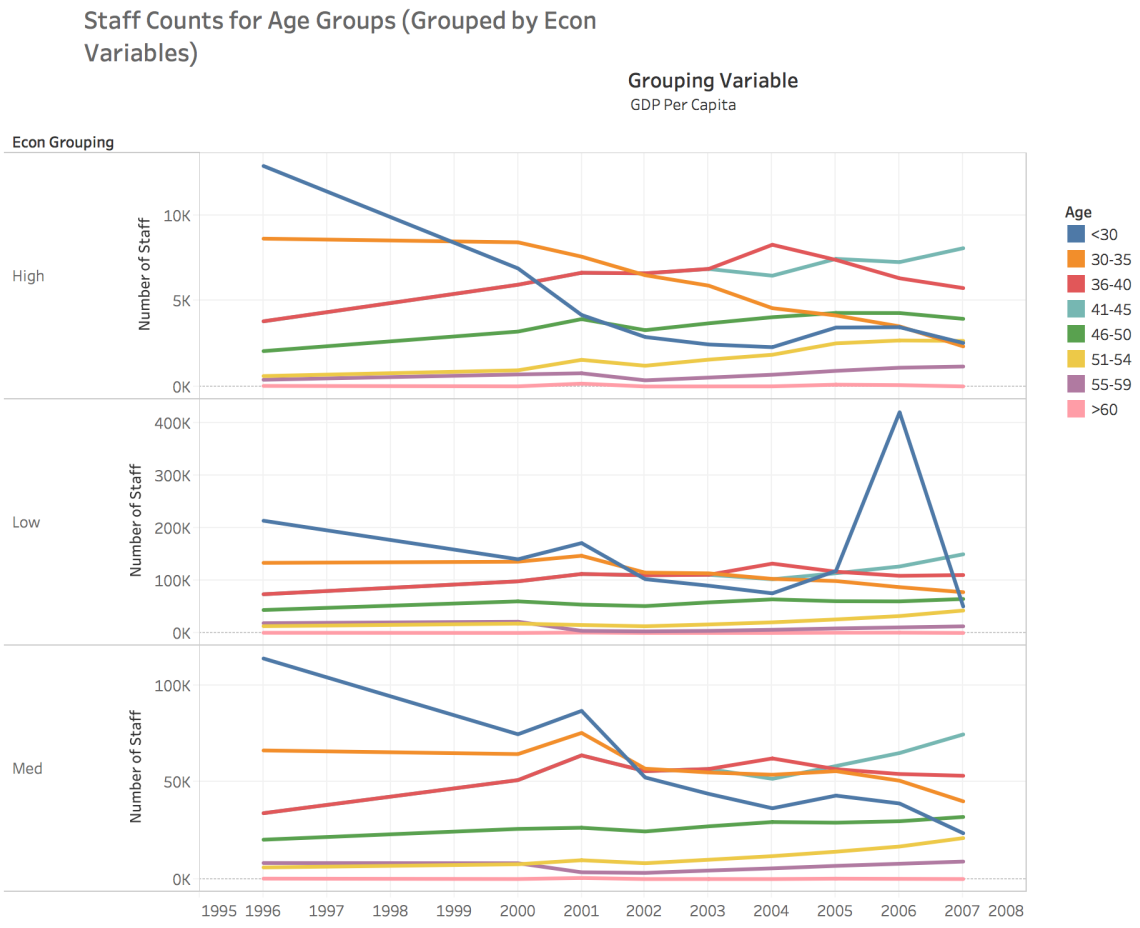


Figure 3: Age Dashboards

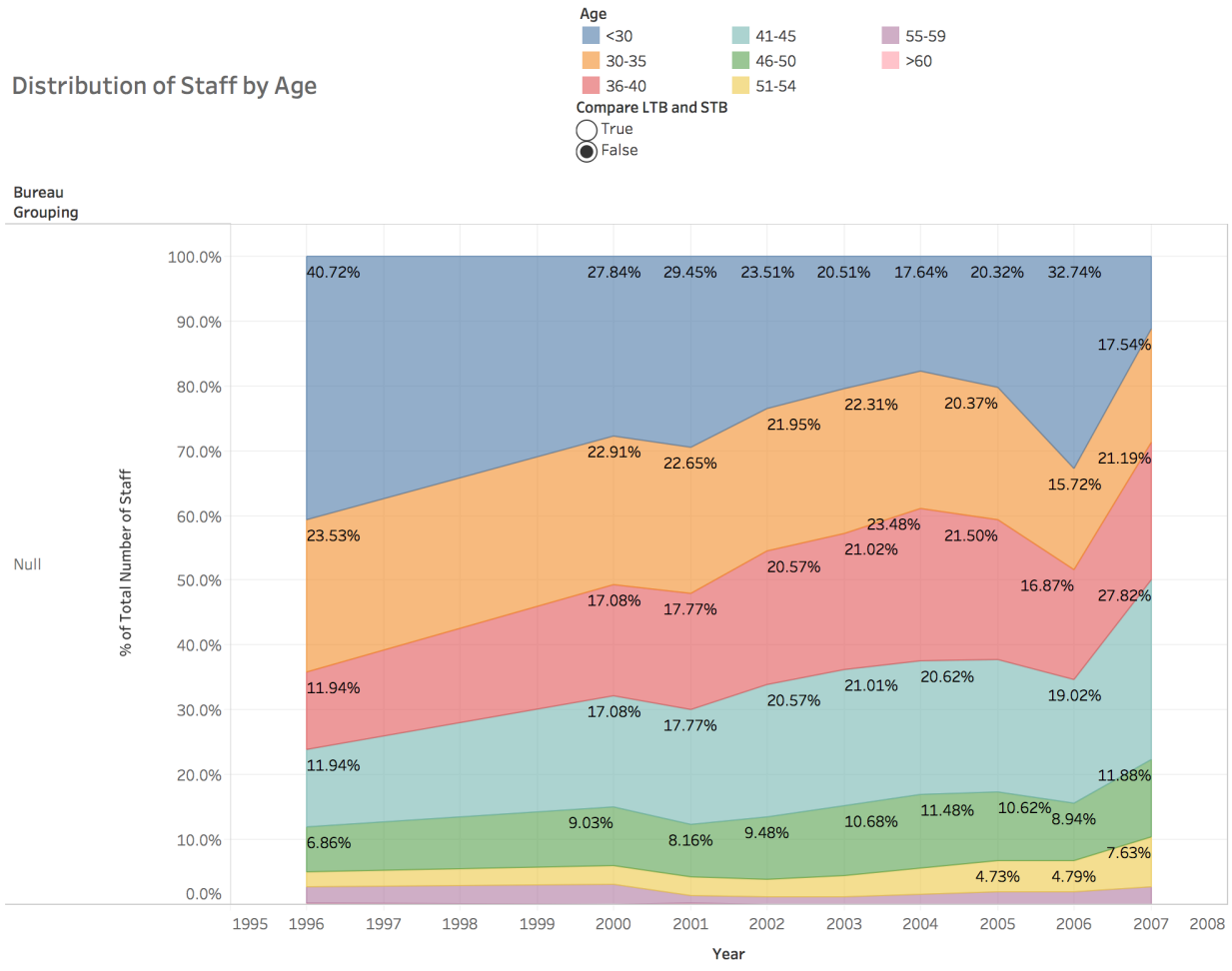


Figure 4: Age Distribution

Age - STB

- Under 30
  - GDP Per capital (6 x 10^-9)
  - Public expenditure (0.036)
- 30-35
  - Urban Area (1.92 x 10^-6)
  - Public expenditure (0.0055)
  - GDP per capital (0.0038)
- 36-40
  - 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
  - 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
  - Urban area: 0.0456
  - GDP per capita: 2.34 x 10^-6
- 30-35
  - 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
  - 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)

Figure 5: Significant Vars