Analysis of Staffing in China Civil Service Agencies

Katie Li, STAT 450

Summary

This report details the visualization and statistical analysis performed on staff counts in two China government agencies, Tax Administration and Environmental Protection. The total staffing numbers for various age, education, and unit levels were provided for each province across several years. For visualization, plots were generated showing how the composition of staff for age, education, and work unit has changed over time. Additionally, staffing was plotted alongside economic variables. Statistical models were performed to test if any economic or demographic variables had significant effects on the staffing numbers. It was found that GDP is consistently a highly significant predictor of staffing. However, due to limitations in the data and the model used, we will explore alternative methods of analysis in future work.

Introduction

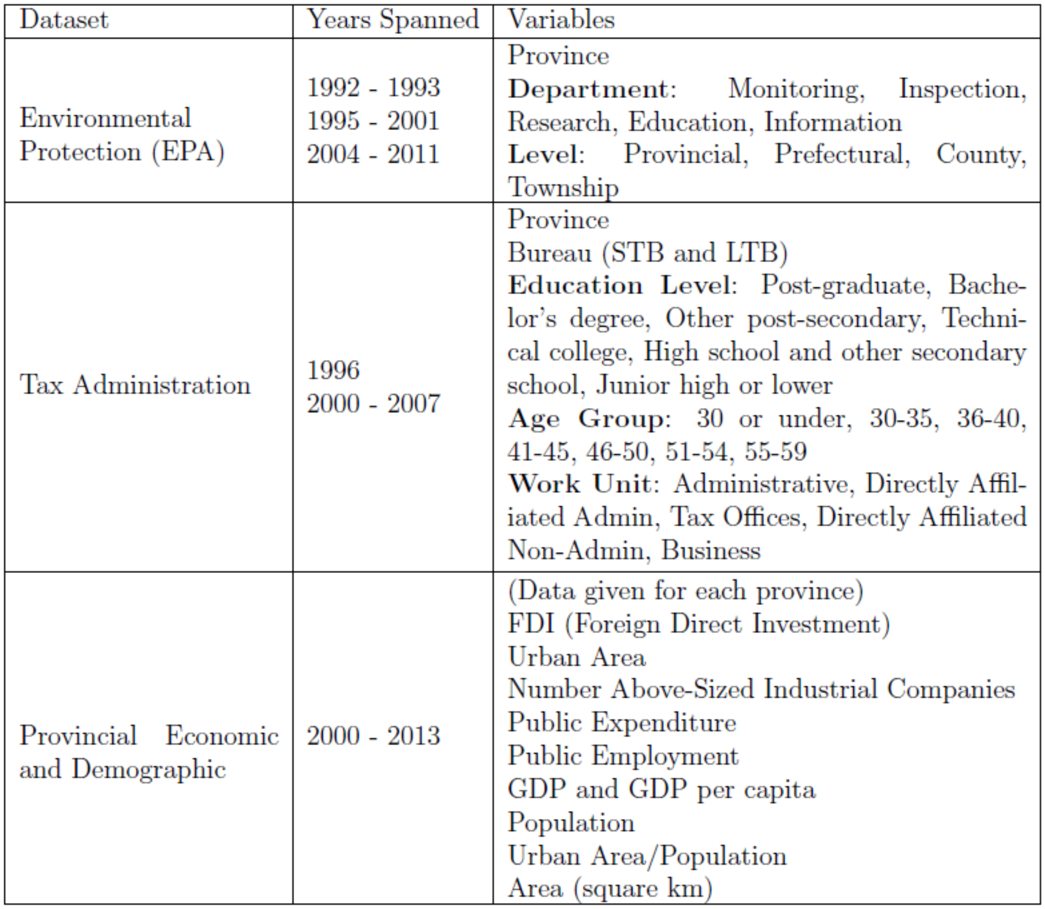
The Tax Administration and Environmental Protection (EPA) bureaus are two of the largest branches in China's civil service. Civil services in China are very decentralized and exist on the provincial, county, city, and prefecture levels. Several years of data on staffing numbers for these two government agencies have been provided. Over the period of data collection, there have been changes in both the composition of staff within a bureau and the overall counts. The objective of the project is to gain insight into these changes. A large component of the case, as requested by the client, is to develop meaningful and appealing ways to visualize the data. Another objective is to determine statistically what changes occurred in the organization of the two agencies and what factors caused staffing levels to change over time.

The report is structured as follows: Section 2 is an overview of the three datasets used throughout the case. Section 3 explains both the visualization and statistical modeling methodology. Section 4 provides and interprets the results from the methods mentioned in Section 3. Lastly, Section 5 details the next steps in the project.

Data

Two separate datasets of staffing numbers for the two agencies of interest were provided. A third dataset including several economic and demographic variables at the provincial level was also given by the client. See Table 1 for an overview of the datasets. The staffing data was hand-collected by research students as part of a study, while the economic indicators were taken from public data published in economic journals. The EPA and Tax datasets are structured differently due to differences in bureaucratic needs. Furthermore, the scope of the data varies throughout the years and across bureaus. For example, we can see in Table 1 that some years were excluded in each dataset.

The Tax dataset breaks down the total staffing counts for each province into groups based on age range, education level, and work unit. This data is given for each of two tax bureaus of interest, state tax (STB) and local tax (LTB). The client provided both the original dataset and a "partially-cleaned" dataset in which he imputed data from 2005 and 2006 to account for temporary workers. The EPA dataset breaks down the total staff into groups based on department and government level.



Dataset	Years Spanned	Variables
Environmental Protection (EPA)	1992 - 1993 1995 - 2001 2004 - 2011	Province **Department**: Monitoring, Inspection, Research, Education, Information **Level**: Provincial, Prefectural, County, Township
Tax Administration	1996 2000 - 2007	Province Bureau (STB and LTB) **Education Level**: Post-graduate, Bachelor's degree, Other post-secondary, Technical college, High school and other secondary school, Junior high or lower **Age Group**: 30 or under, 30-35, 36-40, 41-45, 46-50, 51-54, 55-59 **Work Unit**: Administrative, Directly Affiliated Admin, Tax Offices, Directly Affiliated Non-Admin, Business
Provincial Economic and Demographic	2000 - 2013	(Data given for each province) FDI (Foreign Direct Investment) Urban Area Number Above-Sized Industrial Companies Public Expenditure Public Employment GDP and GDP per capita Population Urban Area/Population Area (square km)

Table 1: Summary of Datasets

Methodology

The Tableau software was used to visualize the data. To make the data compatible with Tableau, we converted all the data files into long format. A useful feature that was utilized throughout the visualization process was Tableau's ability to group variables. Since there are 30 provinces, we wanted to group them to avoid showing 30 lines on each graph. A grouping was made for several economic and demographic variables; as an example, for GDP per capita, provinces were grouped into high, medium, and low categories (all relative to each other). The appendix contains images of all dashboards created.

For the statistical analysis, panel regressions were performed for the Tax Adminstration dataset, using the `plm` package in R. For each level of age range, education level, and work unit, a panel linear model was fitted with the provincial economic and demographic variables as explanatory variables. Here is a sample call to plm, in which we estimate the effects of five covariates on the counts of staff that hold a Bachelor's degree:

```r
plm(Bachelor.s.degree ~ urbanarea + No.of.above.size.industrial.companies +
    public.expenditure + GDPperCapita + GDP, data = edumerged, index = c("Year",
    "Province"), model = "within", effect = "twoway")
```

In the `plm` command, `index = c("state","year")` defines the first variable (Province) as the entity/individual and second (Year) as the time variable. The `plm` call allows the user to choose the model type and effects desired. For the model parameter, a "within" or fixed effects model was chosen. A fixed effect model allows for each province to have an individual effect - each coefficient is interpreted as a within-province effect.

For the effect parameter, we opted to include both individual and time effects, hence the "twoway" selection. It is plausible that there are both sources of variability that vary across province but not over time (individual fixed effects) and sources of variability that vary across time but affect each province similarly (time fixed effects). In further analysis, we would like to run diagnostic tests to confirm whether our assumptions of individual and time fixed effects are appropriate.

Results

Visualization Results

Distribution of age, education level, and work unit for Tax dataset

To look at the composition of staff based on age, education and work unit, we plotted the staff count for each level (as a percentage of total) as a stacked line graph to see how the distribution of age, education, and unit varied over time.

Figure 1 shows the composition of staff based on education level; the staff counts are pooled across all provinces. The percentage of total staff that had a Bachelor's degree grew rapidly,

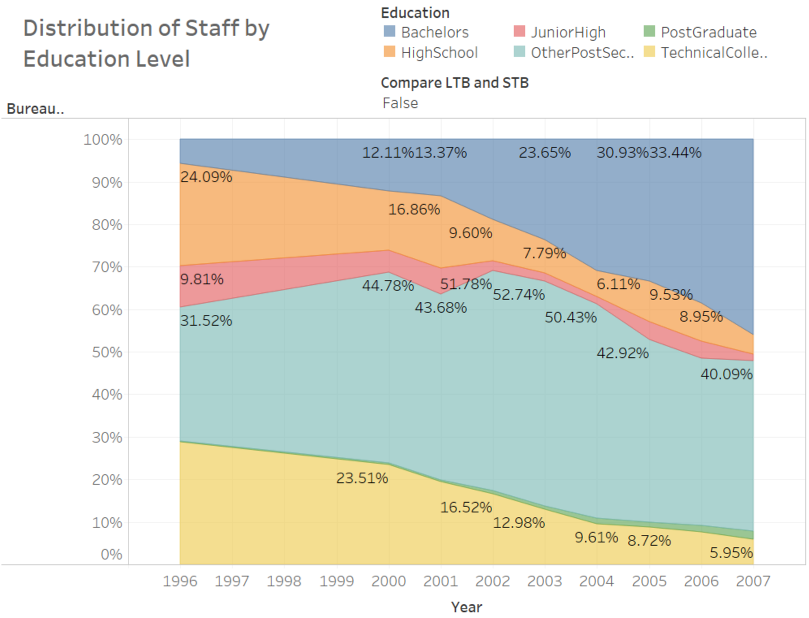


Figure 1: Distribution of Staff By Education Level

from only 5.6% in 1996 to 46% in 2007. On the other hand, the percentage of staff with a Technical college background consistently decreased each year. The education group with the largest percentage of staff is other post-secondary education - it grew from 32% in 1996 to a maximum of 53% in 2003 but then decreased yearly, reaching 40% in 2007.

Figure 2 shows the composition of staff based on work unit. There is one plot for each bureau, LTB and STB. The work unit dataset has years 2000- 2003. In 2000, the Tax Office unit had the most staff, with 63%, however, this percentage decreased yearly and in 2003, the percentage of staff working in a tax office was only 34%. This decrease was most pronounced in the STB bureau group, as the percentage of STB staff in Tax offices decreased from 62% to 27%.

Staffing and economic factors

We also plotted staffing counts against the trends of various economic variables. In Figure 3, four education levels are shown plotted with three economic variables, GDP per capita, number of above size industrial companies, and public expenditure. Of the four education levels shown, the trends of Bachelor's degree and post-graduate seem to be most related to those economic variables. We used panel regression to test this hypothesis.

Panel Regression Results

As mentioned in the previous section, a fixed-effects panel linear model was fitted for each level of age, education, work unit, and for each of the state and local tax bureaus. The table in the appendix summarizes the results of the fixed-effects regression, with each model's

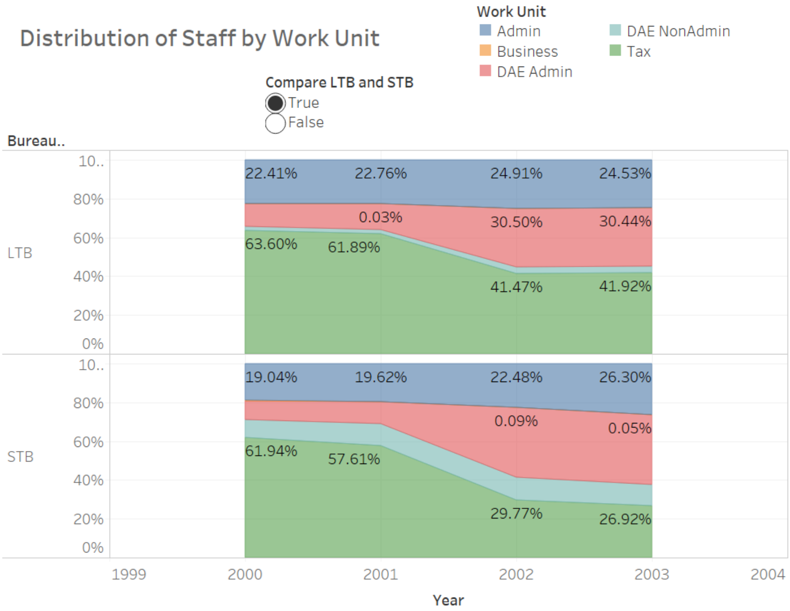


Figure 2: Distribution of Staff By Work Unit

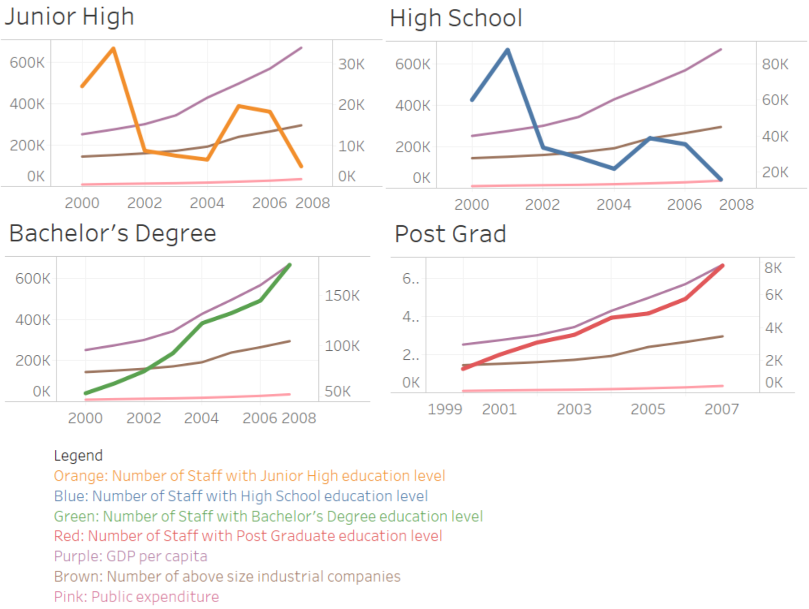


Figure 3: Staffing and Economic Variables for Four Education Levels

significant regressors.

As an example, consider the staffing numbers for the Bachelor's education group in the State Tax Bureau. The least squares equation can be written as:

$$
\begin{aligned}
\widehat{\text{Staff Count}}_{i,t} = &-0.87 * \text{Urban Area}_{i,t} + 0.03 * \text{Number Above Sized Industrial Companies}_{i,t} \\
&+ 1.69 * \text{Expenditure}_{i,t} + 0.30 * \text{GDP}_{i,t} - 0.1 * \text{GDP per Capita}_{i,t} \\
&+ \textit{Province effects}_i + \textit{Time effects}_t
\end{aligned}
$$

Here, i refers to a particular province while t refers to a particular year. Each coefficient can be interpreted as the contribution that a one unit change in the explanatory variable makes on the number of staff - this contribution is the same for all provinces and all time points. In addition to the contribution by the explanatory variabe, each province has its own province effects and each year has time effects that contribute to the overall staff count as seen by the equation.

In the example model, Public Expenditure, GDP, and GDP per Capita are significant predictors of staffing at the Bachelor's education level. In fact, in most models, GDP per Capita and GDP are both significant. However, looking at the coefficients, GDP per Capita has a negative coefficient, while GDP has a positive one. GDP per Capita is just a transform of GDP, so both should not be included in the model. In Figure 3, it appears that GDP is growing at a faster rate than Staff (GDP is the purple line). Since we treat all the regressors as a linear term, it is likely the additive effect of GDP and GDP per capita that is significantly affecting the staffing level. If we try to optimize the model in the future, we will want to remove GDP and only model GDP per capita with a non-linear term.

Some more results from the panel models:

- For age groups 55-59 and 60 and above, none of the regressors were significant predictors of staff counts. This was the case for both State and Local tax bureaus.
- For education, GDP per Capita was a highly significant variable, showing up in 9 out of 12 models.

Generally, it seems that GDP, GDP per Capita, and Public Expenditure have a consistently strong association with staffing level, although their appearance as statistically significant varies depending on which age group, education level, or work unit is being modeled. A larger concern is that the models do not consider the relationship between the levels in age/education/unit themselves. For example, the staffing of the 55-59 age group may be strongly correlated with the 60 and above age group, but the model cannot take that into account. The best we can do is run a linear regression for each level separately, which is what has been done. Due to the limitations of the dataset as well as the limitations of the `plm` package in R, we will want to try other modeling methods in the future.

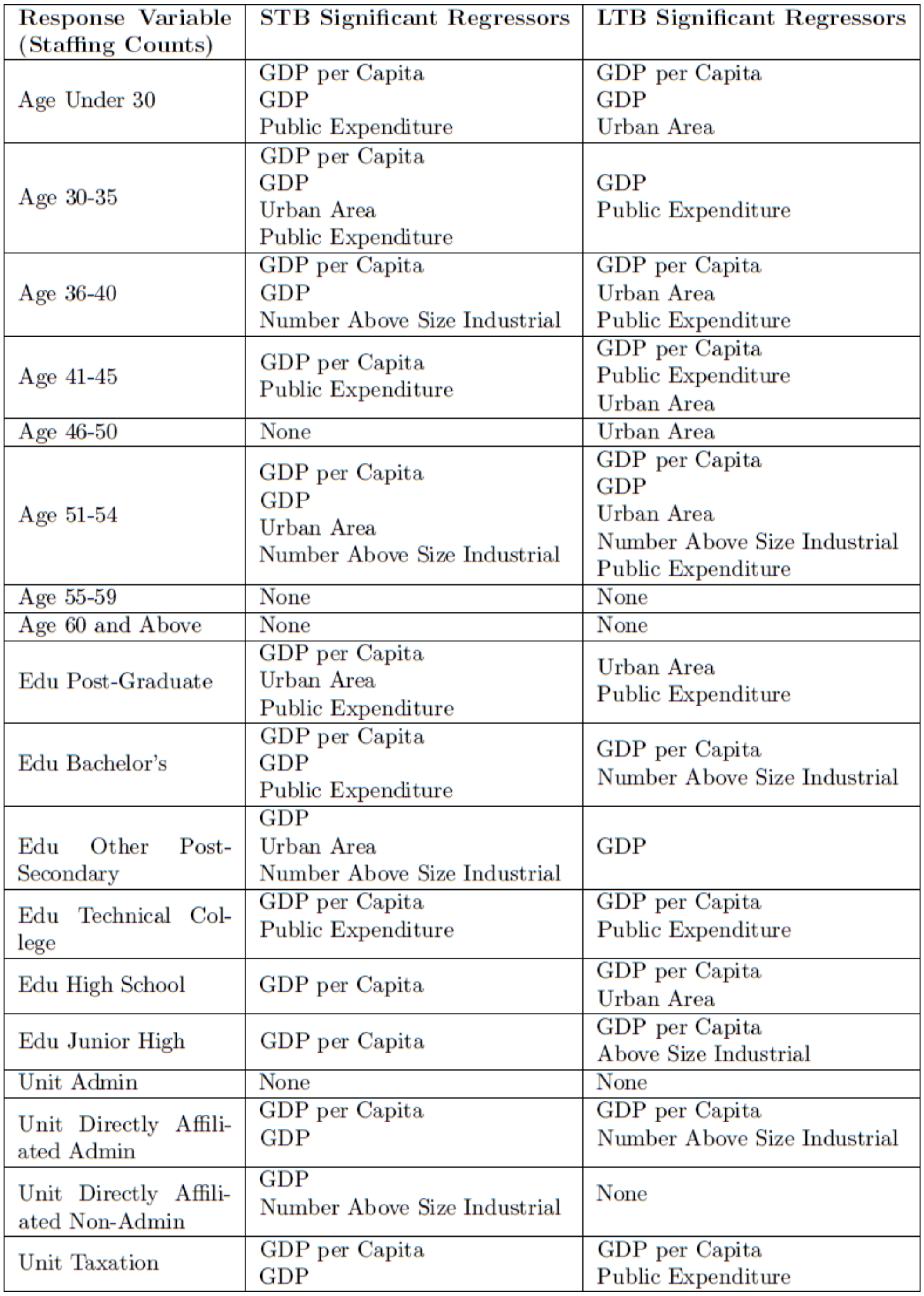
Next Steps

After last Thursday's meeting, the client has indicated that he would like us to focus more on provincial variation rather than trends over time. The client feels that because government staffing tends to be "sticky", looking at changes in staffing over time is not as interesting as looking at differences between provinces or government levels. Because of this, we can explore other models besides panel regression. For example, we can use ANOVA to see if the mean change in staff from 2002 to 2003 across all provinces is the same.

Regarding panel regression, we can try to optimize the models we have found so far. We will want to try the diagnostic tools in `plm` to confirm the parameters we have chosen for effect type. However, panel regression will not allow us to gain insight into provincial variation since the provinces are the replicates in the model. We will try unsupervised learning methods, such as functional clustering. This will allow us to extract information related to which provinces behave similarly in each cluster, what those similar patters are, and what variations between clusters exist.

Appendix

Significant Covariates from Panel Linear Model



Response Variable (Staffing Counts)	STB Significant Regressors	LTB Significant Regressors
Age Under 30	GDP per Capita GDP Public Expenditure	GDP per Capita GDP Urban Area
Age 30-35	GDP per Capita GDP Urban Area Public Expenditure	GDP Public Expenditure
Age 36-40	GDP per Capita GDP Number Above Size Industrial	GDP per Capita Urban Area Public Expenditure
Age 41-45	GDP per Capita Public Expenditure	GDP per Capita Public Expenditure Urban Area
Age 46-50	None	Urban Area
Age 51-54	GDP per Capita GDP Urban Area Number Above Size Industrial	GDP per Capita GDP Urban Area Number Above Size Industrial Public Expenditure
Age 55-59	None	None
Age 60 and Above	None	None
Edu Post-Graduate	GDP per Capita Urban Area Public Expenditure	Urban Area Public Expenditure
Edu Bachelor's	GDP per Capita GDP Public Expenditure	GDP per Capita Number Above Size Industrial
Edu Other Post-Secondary	GDP Urban Area Number Above Size Industrial	GDP
Edu Technical College	GDP per Capita Public Expenditure	GDP per Capita Public Expenditure
Edu High School	GDP per Capita	GDP per Capita Urban Area
Edu Junior High	GDP per Capita	GDP per Capita Above Size Industrial
Unit Admin	None	None
Unit Directly Affiliated Admin	GDP per Capita GDP	GDP per Capita Number Above Size Industrial
Unit Directly Affiliated Non-Admin	GDP Number Above Size Industrial	None
Unit Taxation	GDP per Capita GDP	GDP per Capita Public Expenditure

Dashboards from Tableau

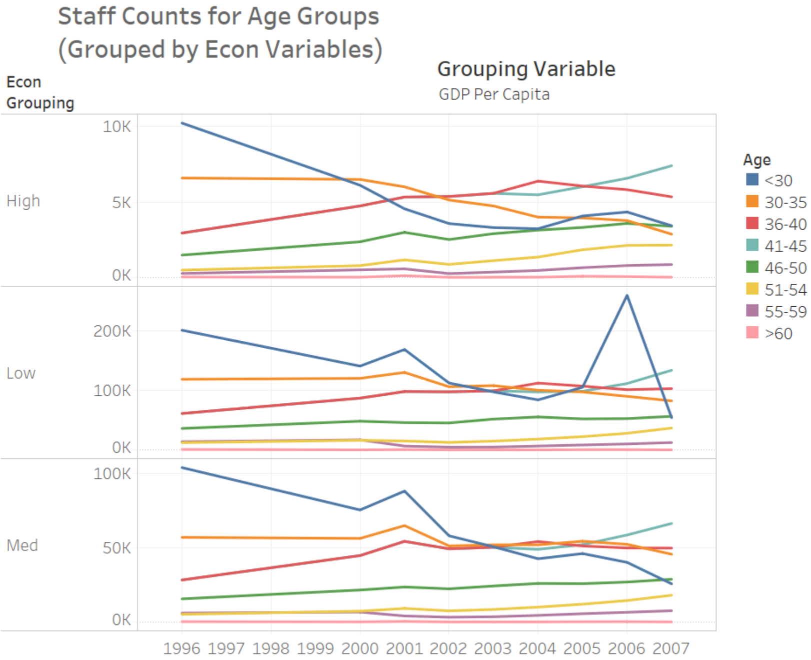


Figure 4: Staffing for Age Groups, Provinces grouped by Econ Variable

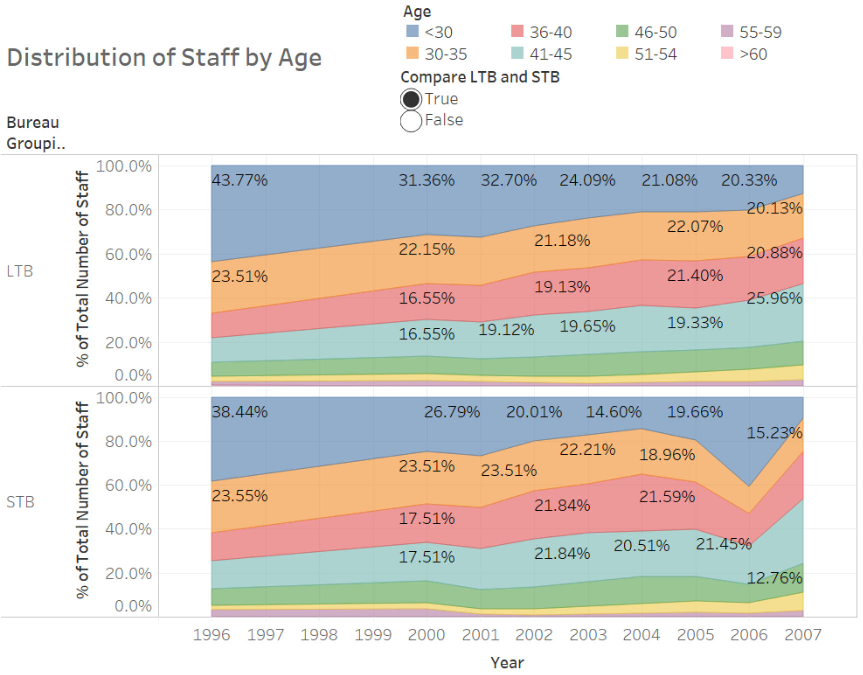


Figure 5: Distribution of Staff by Age Group

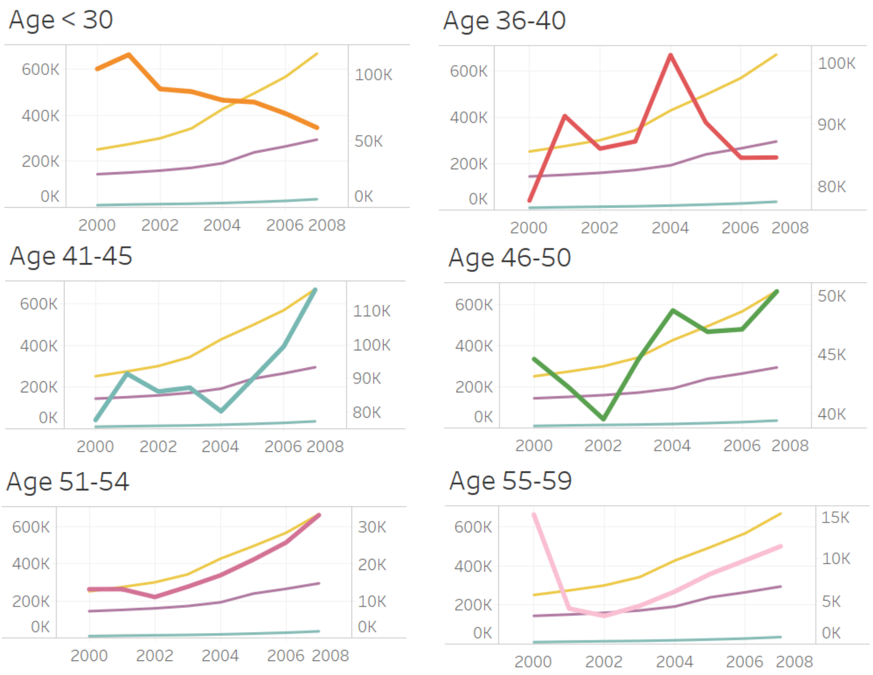


Figure 6: Staff and Economic Variables for Six Age Groups

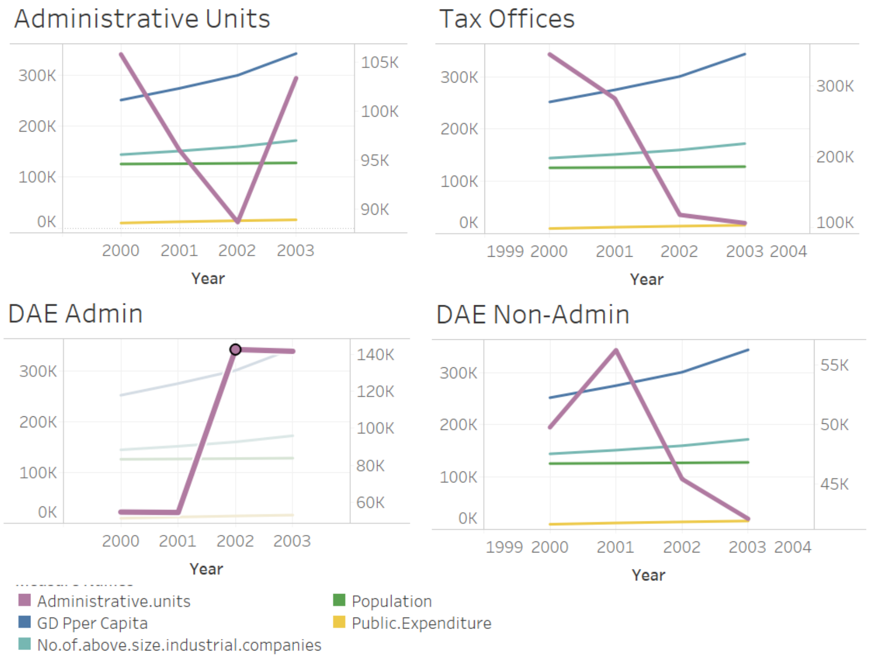


Figure 7: Staff and Economic Variables for Four Work Unit Groups

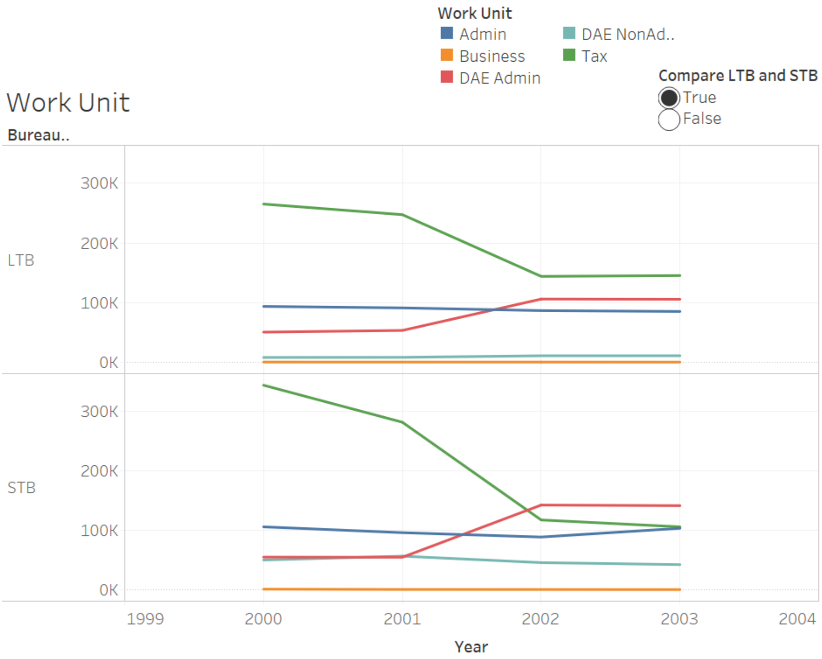


Figure 8: Staffing for Work Unit Groups, Separated by Bureau

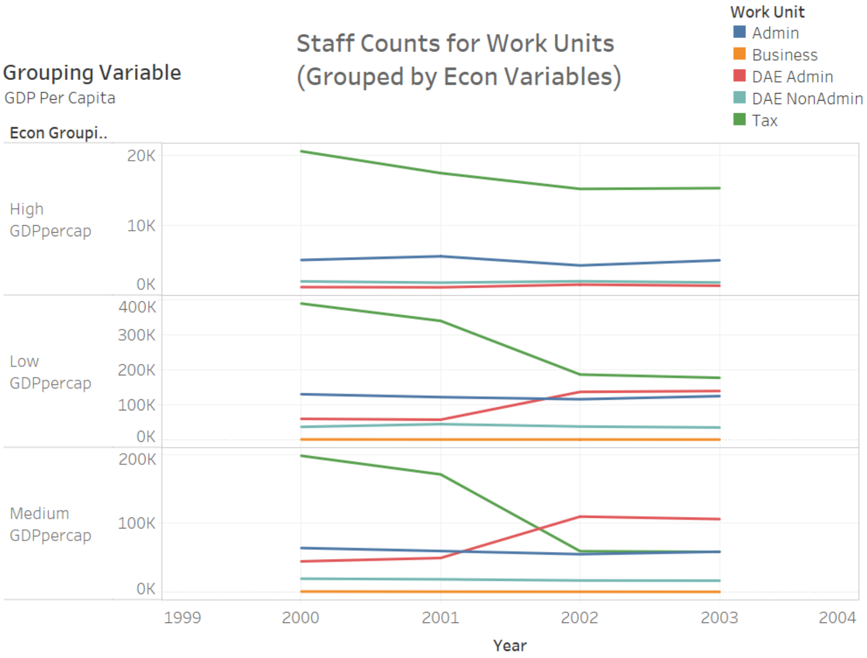


Figure 9: Staffing for Work Unit Groups, Provinces grouped by Econ Variable

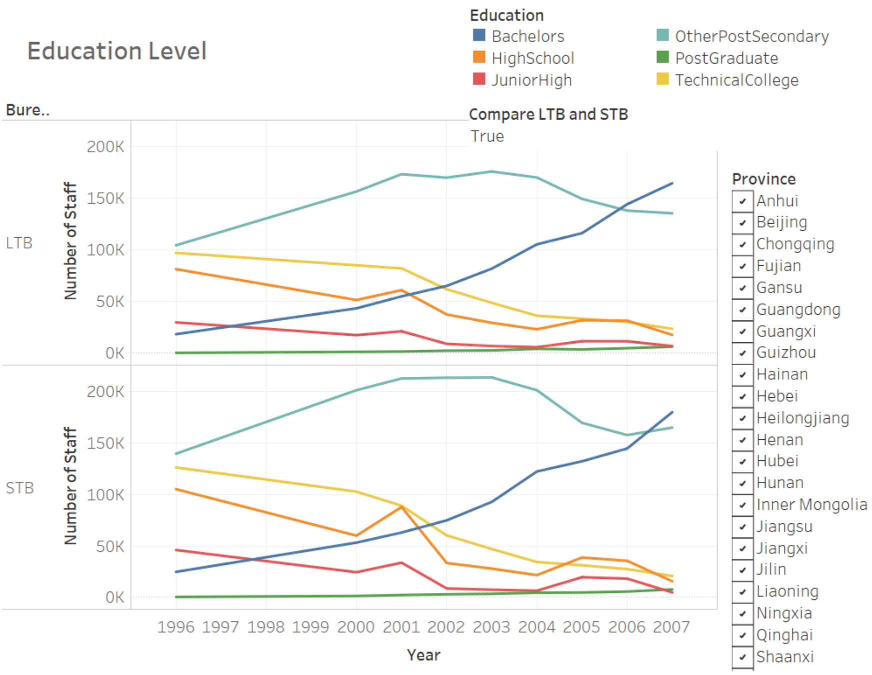


Figure 10: Staffing for Education Groups Groups, Separated by Bureau

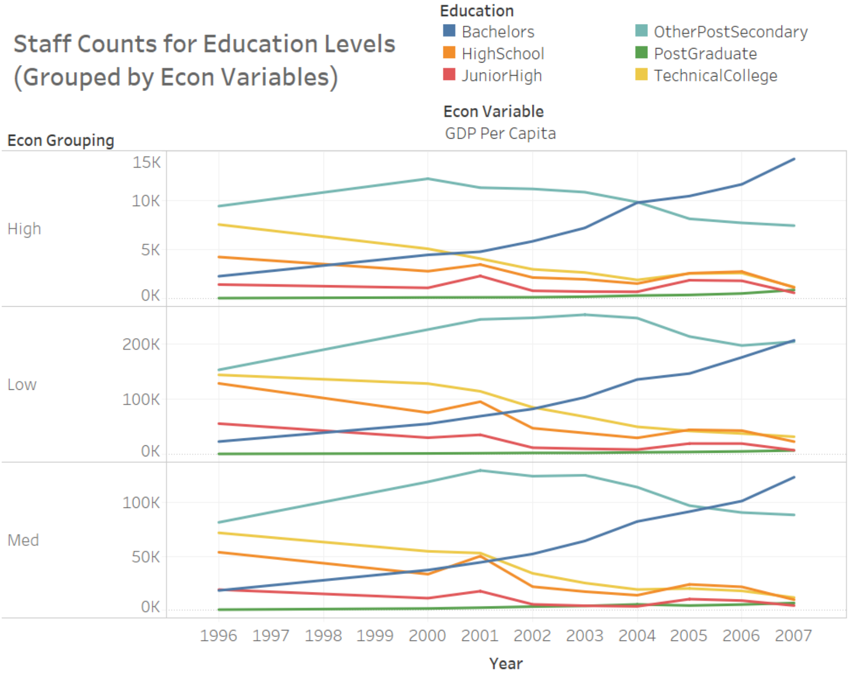


Figure 11: Staffing for Education Groups, Provinces grouped by Econ Variable