STAT 450 : Government Agencies Report 1

Jenny Hui

March 6, 2018

SUMMARY

This report contains an overview of the data visualizations and statistical analysis on staffing values for two largest branches of the China's civil service, namely, Taxation Administration Services (TAX) and the Environmental Protection Bureau (EPB). Panel datasets were provided for both these services, as well as relevant province-specific economic and demographic variable values. Visualizations through Tableau directed our analysis which focused on panel regressions. Based on a preliminary analysis on the TAX datasets, we came to the conclusion that GDP, GDP per capita, the number of above-size industries, and public expenditure are significant variables in explaining the staffing variations.

INTRODUCTION

As a country with the world's largest population and is constantly economically expanding, it is of interest to examine the compositon and structure of China's civil services. Data was collected within a 10-year span on the staffing numbers for two of China's largest civil service branches being the Tax Administration Services (TAX) and the Environmental Protection Bureau (EPB). The general objectives of this open study and analysis is to: 1) create clear and informative data visualizations, 2) formulate hypotheses to test the significance of various factors in the staffing changes, and 3) model the staffing values at various levels. We will be approach these objectives first by creating dyanimic data visualizations through the Tableau software. The visual results will then inform the direction of the statistical analysis, which will be focused on panel regressions given the structure of the data. The rest of the report is presented in the following format: 1. Data Description, 2. Methods, 3. Results & Conclusions, 4. References, 5. Appendix, and 6. Figures.

DATA DESCRIPTION

The provided data consists of following three panel datasets: 1. TAX This dataset contains relevant data pertaining to the staffing details of the Tax Agencies. The response variable is the staffing numbers which, as a panel dataset, is shown to vary over time. The data also contained many multi-level factors such as Age, Education Level, Government Level, Province, Department, and Bureau. After an initial look-through, it appears that the time range of the available staffing data varies depending on which factor we are looking at. In addition, we were provided with another dataset containing Tax Agency staffing numbers on a National level which were separated into similar factors mentioned above.

2. Environment Protection Bureau (EPB) This dataset contains relevant data pertaining to the staffing details of the Environmental Protection Bureau. The EPB dataset is smaller compared to the TAX data, spanning from 1992-1993 and then from 1995-2011. The response variable and multi-level factors remains the same, with the exception of Bureau.
3. Economic & Demographic Variables (EDV) As its name suggests, this dataset contains various economic and demographic variables such as GDP, FDI, Area/KM2 for each of the 30 provinces in China.

With the exception of the EPB, there is a noticeable amount of missing data spread throughout the different datasets.

METHODS

Due to the complexity of the provided datasets, we had decided to first perform an exploratory data analysis via Tableau. Through the software, we were able to generate dynamic visualizations of overall trends and patterns of the different dataset, as well as the interactions between them; this is one of the key objectives for this study. (Refer to Figures 1 - 3 for sample visualizations)

Our approach to the data visualizations can be generalized into two themes: between-datasets and within-datasets. Visualizations belonging to the 'between group explores the relations between TAX and EPB staffing numbers, and the values for different economic and demographic variables. Some motivating questions may include whether a rise in provincial GDP per capita correspond with an increase in staffing for that province, or whether decreases in provincial public expenditures would correspond with an increase in staff with post-graduate education. Visualizations pertaining to the 'within' group explores the trends and patterns found within the different datasets. An item of interest here may be how the decomposition of staffing into different education levels change from 1996 to 2007. Furthermore, do these changes occur consistently for all provinces or are there some provinces that deviate from the trend? Links to the Tableau workbooks containing the full set of data visualizations can be found in the Appendix.

Based on the results from the visualizations, we had then began a preliminary analysis on the TAX dataset. Two-way linear model with Province and Time as fixed effects were created for each level for the factors Age, Education Level, and Role, using the plm package for R. These models provided indications of significant EDVs that may have contributed to the changes in the staffing values. These results were then used to hone existing data visualizations and will be a key component in future analyses.

RESULTS & CONCLUSIONS

In addition to the generalized results that can be directly read off of the visualizations, our preliminary analysis of the TAX dataset showed that there are four economic / demographic variables that are significant in explaining the variations in the TAX staffing values, namely, GDP, GDP per capita, the number of above-size industries, and public expenditure. Although in varying effects, these four variables are significant to most of the different levels for each categorical factor on the dependent variable (staffing values). A complete summary of the p-values are provided in the link in Appendix 2. These results will be a starting point for future analysis.

The next steps for this project will be focused on the statistical analysis component; we would like to continue formulating and testing new hypotheses to explain what we observe from the visualizations. Based on the most recent meeting however, it has also been recommended to prioritize finding the dimensions of variation within the different levels of the dependent variables. For example, consider Figure 4 which shows the distribution of the ratio of staff for various education levels, separated by the local and state tax bureaus. A hypothesis of interest may be whether the proportion of change in staffing numbers for the different education levels are varying in the same way for each province. Appropriate functional data analysis techniques will be evaluated and considered as we change the pivot the focus of this project.

REFERENCES

1. Documentation for the plm package in R : https://cran.r-project.org/web/packages/plm/plm.pdf
2. Tableau software : https://www.tableau.com/

APPENDIX

1. Preliminary TAX Analysis - Rcode : https://github.com/wenzhengzzz/STAT450-550-Project-GovtAgencies/blob/master/stat450/Visualizations/visualization_notes/initial-TAX-analysis.R
2. Preliminary TAX Analysis - Results : https://github.com/wenzhengzzz/STAT450-550-Project-GovtAgencies/blob/master/stat450/Visualizations/visualization_notes/Tax-Personel-Significant-Factors.docx/
3. Visualizations

a. TAX : https://github.com/wenzhengzzz/STAT450-550-Project-GovtAgencies/blob/master/stat450/Visualizations/Age-Unit-Edu-Final.twb
b. EPB : https://github.com/wenzhengzzz/STAT450-550-Project-GovtAgencies/blob/master/stat450/Visualizations/EPB-vis.twb
c. EDV : https://github.com/wenzhengzzz/STAT450-550-Project-GovtAgencies/blob/master/stat450/Visualizations/demo-econ-vis.twb

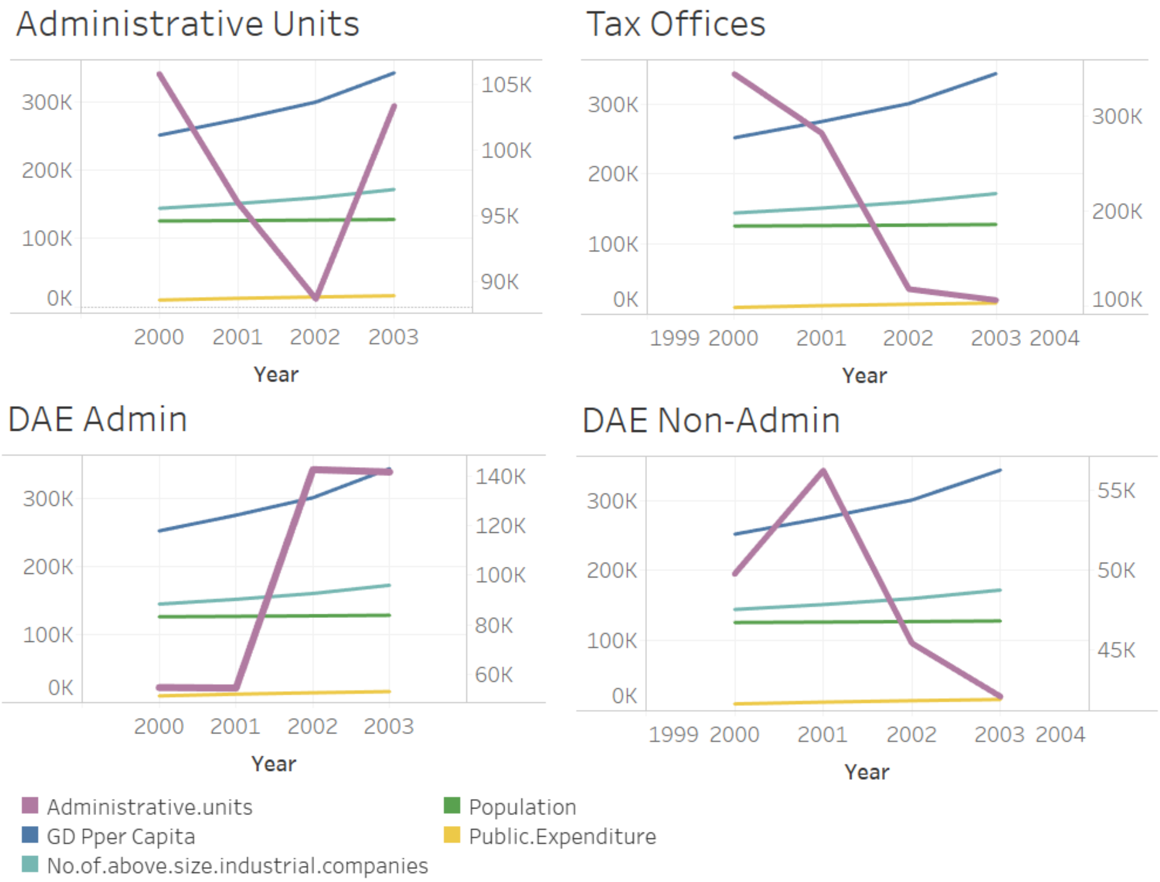

Figure 1: Sample “between-dataset” visualizations : TAX & EDV

FIGURES

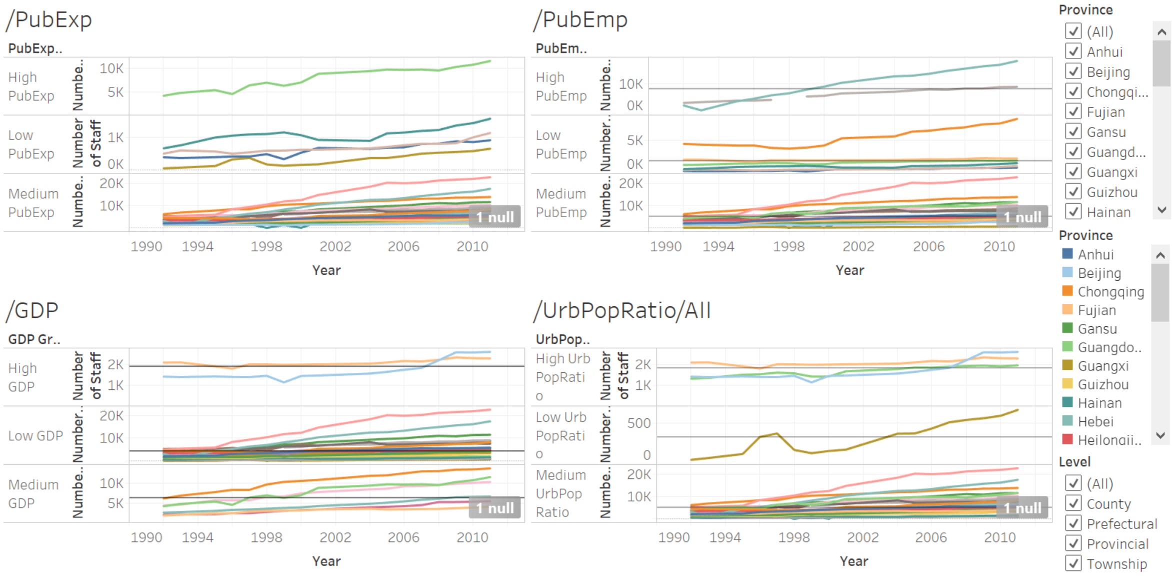

Figure 2: Sample "between-dataset" visualizations : EPB & EDV

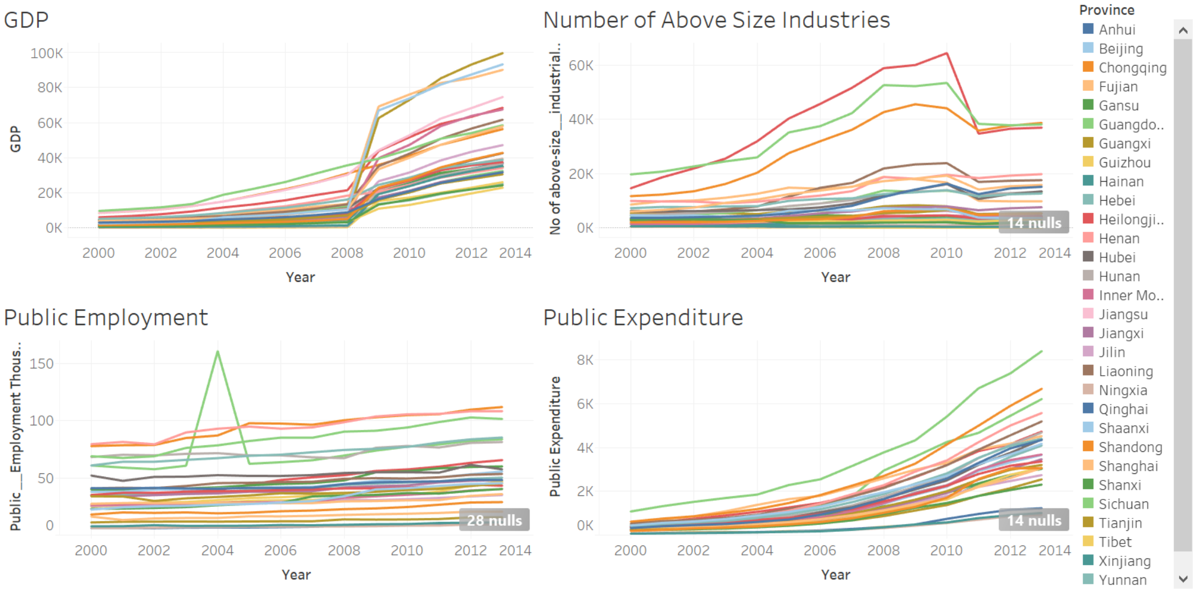

Figure 3: Sample "within-dataset" visualizations.

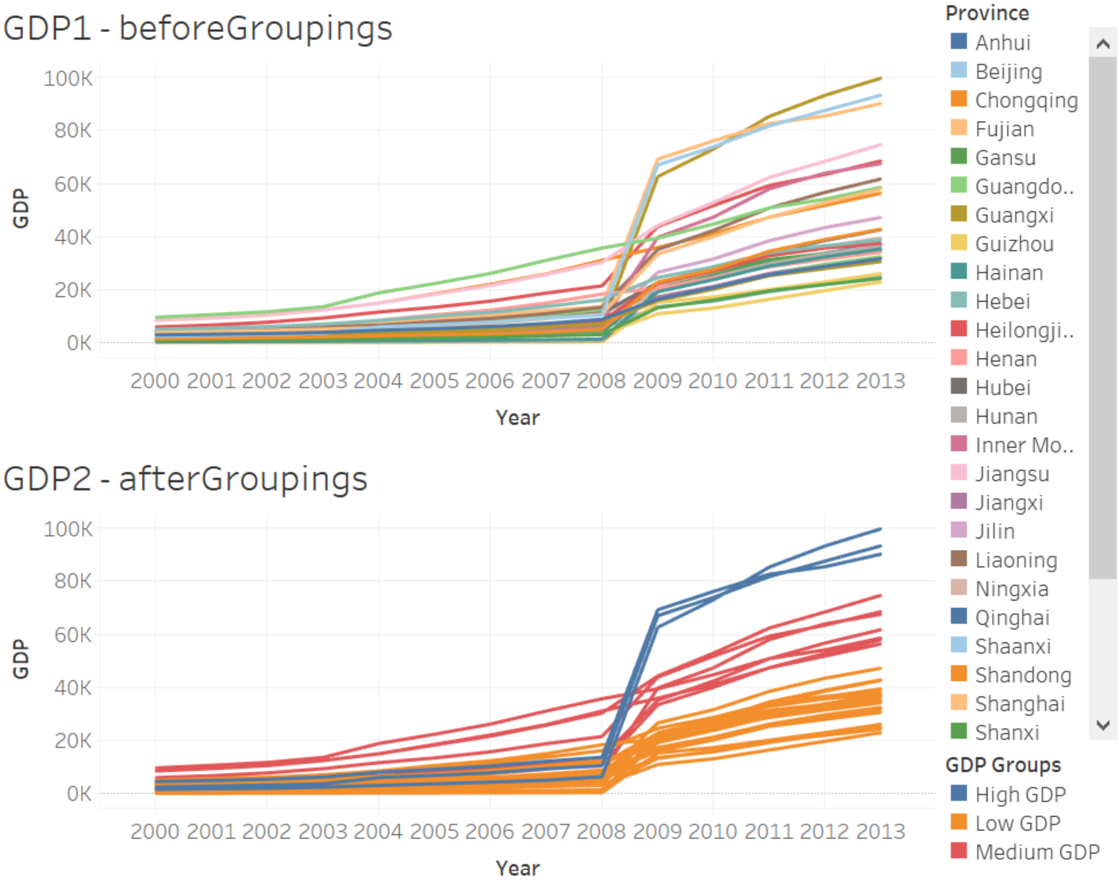

Figure 4: Sample GDP Groupings.

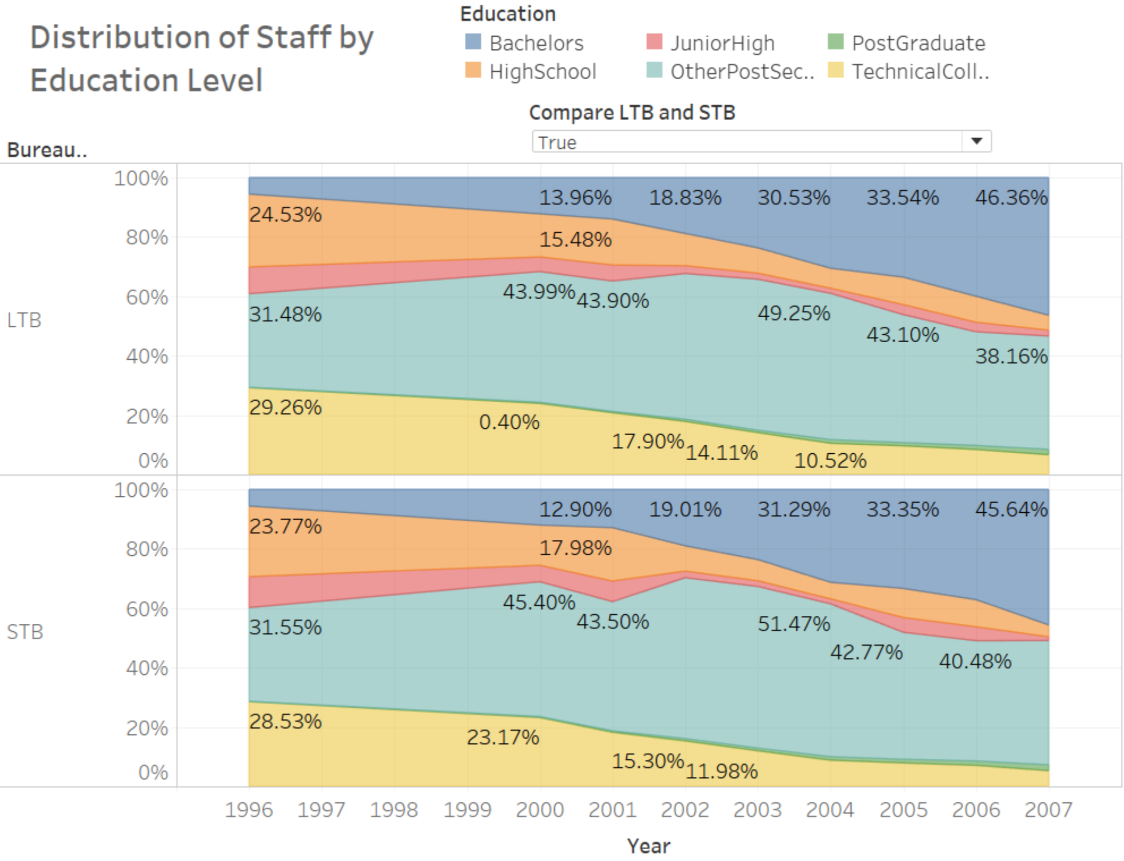

Figure 5: TAX Education distribution.