STAT550 Final Report

Government Agencies

Wenzheng Zhou, Qiaoyue Tang, Shanshan Pi

Department of Statistics

University of British Columbia

Apr.15, 2018

Abstract

Tax Administration Services (TAS) is one of the largest branches of China's civil government agencies. The main objectives of this study are to compare changes in staffing number, identify similar trends, and visualize them on a map of China in a ShinyApp. This report presents the Non-negative Matrix Factorization (NMF) method for identifying patterns on the changes of staffing number in Chinese provinces over 8 years (2000-2007), gives instructions on how to use the ShinyApp to find provinces with similar trends, and interprets the map and plots.

1 Introduction

This project is based on the data collected from one of the largest Chinese government agencies: Tax Administration Services (TAS). The data potentially reflect large changes in policies, bureaucratic organization and personnel thus showing certain patterns and trends on the changes of staffing number in the provinces.

This project aims to identify provinces with similar patterns using algorithms and framework from Nonnegative Matrix Factorization (NMF) method and visualize them in a ShinyApp. The provinces are coloured by clusters in a map of China and the staffing numbers over selected years are plotted against time.

This report is structured as follows: Section 2 illustrates how to use the ShinyApp and explains how to use filters in the app. Section 3 gives the explanations of the results shown in the display panel. Further discussions and suggestions can be found in Section 4.

2 Instructions

The name of our shiny app is STAT450/550 Project: Chinese Government Agencies. There are two parts in the interface: operational part on the left side and display part on the right.

2.1 Operational part

In the operational part, you can apply the following four filters in your search:

- *Year*: A slider filter range from the year 2000 to 2007. Use it to select the range of years for the clustering.
- *What to cluster?*: Two radio buttons, one is "Total number of staff", another one is "Staff Ratio". Choose one of them as the filter criterion by clicking the corresponding button.
- *STB or LTB?*: Two radio buttons, one is "STB", another one is "LTB". Choose one of them as the filter criterion by clicking the corresponding button.

- *Number of Clusters?*: A select list of "2", "3" and "4". Choose the desired number for the clustering from the list by clicking on the numbers.

After selecting all the filters, click the *search* button to see the results of the clustering. The results will be displayed on the right side of the web browser.

2.2 Display part

In the display part, there are three buttons. Click each one of them to see the corresponding panel:

- *Map of China*: In this panel, the clustering results are shown on the map of China. Provinces are shown in different colors which correspond to different clusters. For each cluster, the number of provinces and the province names are shown in the text box below.
- *Plots for Different Clusters*: This panel contains trend plots for all the provinces after applying the selected filters and is partitioned into different clusters.
- *ReadMe*: A simple introduction to the shiny app.

3 Interpretations

3.1 Clustering Method

We perform a cluster analysis on both the total staff number and the normalized staff ratio (staff number/population). The clustering method is based on the result of Non-negative Matrix Factorization (NMF). NMF is a technique of decomposing a large matrix into the product of two smaller matrices and can be considered as a clustering method. The two smaller matrices contain information on the cluster patterns and weights respectively. For example, we assign an object to cluster 1 if it has the largest weight on cluster pattern 1 comparing to the weights on the other cluster patterns. The reason to provide the choice of clustering by the normalized staff ratio is to remove the absolute size effect. For example, some provinces may have a larger population thus having a larger staff number. The size differences between provinces become a dominating factor in the clustering method. The normalized staff ratio is between 0 to 1 for all provinces thus can capture other useful features if interested. NMF has the property that both the input and the results only contain non-negative elements. It is advantageous for our dataset because both the total staffing number and the normalized staff ratio contain non-negative numbers only.

3.2 Map of China Panel

In the *Map of China* panel, we may visualize the clustering result on the map and read the number of provinces and the province names in each cluster. As shown in

Figure 1, different clusters are marked by different colours, and provinces marked by the same colour belong to the same cluster. Sometimes the map shows a cluster of NA and the corresponding region is shaded gray in the map. It is because the original dataset does not contain information on certain provinces and we mark them as "not available". The table below the map shows the names of the provinces in each cluster. As we change the filters on the left-side panel, the clustering map and the table change correspondingly. For example, as shown in *Figure 2*, when we change the number of clusters to 3, the map shows three different colours representing three clusters. The table below change accordingly to show province names in these three clusters.

3.3 Trend Plot Panel

The *Plots for Different Clusters* panel contains trend plots for different provinces after applying the filters and is partitioned into different clusters. The clusters and provinces in each cluster are the same as shown in the *Map of China* panel. For example, in *Figure 1* we demonstrate the clustering result with the choices of *2 clusters, LTB and Total Staffing Number*. We see that *Xinjiang* and *Shanghai* belong to cluster 1 and all the other provinces belong to cluster 2. In *Figure 3*, the trend plots for *Xinjiang* and *Shanghai* are plotted together on the left-side under cluster 1, and the trend plots for all the other provinces are plotted on the right-side under cluster 2. The trend plots are designed to visualize how different clusters behave differently. From *Figure 3*, we observe that the two provinces in cluster 1 both have a flat line between 2000 to 2005 and a peak in 2006. The other provinces in cluster 2 seem to have a common peak in 2001 and a period of decrease between 2004 to 2006. We may note that some of the provinces plotted in cluster 2 (the few lines in the bottom) seem to be flat during the whole period. It indicates that 3 clusters could be more appropriate, depending on the investigator's preference over the pre-specified number of clusters to look at. As another example, *Figure 4* shows the clustering result with the choices of *3 clusters, LTB and Staff Ratio*. We observe that provinces in cluster 1 seem to have a common peak between 2000 to 2002; provinces in cluster 2 seem to have a peak in 2006; provinces in cluster 3 seem to have a decrease between 2004 to 2006.

4 Discussions and Recommendations

The NMF method is a quite stable clustering method, but the results can still change from one attempt to another. We suggest looking at the trend plots for different clusters and see if the clustering is able to capture unique characteristics within a cluster. If the results are not satisfying, try another attempt would help. It is also important to note that NMF is not a classical clustering method and may not be able to give the exact number of clusters specified. If the actual number of clusters is less than the pre-specified number of clusters, it means the data are not significantly different enough to be separated into more clusters.

So far we know how the filtered provinces can be separated into different clusters, but we are not sure what makes them being clustered into the same group. A further analysis can focus on investigating how the trend of staffing number/ratio in one group is different from another group. In addition, if time allows, missing values should be imputed and included in the dataset since we may lose information if ignoring them.

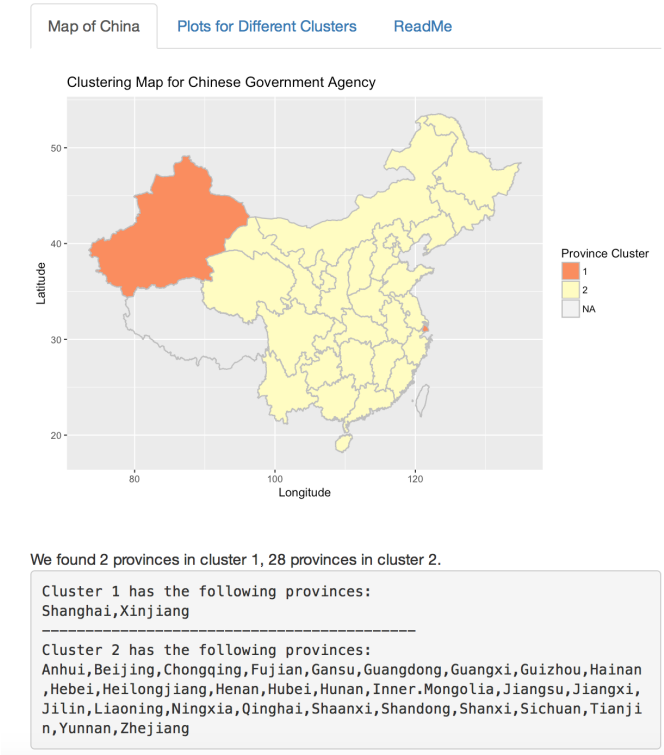

Figure 1: Map of China panel with filters 2 clusters, LTB and Total Staffing Number.

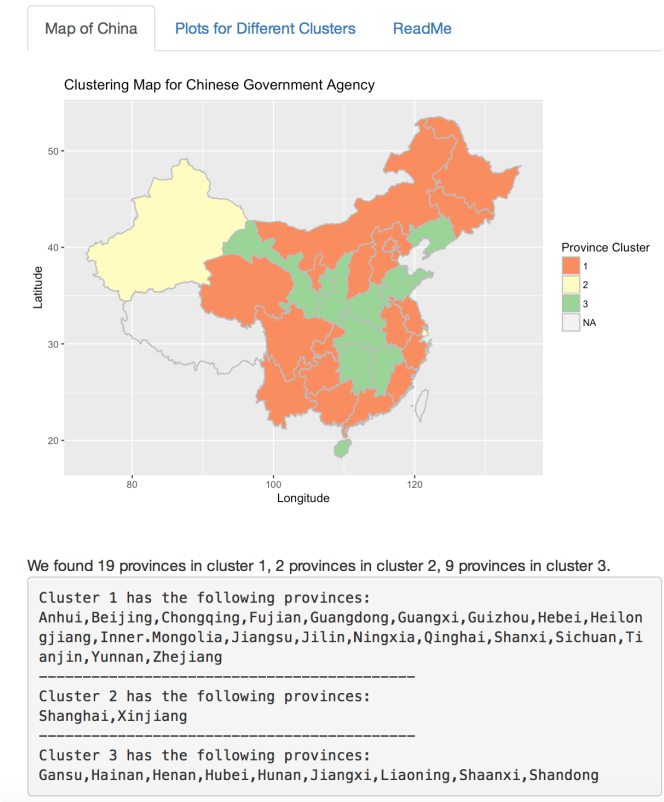

Figure 2: Map of China panel with filters 3 clusters, LTB and Staff Ratio.

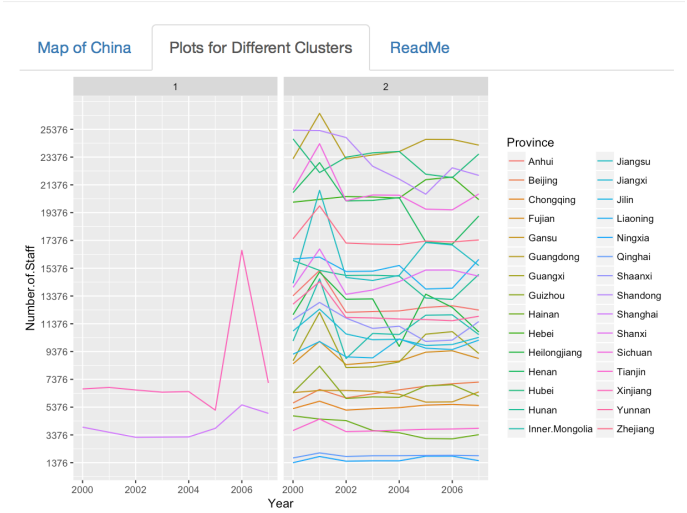

Figure 3: Plots for Different Clusters panel with filters 2 clusters, LTB and Total Staffing Number.

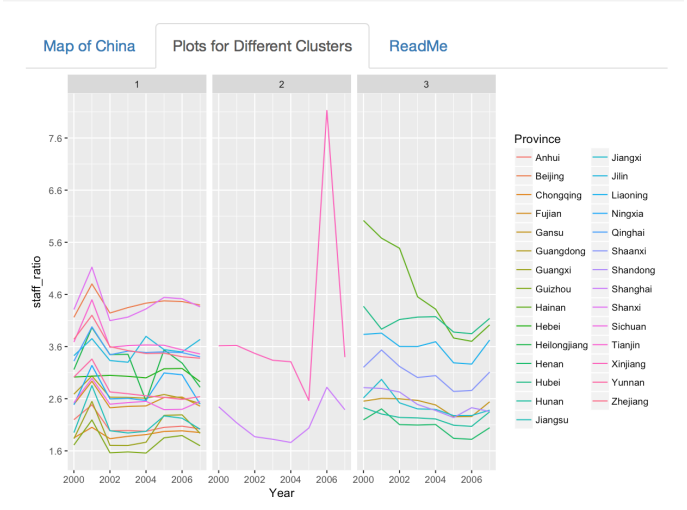

Figure 4: Plots for Different Clusters panel with filters 3 clusters, LTB and Staff Ratio.