Abdullah Farouk

GOVERNMENT AGENCIES

April 10, 2018

Student ID: 71244370
University of British Columbia
Department of Statistics

SUMMARY

In this report, we propose three methods to understand changes in employee composition (i.e. employee age and type of work done) of China's Tax and Environmental Protection agencies. We recommend the use of a Polar Area diagram to identify interesting seasonal employment trends, an interactive heat map to elegantly display changes in employment patterns through time and across provinces and a Poisson regression model to study correlations between hiring practices of the two agencies. We conclude with a discussion of the limitations of the data provided (e.g. missing values) and how they can be handled.

INTRODUCTION

The Government of China is responsible for the well being of its people. This includes protecting China's air, water and land from pollution. These tasks are overseen by the Ministry of Environmental Protection of the People's Republic of China (EPA). The Chinese government relies on taxes collected by the State Administration of Taxation (SAT) to fund the EPA so it can meet its mandates. This requires an abundance of manpower at the SAT and the EPA, whose employment patterns are the core focus of this report. In particular we outline methods to:

- Identify key trends in employment patterns within the EPA and the SAT.
- Visualize changes in employment within these two organizations over time.
- Asses if employment patterns of one organization affects those of the other.

In the upcoming sections we proceed as follows. First, we describe variables in the datasets under study. We then outline different visualization techniques appropriate for these datasets. Finally we discuss the limitations of these datasets and provide solutions to overcome them.

PROPOSED METHODS OF ANALYSIS

In this section we proceed as follows:

- We provide a brief description of the variables in the datasets provided.
- We then propose models that address the points raised in the introduction.

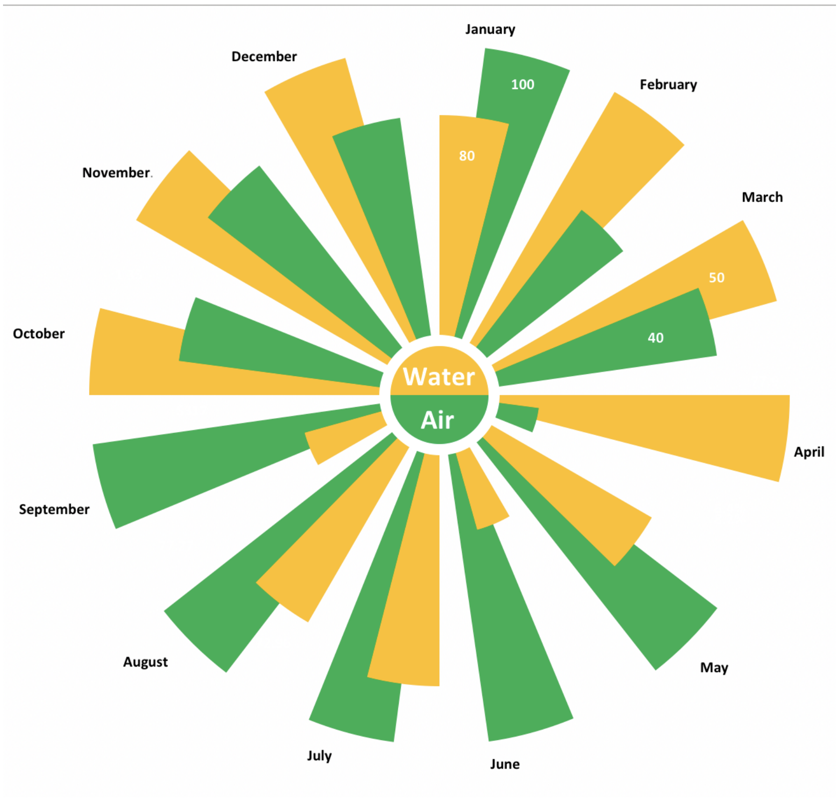
Data Description

There are two datasets under study in this report. The first dataset contains information on the age, positions held and education levels of SAT employees from 1996 to 2013. The second dataset holds records of the number of people employed by the EPA to do different tasks at the national, provincial and federal level from 1992 to 1996.

Now we present different methods we feel appropriately answer the questions raised above. We begin with a description of each method followed by a simple example of their use.

Polar Area Diagram (PAD)

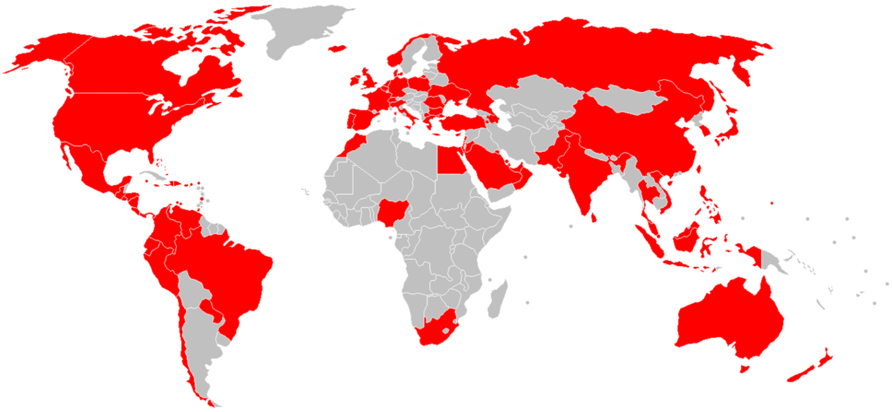
PADs are great at revealing cyclical patterns in datasets under study. This makes them very useful for displaying trends in the EPA's employee count across its different teams, over time. The diagram below is an example of how a PAD can be used on the EPA employment data. It is drawn using simulated values of the total number of people hired by the EPA, across its Water and Air protection teams, throughout the twelve months of 2014.



The diagram above makes it very easy to spot seasonal relationships in the data. For instance, we see that the number of people hired into the Air Protection team is highest during the summer months, whilst the number of people hired into the water Protection team is highest during the fall and winter seasons.

Interactive Heat Map of Chinese Provinces

An interactive heat map is another elegant way of highlighting changes in the number of people employed by the EPA and SAT through time and space. The diagram below is an example of what a heat map looks like. We can use a similar heat map of China to picture changes in the SAT employment structure, across its different provinces over time.



For the purposes of this study it is best to think of this diagram in a dynamic manner. For instance the intensity of the color could be used to represent how many people are employed by the SAT in a particular province (a lighter shade represents fewer people employed). We can then vary the intensity from one year to the next. This allows us to see which provinces have experienced a large increase or decrease in the number of people employed by the SAT over time.

The methods discussed above illustrate how employment trends of the EPA and the SAT can be visualized. Now we move on to discussion of a method that allows us to statistically check for some form of inter-dependency between the hiring practices of these two organizations.

Poisson Regression Models (PRM)

PRMs are a class of statistical models used to study variables that are a count of something. They allow us to explain the variation observed in these variables using models that resemble linear regression models. This makes them an ideal candidate for investigating inter dependencies in employment trends (essentially count data) of the EPA and the SAT. Here is a simple example of how this can be done. Suppose we want to know whether the number of 25 - 35 year olds hired by the EPA has an effect on the number of 25 - 35 year olds hired by the SAT in the Jilin Province. We could use a Poisson regression model to answer this question. Our response variable would be a count of 25 - 35 year olds employed by the SAT in Jilin and our explanatory variable would be a count of 25 - 35 year olds hired by the EPA in different provinces. We can then use statistical techniques to determine if the association between these two variables, computed by our PRM, is meaningful.

Limitations of the data

In this section we discuss features of the datasets provided that limits the analyses that can be carried out. In particular we address two main difficulties and potential remedies to them.

Small Overlapping Windows of Time

There is only one year in common between the two datasets. A lack of data over the same time period makes it difficult to study inter dependencies in the employment structures of these organizations. This can be resolved by gathering more data over similar time periods.

Missing Values

Missing values make it difficult to holistically visualize trends over time. They also make it difficult to use PRMs to identify causal relationships of interest. There are two ways to deal with missing values. The easiest solution is to delete data points that are missing values in one or more variables. This, however, will reduce the size of the dataset. Due to this, statistical procedures that test for dependencies in employment trends of the EPA and the SAT will not be able to detect them if they are subtle.

The second solution is to fill in missing values with predictions of what they should be (imputation). Multiple regression models (regression models with many explanatory variables) can calculate the value of a missing data point by exploiting the relationship between a variable that is missing values and other variables in a dataset. Although, a disadvantage of this method is that the use of such relationships make any results obtained representative of only the sample analyzed. This makes it difficult to generalize findings.

Conclusion

In this report we recommend three methods we feel appropriately address the goals of this study. The Polar Area Diagram helps draw attention to cyclical trends, the interactive heat map showcases changes over time and the Poisson regression model quantifies the level of interdependency in employment patterns of the EPA and the SAT. We feel each of these methods present potential to reveal insightful information about the data at hand.