VISTAS - Visualising Industry Skill TAlent Shifts

Cheryl Pay Wei Lin
Singapore Management University
cheryl.pay.2019@mitb.smu.edu.sg

Chong Jia Jun Louis
Singapore Management University
louis.chong.2019@mitb.smu.edu.sg

Lau Wei Han Amos
Singapore Management University
amos.lau.2019@mitb.smu.edu.sg

ABSTRACT

[write abstract for research paper such as data viz methods, evals, user interface design] VISTAS - Visualising Industry Skill TAlent Shifts is a shiny app that aims to provide insights to talent, migration and skill trends in an interactive and user-friendly way... (no more than 300 words)

Abstract

Numerous studies have been conducted to analyse the impact of employment growth and migration on GDP per capita. However, these studies usually use datasets aggregated at country level. More can be done to deep dive into the relationship between GDP per capita, employment growth and migration within each industry to help individuals and countries better understand the rapidly evolving labour market and its impact on country's economic growth. LinkedIn and World Bank Group have partnered to release a useful set of data from 2015 to 2019 on the employment growth and migration for different industries and skills in various countries. Together with GDP per capita data for each country, countries can study the impact of employment growth and migration for each industry and skill on GDP per capita, find out which are the key skills and whether they have been gained or lost to other countries. At the same time, individuals can analyse the labour market, identify the highly sought-after skills and countries that hold better employment opportunities for each skill. This boosts the competitive edge of countries and individuals in the labour market. To aid individuals and countries in the study, we have designed and developed VISTAS (Visualising Industry Skill TAlent Shifts), an interactive and user-friendly visual analytics dashboard.

1. INTRODUCTION

The LinkedIn and World Bank Group have partnered and released data from 2015 to 2019 that focuses on 100+ countries with at least 100,000 LinkedIn members each, distributed across 148 industries and 50,000 skill categories. This dataset aims to help government and researchers understand rapidly evolving labor markets with detailed and dynamic data. It comprises growth rate of employment in each industry in each industry, growth in number of people from each industry and skill in each country and top 10 skills required in each industry. As an extension, we included macroeconomic data (GDP per capita) and population data from the World Bank.

Aside from being able to download the data in csv format from the LinkedIn-World Bank partnership webpage, simple data visualisations are available. However, they are limited in their variety and interactivity because they only allow one parameter of choice. Information is presented in one default way with no other option and the user cannot conduct comparisons "at one glance."

- Talent Migration: The available visualisation only shows top 10 countries on the map, and top 5 countries, industries and skills in the table. A user may want to know more than just the top N for these metrics. The use of a map and table makes it difficult to visualise the inflow vs outflow of the selected country. A visualisation that allows for comparison between inflow and outflow will be useful to understand if a country is gaining or losing talent.

<Figure 1: Insert screenshot showing "Talent Migration">

- Industry Skills Needs: The interactive visualisation in Industry Skills Needs panel allows users to pick an industry group and industry, but only the skills valued in Year 2019 are shown. No option to change the year is provided. One is hence unable to see the change in skills over the years in an industry. By presenting the changes over the years, the user is able to review trends and gather insights to the industry skills shift and sought-after skills in a particular industry.

<Figure 2: Insert screenshot showing "Industry Skills Needs">

In addition, studies on GDP per capita, employment growth and migration are usually done at country level and there are limited studies to analyse the relationship between GDP per capita growth and employment growth or migration in each industry. These analyses are also shown as static charts in reports, lacking interactivity.

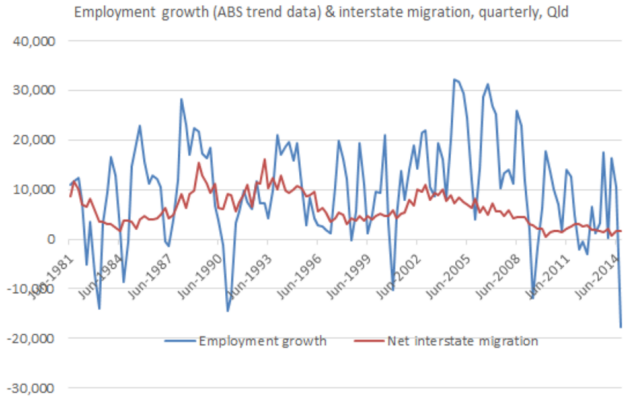

Figure 1: Figure 3: Employment Growth vs Interstate Migration

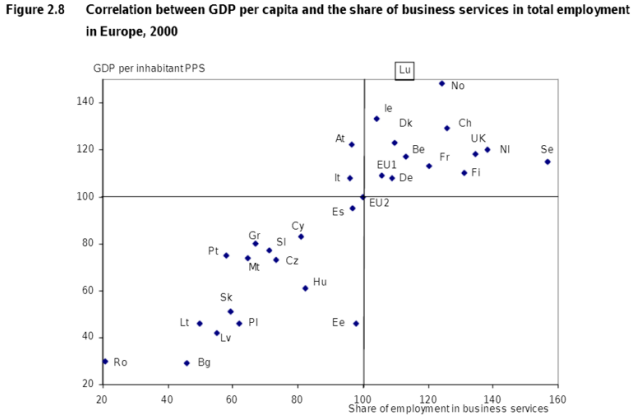

Figure 2: Figure 4: GDP per capita vs Share of employment in business services in Europe in 2000

Figure 3 shows the relationship between employment growth and interstate migration (Tunny, G., 2015). Here, analysis is done at state level, but not industry level.

Figure 4 shows the relationship between GDP per capita and share of employment in business services in 2000 for countries in Europe (Kox, H. & Rubalcaba, L., 2007). However, this does not show how a change in migration or employment growth in an industry will cause the change in GDP per capita.

As such, we have developed a dynamic, interactive and user-friendly visual analytics dashboard as a R Shiny application, titled VISTAS (Visualising Industry Skill TAlent Shifts). It allows individuals and countries to view their competitive advantage and understand the evolving labour market across the world. Comparisons can be done at both country-level, industry-level and skill-level.

This paper reports on the effort to design the VISTAS application and consists of six sections. Section 1 provides a general introduction of the paper. Section 2 provides the motivation and objectives. Section 3 provides a review of analytical techniques for visualising and analysing the data. This includes scatterplots, slope graphs, treemap and more. The interface design and R package selection of VISTAS are discussed in Section 4 and 5 respectively. Section 6 demonstrates and discuss the insights from the use of VISTAS for analysis. Lastly, the paper concludes by highlighting the future direction of research.

2. MOTIVATION AND OBJECTIVES

Through the VISTAS application, we wish to provide individuals and countries with insights into various interest areas to benchmark themselves against the global landscape. We envision that our application will help individuals and countries answer questions on employability, employment opportunities, migration and skill trends.

As mentioned in Section 1, instead of static charts that highlight observations and present them directly, we want to enable users to discover the insights themselves. Hence, our objective is to create a simple tool for users to analyze the data in various ways, including conducting statistical analyses of regression and correlation. The application will support analytical requirements that will be explained in Section 3. It will give users flexibility in selecting their own variables for every visualization. One can choose the values they want to see on their visualizations i.e. x and y variables, categorization (colour by region) and filter using different variables.

3. ANALYTICAL TECHNIQUES AND VISUALISATIONS

The VISTAS application supports various analytical techniques that will produce the visualisations listed below. As mentioned earlier, this will enable users to discover deeper insights on the labour market.

1. Regression Plot: The regression plot and histogram visualise the relationship between two variables (among GDP per capita growth, employment growth, industry migration or skill migration) and their marginal distributions.
2. Scatter Plot: The interactive scatter plot visualises the values for two variables (among GDP per capita growth, employment growth, industry migration or skill migration) and is coloured by year.
3. Correlation Matrix: The correlation matrix visualises the strength of relationship between pairs of variables. Four variables are used i.e. GDP per capita growth, employment growth, industry migration and skill migration.
4. Choropleth Map: The choropleth visualisation is a thematic map where countries are coloured by country, industry or skill migration values.
5. Chord Diagram: The chord diagram is a graphical method of displaying the inter-relationships between data in a matrix. The data are arranged radially around a circle with the relationships between the data points drawn as arcs connecting the data. In this case, the relationship is the net migration between two countries.
6. Slope Graph: The slope graph allows users to compare changes usually over time for a list of categorical variables. The change in country, industry or skill migration for various countries over the years are shown.

7. Treemap: The treemap displays hierarchical data as a set of nested rectangles. Each group is represented by a rectangle, which area is proportional to its value. It shows the net country, industry or skill migration by the colour of the rectangles, and either the population or GDP per Capita by the size of the rectangles. The rectangles can be nested by region or income group of the countries, and clicking on a rectangle will show the individual countries within that rectangle.
8. Geofacet Plot: The geofacet plot takes data representing different geographic entities and apply a visualisation method to the data for each entity, with the resulting set of visualisations being laid out in a grid that mimics the original geographic topology as closely as possible. It shows the net country, industry or skill migration. If it is based on industry or skill migration and multiple industries or skills are selected, bar charts will be shown for each country.

4. USER INTERFACE DESIGN

The VISTAS application is built with several tabs, one per visualisation described in Section 3. Each tab is accompanied by a tooltip (seen when hover over), which describes the purpose of each tab and the visualisation shown in each. Within each tab, there is also a note on the side panel to explain further details of each visualisation.

The introduction tab gives an overview on the VISTAS application and provides a link to the step-by-step user guide. As for the various visualisations, they are divided into two groups - Statistical Analysis and Migration Analysis. The regression plot, scatter plot and correlation matrix are grouped under statistical analysis, as they will produce statistical results when variables are compared with one another. The choropleth map, chord diagram, slope graph, treemap and geofacet plot are grouped under migration analysis.

On the side panel in each tab, users may select different variables, filters, grouping and colours to plot various visualisations. Action buttons are added to delay reactions and minimize lag in the application. The main panel in each tab will then show the visualisation. Proper dimensioning is done to ensure the visualisations are not distorted and the positioning is not compromised when the window width is reduced.

5. APPLICATION

Next, we will explain the R packages used to build each function of the application and element of the visualisations. Comprehensive evaluation was done to decide on the selected R packages for the VISTAS application. However, we also noted some limitations of the packages.

Our application was built around the tidyverse universe of packages and shiny package. tidyverse is an opinionated collection of R packages designed for data science. shiny is an R package that makes it easy to build interactive web apps straight from R.

5.1 System Architecture

The following packages were used for the system architecture and data wrangling.

5.1.1 tidyverse (with readxl)

The tidyverse universe of packages was used extensively throughout the project. It was mainly used for data wrangling operations as it provided a grammar for data manipulation that was easy to comprehend and use.

readxl, also part of the universe but not loaded automatically, was used to read our data files which were in Excel format.

5.1.2 shiny (with shinydashboard, shinydashboardPlus, dashboardthemes, shinyWidgets, shinyjs, shinycssloaders, shinyBS, shinyalert)

The application was built using shiny. Other packages were used to extend the functionality of the base shiny package, as elaborated below:

- shinydashboard allowed us to have a dashboard layout for our application, providing an easy way for users to navigate around our application.
- shinydashboardPlus extended shinydashboard and was used in our application for the flipBox, a box which would flip over when the user hovered over it. This allowed us to show the individual visualisations in our introduction page, while providing further description when a user hovered over the image of the visualisations.
- dashboardthemes was used to theme our application from the default dashboard colour scheme.
- shinyWidgets was used mainly for its pickerInput. The picketInput is a drop-down selection input with more features.
- shinyjs was used for us to call javascript functions in our application. The ability to use javascript allowed more advanced manipulation. For example, it allowed us to disable the zoom function of our choropleth map, which was not easily achievable by the leaflet package.
- shinycssloaders was used to provide loading animations on some of our visualisations that took longer to load.
- shinyBS allowed us to include Bootstrap-like popovers into our application. These popovers provide users with helpful hints on most of the inputs throughout our application when they hover over them.
- shinyalert allows us to create modal dialogs to provide useful information such as information on the visualisations and a help guide to the user.

5.2 Data Visualisation

The following packages were used for our visualisations in the application.

5.2.1 ggplot2 (from tidyverse)

ggplot2 was the main package for plotting of graphics, including scatter plot. It provides a user-friendly way of mapping data variables to aesthetics.

5.2.2 plotly

plotly makes highly interactive, graphically appealing, web-based graphs and can be used together with ggplot2 to show the value of each point on the scatter plot.

5.2.3 ggstatsplot (with ggscatterstats, ggcorrmat)

ggstatsplot was used to create graphics with details from statistical tests included in the information-rich plots themselves. It is an extension of ggplot2 package and preferred over ggplot2 package to create regression plot with statistical results and marginal distributions for its simpler and faster data exploration. As there are very few variables (only 4), it is also preferred over corrplot package to create a simple correlation matrix.

5.2.4 parameters

parameters was used to process the parameters of statistical models e.g. compute p values, confidence intervals and coefficients. Given that some of the statistical results are already shown via the ggstatsplot package, parameters package is preferred over the olsrr package for its simple and streamlined report of statistical results.

5.2.5 leaflet

Leaflet is a javascript library used to create interactive maps. The R package named leaflet makes it easy to use Leaflet in R. While there are many other ways to plot interactive maps in R such as tmap, we chose leaflet as it had more features and provided a way to manipulate the layers without re-rendering the base map.

5.2.6 chorddiag

chorddiag is a package that allowed the generation of chord diagrams. It provides an interface to the d3 library's chord diagram, allowing us to generate d3.js chord diagrams easily.

circlize is another package that allowed the generation of chord diagrams. However, we chose the chorddiag package over this as chorddiag allowed mouseover of the chords to display label values. This functionality was important as our chord diagram may be too cluttered if a user chooses too many countries to display.

5.2.7 CGPfunctions with ggallin

CGPfunctions provides us with the newggslopegraph function. This function provides an easy way for us to plot slope graphs using ggplot2. As an extension of ggplot2, we are able to manipulate the output easily using ggplot2 grammar.

ggallin provides us with the pseudolog10_trans function, a log-like function that works with positive and negative values. This allowed us to apply a log-like transformation for our slope graph if the range of values are too wide.

5.2.8 treemap and d3treeR

treemap allowed us to easily generate treemaps by providing our data and columns to index by.

d3treeR extended the treemap package to allow us to plot treemap objects in a d3.js treemap. Using d3.js treemaps allowed interactivity, where the user can click on the map to go down the hierarchy.

We also considered using plotly for generating of our treemaps. However, extensive data manipulation is required to format the data in a parent-child relationship structure before we can draw treemaps in plotly. The treemap package did not require this manipulation, as it allowed us to specify column names to provide hierarchy to our data e.g. region followed by country.

5.2.9 geofacet

geofacet extends ggplot2 to allow us to create geographically faceted visualisations in R. This allowed us to plot subplots for each country that are arranged in a grid that mimics geographical topology.

As an extension of ggplot2, we are able to manipulate individual plots in a similar function, thus providing a lot of flexibility on the type of plots we would like to show.

5.3 Other packages used outside our application

The following are the packages we used outside of our application e.g. for data wrangling.

5.3.1 rnaturalearth

rnaturalearth was used to generate spatial data for our world map. As these spatial data had to be downloaded, we first downloaded and saved the files into rds format for loading into our application. This allowed our application to load faster, as compared to if we had to download the spatial data every time the application loads.

5.3.2 sf

The simple feature (sf) package was used to manipulate the spatial data we downloaded for our world map. This allowed us to perform sanity checks on the downloaded data and fix it accordingly. For example, the names of certain countries downloaded using rnaturalearth did not correspond to those in our datasets. We had to rename some of these countries to ensure the names matched up.

5.4 Limitations of Certain Packages

5.4.1 d3treeR

We were unable to get d3treeR working with other packages such as chorddiag due the incompatibility of the d3.js versions. d3treeR was using d3.js version 3.x while chorddiag was using d3.js version 4.x. This is an inherent problem with how htmlwidgets uses the highest version of d3.js, which may not be backward compatible. We had to fork this package from d3treeR and manually rename the d3 object, i.e. from d3 to d3_3. While the ideal option would be to upgrade the package to work with a newer d3.js version, this was a quick fix to allow us to use both packages together. The forked package with the renamed d3 object can be found at https://github.com/moomookau/d3treeR.

5.4.2 CGPfunctions

We used the ggnewslopegraph function of CGPfunctions to draw slope graphs in our application. While not really a limitation, we were unable to colour the slope graphs by a grouping variable e.g. Income Group or Region. We had to fork this package from github and make changes to enhance the existing package to achieve this. The forked package with additional functionality can be found at https://github.com/moomookau/CGPfunctions.

6. INSIGHTS FOR CASE STUDY

Using the VISTAS application, we will have a case study on countries in East Asia and Pacific, in particular Singapore. The focus will be on the Information Technology and Services industry and Data Science skill, as it is of interest to SMU MITB students.

What has the audience learned from your work? What new insights or practices has your system enabled? A full blown user study is not expected, but informal observations of use that help evaluate your system are encouraged.

The relationship between GDP per capita growth vs industry employment growth, industry migration or skill migration is not strong. This is shown in the low R-squared values (below 0.02) in Section 3.3. Even when multiple variables are used to build a multiple regression model with GDP per capita growth and the less important variables (by p-value) as well as the correlated variables are removed in Section 3.5-3.6, the R-squared values are still low. Furthermore, in Section 3.4, we can see the weak correlation between GDP per capita growth and the other variables. Hence, even though we observe a slight negative relationship between GDP per capita growth and the other variables in Section 3.3 i.e. lower GDP per capita growth with higher employment growth and migration, we are unable to provide a firm conclusion on how GDP per capita growth is impacted by migration and employment growth in a country. Further study can be done to understand how GDP per capita growth is influenced by a combination of many other variables or how different filters used (e.g. year, industry) can change the results. The relationship between industry employment growth vs industry or skill migration is stronger, as seen in the R-squared values in Section 3.3 i.e. 0.429 (industry employment growth vs industry migration) and 0.248 (industry employment growth vs skill migration). There is also moderate correlation between industry employment growth vs industry or skill migration, as seen in Section 3.4 i.e. ~0.6 (industry employment growth vs industry migration) and ~0.5 (industry employment growth vs skill migration). Even though the R-squared and correlation results are still not ideal, the results are able to give a glimpse of how much employment growth changes with industry or skill migration. Ideally, the employment growth to industry or skill migration ratio should be positive (i.e. higher employment growth with higher migration, as seen in Section 3.3) and high value to give job seekers confidence that they are more likely to find a job in certain industry and country if they have certain skill. Further study can be done to compare the ratio for different industries and skills. This will give job seekers knowledge of which industries and skills to focus on.

7. CONCLUSION

Future work - ....

We have identified two main limitations of our project which arise due to the LinkedIn dataset: Our data does not capture all the countries due to the penetration rate of LinkedIn; some countries do not use LinkedIn. Our data is not representative of the entire migration landscape as LinkedIn data has better coverage fo white-collar workers in knowledge-intensive sectors.

8. ACKNOWLEDGEMENT

The authors wish to thank Professor Kam Tin Seong for his guidance throughout the tenure of the project.

9. REFERENCES

1. Duca, D. (2019, July 9). Using LinkedIn for social research. Retrieved February 28, 2021, from https://blogs.lse.ac.uk/impactofsocialsciences/2019/07/09/using-linkedi

2. World talent migration trends. (2019, August 22). Retrieved February 28, 2021, from https://www.bayut.com/mybayut/world-talent-migration-trends/

3. People on the move: Global migration's impact and opportunity (2016, December). Retrieved February 28, 2021, from McKinsey Global Institute website: https://www.mckinsey.com/~/media/McKinsey/Industries/Public%20and%20Social%20Sector/Our%20Insights/Global%20migrations%20impact%20and%20opportunity/MGI-People-o pdf