Data Visualisation and Its Application in Official Statistics

Olivia Or
Census and Statistics Department, Hong Kong, China
ooyor@censtatd.gov.hk

Abstract

Data visualisation has been a growing topic of research among statistical practitioners and academic. Its application for official statistics is apparent as seen in the many research projects in various national statistical offices (NSOs). As digital-age users are becoming more accustomed to graphics with interactivity rather than plain numbers or tables, there are opportunities for complex official data series to be presented in visual and dynamic formats to facilitate understanding and interpretation. Visualisation of raw data also offers a quick and effective way to identify abnormal patterns, such as outliers or missing values, during internal processing. In this paper, various data visualisation tools relating to official statistics production and dissemination will be discussed and illustrated with examples from NSOs.

Key Words: Data dissemination, graphical presentation, exploratory data analysis

1. Introduction

Data visualisation is an important field in statistics. As noted by Statistician Edward Tufte, “often the most effective way to describe, explore, and summarise a set of numbers – even a large set – is to look at pictures of those numbers” (Tufte, 2001). Presentation of data in graphical formats assists cognition, so that the data displayed can be more easily understood. In addition, visualisation can also be used to communicate stories, ideas and insights surrounding the data, thereby empowering the readers to better interpret and make use of the data. With the proliferation of computing technology, large datasets can now be conveniently presented in various graphical formats.

The use of data visualisation in official statistics is obvious. Many National Statistical Offices (NSOs) have been researching, designing and implementing effective visual formats to present their data and statistics. For instance, the Office of National Statistics (ONS) of the United Kingdom (UK) has set up a Data Visualisation Centre, whereas Statistics New Zealand has listed visualisation developments as one of their business priorities.

This paper presents the different applications of data visualisation in official statistics, as seen from the examples in various NSOs.

2. Visualisation for Quality Assurance

Before processing, raw data collected need to be verified and scrutinised to ensure quality. It is often of interest to understand the data distribution and characteristics before further processing. Traditionally, validation rules, inspection of summary data or graphs such as box plots and scatter plots are used to check for irregularity or abnormal patterns, such as outliers and missing values. However, it would be harder if the datasets are multivariate, large and complex, as in the case of survey data collected and processed by NSOs.

Nowadays, modern technologies enable large multivariate datasets to be explored using new visualisation techniques, so as to identify the most suitable editing and imputation schemes to be applied. In this section, several visualisation tools adopted by NSOs to facilitate data exploration and quality assurance will be presented.

i. Multivariate Data

Invented by Computer Science Professor Ben Shneiderman of University of Maryland in 1992, "treemap" is a graphical tool for visualising distributions of large and complex hierarchical data. It provides a "top-down" view of aggregated data with hierarchical structure, and displays them as a set of recursively divided rectangles to represent the hierarchy, with areas and colours showing two other attributes. Patterns can be easily revealed in treemaps for further study and analysis.

Statistics Netherlands has proposed the application of "treemap" for checking quality of hierarchical data during statistics compilation process (Tennekes, de Jonge, & Daas, 2012). One of their applications is for detecting unexpected changes of total value added of different economic sectors as compared to last year (Figure 1). Treemaps can also be constructed with zooming and interactive capabilities to facilitate drilling in.

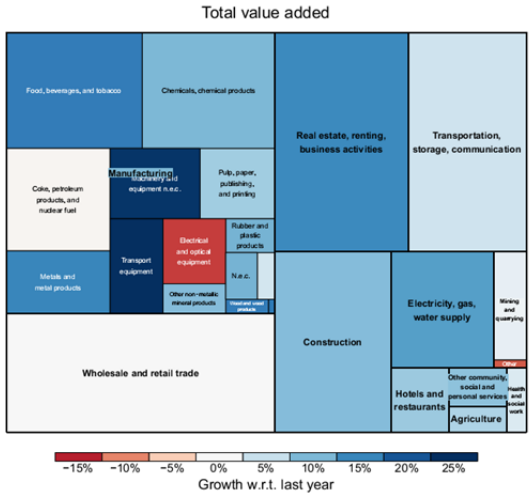

Figure 1 Treemap of Dutch Total Value Added by Economic Sector. Sizes of rectangles represent the total value added, while colour shows the growth with respect to last year.

In contrast to treemap, a tableplot is a bottom-up tool that summarises a large multivariate dataset and presents the distributions of variables in bar charts (for numeric variables) and stacked bar charts (for categorical variables), sorted by a selected variable (Malik, Unwin, & Gribov, 2010). Records are divided into "row bins" according to the order of the sorting variable, and values of each "row bin" are plotted. Abnormal data distributions, relationships between variables and patterns of missing values can be spotted easily. Statistics Netherland implemented static tableplot on the raw data of structural business statistics, as illustrated in Figure 2. Irregularity can be spotted right away. Similar to treemaps, interactive capabilities can also be added to tableplots to facilitate data exploration.

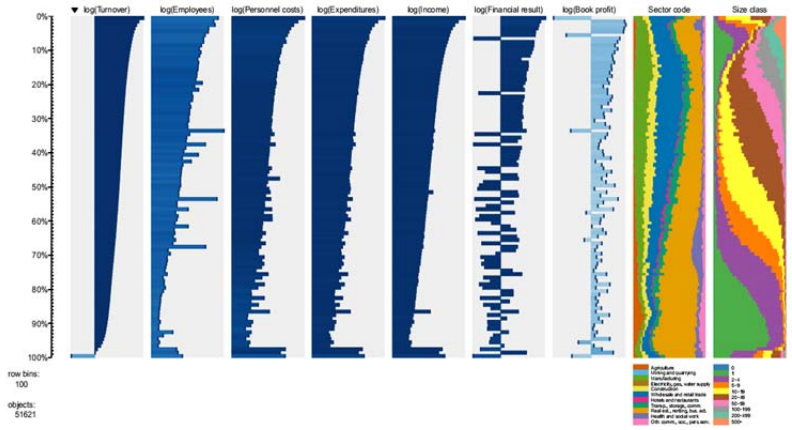

Figure 2 Tableplot of Dutch Structural Business Statistics (Tennekes, de Jonge, & Daas, 2011)

ii. Missing Values

The United Nations Industrial Development Organisation (UNIDO) has explored the use of visualisations in understanding missing value distributions. One of the graphical tools they used is "aggregation plot" (UNIDO, 2012). Figure 3 shows an example, in which missing values across combinations of variables are shown in red. Further checking can be conducted to understand why certain combinations of variables result in a large number of missing values. Also, spine plots can be used to study the proportion of missing values in variable A and how it distributes with another variable B (Templ & Filzmoser, 2008). An example by UNIDO is shown in Figure 4 (UNIDO, 2012).

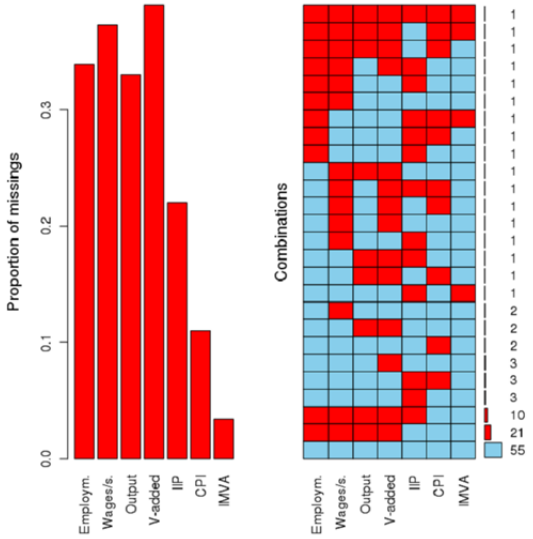

Figure 3 Aggregation Plot of UNIDO Industrial Statistics Database

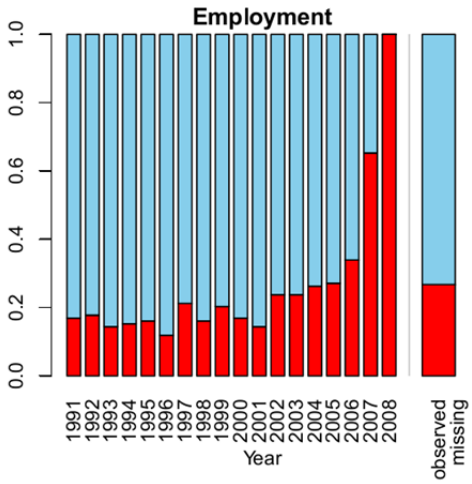

Figure 4 Spine Plot of UNIDO Industrial Statistics Database (color coding of missing and available data in variable "employment" by year)

3. Visualisation for Public Understanding

As noted by Bill McLennan, former head of the NSOs in Australia and UK , "to be useful official statistics need to be used" (Statistics New Zealand, 2012). Indeed, all NSOs subscribe to the same mission to provide statistics for use by the community, whether experts or laymen, in an accessible and comprehensible manner.

Visualisation is an effective way to present official statistics in an impactful way that empowers readers to see trends, patterns and insights behind the data, and encourage them to interpret and think deeper. Many NSOs have explored the use of sophisticated visualisation tools to "present statistical information in easy-to-understand forms, and provide deeper context to the stories" (Statistics New Zealand, 2012), so as to generate more public interest and more uses of official statistics.

i. Extending Audience Base

For the general public, statistics may sound like a nightmare full of numbers and formulae that make no sense to them. To raise their interests in data and statistics, the best way is to engage them through stimulating and "user-friendly" data stories. Information graphics (infographics in short) can be an effective way to concisely display the key points and messages behind large amount of complex data and analysis to the general public, especially laymen who do not have the skills or time to understand large and complex datasets.

Infographics are best used by NSOs as publicity materials to reach out to a large group of audiences. As a pioneer, Statistics New Zealand has developed a number of infographics to supplement their published articles. One of them, which summarises

the prices of haircuts over years, is shown in Figure 5. Through this infographic, readers can better appreciate how haircut prices had changed with inflation and service scopes over time. On top of graphics, textual explanation is provided, and an in-depth analysis is also available in an accompanying article for those readers who would like to read more details on this topic.

Seeing this as a key trend, the ONS of UK has also initiated to present infographics and headline stories on the ONS website as one of their publication strategy. They published an infographic on labour market overview, including comparison of full time and part time workers and earnings by industry (Figure 6). Even the general public with little knowledge on the topic can quickly grasp the latest position of the labour market through the key data points and explanations provided.

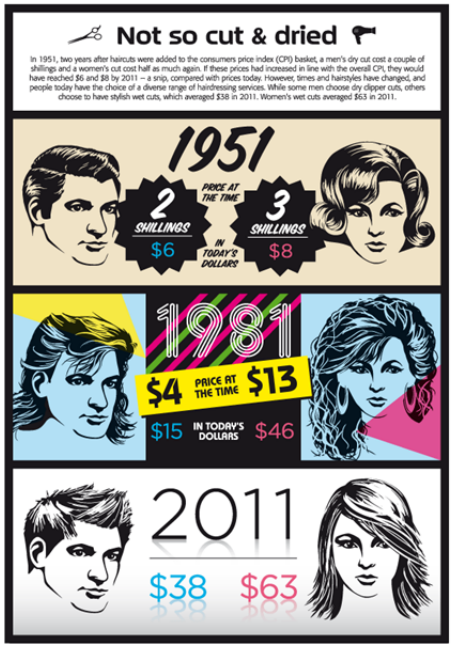

Figure 5 Infographic on Prices of Hair Cuts by Statistics New Zealand

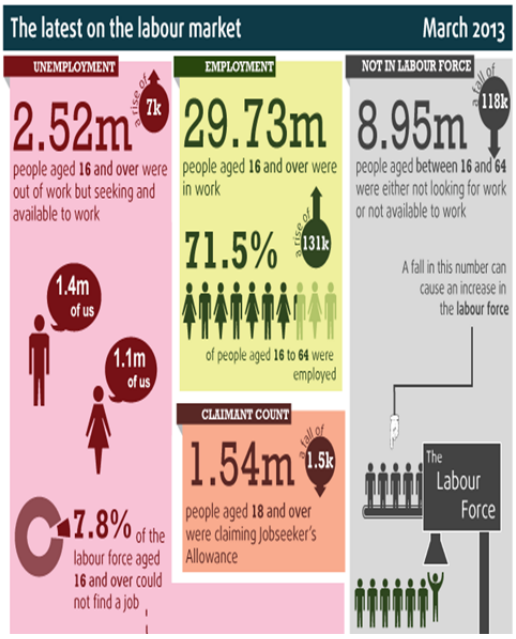

Figure 6 Extract of an infographic on Labour Market produced by the Office of National Statistics of the United Kingdom

ii. Exploring Trends and Relationships

In addition to grasping key concepts, sometimes readers would like to study deeper on trends, relationships and comparisons among certain statistical data series. With the advancement in computing power and web technologies, nowadays official statistics can be presented in animated or interactive formats to facilitate understanding.

One of the most common applications of animated or interactive visualisations is to show changes of population structures over time. For instance, the Census and Statistics Department of Hong Kong Special Administrative Region presented the statistics of its 2011 Population Census in animated and interactive format (Figure 7). Changes of population structure, composition and growth over time are either animated or equipped with interactive functions, allowing users to visualise the changes with a few clicks. Such statistical graphs can be appreciated a lot more easily than static charts or tables, allowing users to instantly see the trends.

The Australia Bureau of Statistics (ABS) also published their 2011 Census data with an interactive tool called "SBS Census Explorer" (Figure 8), allowing users to explore and compare demographic data by different attributes such as language spoken,

country of birth, household income, etc.

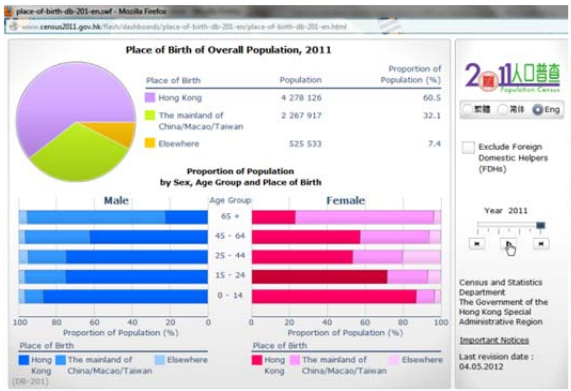

Figure 7 Interactive Visualisations of Hong Kong 2011 Census Data

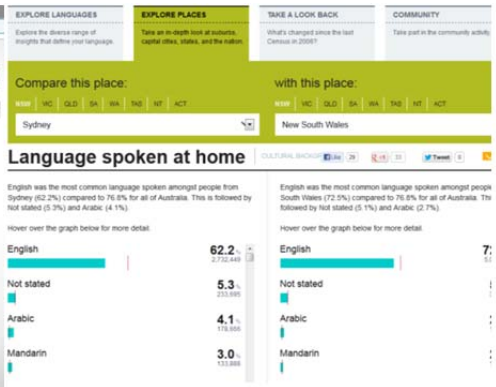

Figure 8 SBS Census Explorer of Australia 2011 Census Data

iii. Analysing Geospatial Data

Many official statistics series contain geographic components, such as population characteristics by residential area. Presenting these geospatial data series in graphs facilitates patterns and insights about the data across different regions to be easily observed. With the emergence of free and user-friendly Geographic Information Systems (GIS) software such as Google Maps, and with more and more users adapting to scrolling through digital maps to look at information, geospatial visualisation is gaining popularity in presenting geospatial official data.

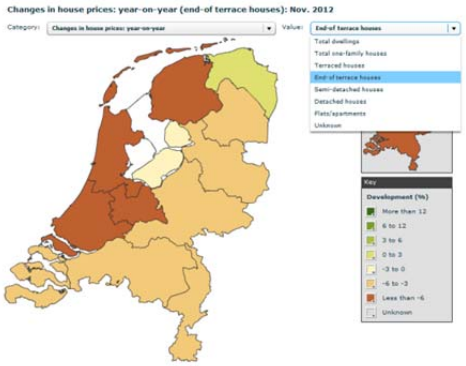

Figure 9 Dutch Housing Market on the Map by Statistics Netherlands

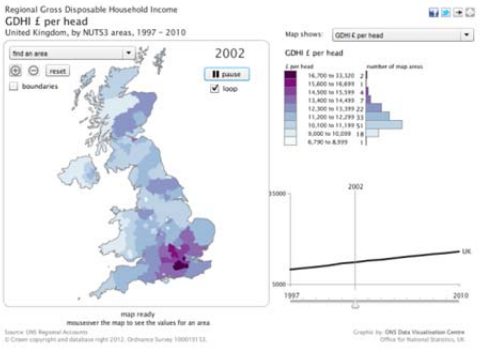

Figure 10 Animated Map of Regional Gross Disposable Household Income by ONS of UK

Statistics Netherlands published a series of interactive maps to illustrate economic activities and housing situations. Figure 9 shows a choropleth map highlighting the changes in selling prices/transactions of existing houses in different parts of the Netherlands as compared with the same month in previous year, using different colour shades. Another common geospatial graph is cartogram, in which the size of different areas on the map is proportional to a certain statistical measurement.

Geospatial visualisations can be developed with animation as well. The ONS of UK has developed an animated map to show the regional gross disposable household income (Figure 10), enabling users to see the changes over time by region.

4. Conclusion

Data visualisation is a useful tool to facilitate understanding and usage of data. The examples presented in this paper show that NSOs are placing more focus on using graphical ways in processing and disseminating official statistics. Internally, data visaulisation facilitates understanding of raw data collected, and enables identification

of abnormal patterns and missing values to ensure data quality. In data dissemination, visualisation can effectively draw readers' attention, stir their interest, communicate the key observations and stimulate readers' thinking, as graphics are a lot more appealing and comprehensible than plain numbers, texts or tables. In this sense, visualisation is a useful tool to promote statistical literacy. On the other hand, to ensure that visualisation is properly used, NSOs can also educate the media on how to report and present official statistics in effective and unbiased visualisation.

The goal of data visualisation is to present the data and stories surrounding the data with clarity and integrity, without any distortion or misrepresentation. By making use of visualisation tools to attract, engage and empower users, official statistics can become something that all users, even laymen, can understand, think through and make good use of.

References

Census and Statistics Department, Hong Kong. (n.d.). *Interactive Visualizations*. Retrieved from Census Results, 2011 Hong Kong Population Census: http://www.census2011.gov.hk/en/interactive-visualisations.html

De Róiste, M., Gahegan, M., Morrison, P. S., Ralphs, M., & Bucknall, P. (2009). *Geovisualisation and policy: exploring the links.*

Forbes, S., Goodyear, R., & Ralphs, M. (2011). *Visualising Official Statistics, Statistics New Zealand Working Paper No 11-02.*

Malik, W. A., Unwin, A., & Gribov, A. (2010). An Interactive Graphical System for Visualizing Data Quality–Tableplot Graphics. In *Classification as a Tool for Research.* Springer.

Office of National Statistics, United Kingdom. (n.d.). *Interactive Content*. Retrieved from Office of National Statistics: http://www.ons.gov.uk/ons/interactive/index.html

Office of National Statistics, United Kingdom. (n.d.). *Labour Market Infographic Summary, March 2013*. Retrieved from http://www.ons.gov.uk/ons/infographics/labour-market-summary-mar-2013/index.html

SBS. *SBS Census Explorer*. Retrieved from Census 2011 Australia: http://www.sbs.com.au/news/census/

Shneiderman, B. (n.d.). *Treemaps for space-constrained visualization of hierarchies.* Retrieved 12 2012, from http://www.cs.umd.edu/hcil/treemap-history/index.shtml

Spence, R. (2001). *Information Visualization.* Pearson.

Statistics Netherlands. (n.d.). *Interactive Infographics*. Retrieved from CBS: http://www.cbs.nl/en-GB/menu/publicaties/webpublicaties/interactief/default.htm

Statistics New Zealand. (2012). *Annual Report of Statistics New Zealand for the year ended 30 June 2012.*

Statistics New Zealand. (n.d.). *Not so cut and dried: 60 years of tracking haircut prices in the CPI*. Retrieved from Statistics New Zealand: http://www.stats.govt.nz/browse_for_stats/economic_indicators/CPI_inflation/haircuts.aspx

Templ, M., & Filzmoser, P. (2008). *Visualization of Missing Values using the R-Package VIM.*

ten Bosch, O., & de Jonge, E. (2008). Visualising official statistics. *Statistical Journal of the IAOS.*

Tennekes, M., de Jonge, E., & Daas, P. (2011). Visual Profiling of Large Statistical Datasets. *New Techniques and Technologies for Statistics.*

Tennekes, M., de Jonge, E., & Daas, P. (2012). Innovative Visual Tools for Data Editing, Statistics Netherlands. *United Nations Economic Commission for Europe, Conference of European Statisticians.* Oslo, Norway.

Tennekes, M., de Jonge, E., & Daas, P. (2013). Visualizing and Inspecting Large Datasets with Tableplots. *Journal of Data Science*.

Tufte, E. R. (2001). *The Visual Display of Quantitative Information.* Graphics Press.

UK Statistics Authority. (2012, 10 26). *Statistics: Communicating and Publishing Statistics*. Retrieved from http://www.publications.parliament.uk/pa/cm201213/cmselect/cmpubadm/writev/commstatistics/m08.htm

United Nations Industrial Development Organization. (2012). *R in the Statistical Office: Part II.*

Yau, N. (2011). *Visualize This, the FlowingData Guide to Design, Visualization, and Statistics.* Wiley.