Data visualisation and official statistics: providing new evidence and enhancing understanding.

Forbes, Sharleen

Victoria University of Wellington
P.O.Box 600
Wellington 6011
New Zealand

Statistics New Zealand
P.O.Box 2922
Wellington 6140
New Zealand

email: sharleen.forbes@vuw.ac.nz

Abstract

Recent data visualization developments provide new tools for displaying official statistics. A range of these will be displayed, from multi-dimensional static graphs to dynamic and interactive graphs, including geo-visualisations. These tools not only enhance the range of outputs of national statistics offices but can also be used for motivational and educational purposes. The use of a price-kaleidoscope to help interpret the Consumer Price Index and of dynamic population pyramids to demonstrate the momentum effect in demography is given. A number of both actual and potential policy uses in New Zealand using geo-visualisation tools with official statistics are also discussed.

Introduction

Visual tools provide one way to students to gain experience of statistical concepts. A simple non-graphical and technology free example is the portrayal of the mean as the 'balancing point' or centre of gravity of a set of data by placing small blocks on a ruler. Changes in the standard deviation (as a measure of the thickness of 'spread') are also easily demonstrated in this way. There is a wide body of literature on the development and use of visual graphics for communicating and analysing statistics (e.g. Tufte, 1983; ten Bosch & de Jonge, 2008; Hidalgo, 2010). Paparistodemou & Meletiou-Mavrotheris (2008) used dynamic statistics software to enhance the learning of statistical inference by young school students while Dominguez-Dominguez & Dominguez -Lopez (2010) contend that the visual approach is '*learning by playing and do-it-yourself'*. Examples that demonstrate the potential use of recent visualizations to develop conceptual understanding when teaching official statistics are given in this paper.

Data visualization software that provides new methods for access to and interpretation of official statistics is now readily available (Statistical Journal of the IAOS, 2008; Forbes et al, in press). These tools allow us to explore some of the policy issues faced by government in new ways. In particular, most official statistics have an associated geography and geo-visualisations can be used to make manageable, display and explore some of the very large and complex data sets that national statistics offices handle. As de Róiste et al (2009) state geo-visualisation is '*conceptualised as a means of visually representing spatial data to better explore and understand patterns and relationships in the underlying information*'. Exciting new open source Geographic Information Systems (GIS) such as Google Maps, Quantum GIS (http://www.qgis.org), etc. have removed many of the former barriers to access to sophisticated technology. National statistics offices have also created their own geo-visualisations (for example, Statistics New Zealand's Commuterview product discussed below). The latest geo-visualisations combine graphs and maps with statistical analysis. One of these, GeoVista (open-source software created at Penn State University, http://www.geovista.psu.edu/grants/cdcesda) is used to demonstrate that, while these tools could be viewed as primarily for specialist researchers or academics, they also provide a way for policy analysts to explore the multivariate characteristics of official statistics data and visually link these characteristics to their underlying geographic patterns. Identification of geographic clusters is important when policy interventions are being designed, and many issues can be explored using these tools. Some policy uses of geo-visualisations in the New Zealand context are given below.

In summary, the paper focuses on two diverse applications of data visualization: enhancing the understanding of official statistics and providing new evidence for policy makers from existing data.

Teaching (official) statistics concepts using data visualization

Example 1: Stepping from simple to multiple regression

Conceptually there is a large step from simple bivariate regression to multivariate models that may or may not incorporate interactive terms (one variable acting as a modifier on another). A visual aid to assist student's understanding is a three-dimensional (3-D) display (in figure 1a a pin-graph where the heads of the pins form a 3-D scatterplot created using the 'R' package (www.r-project/org). These graphs enable students to see and discuss the degree of interaction present and decide whether to only include main effects or to add interaction terms in regression models by looking at the consistency of pattern (for example, across highest education classes in figure 1a). An extension from three- to four-dimensions is achieved by the use of colour in figure 1b and can be used to discuss whether separate models should be fitted when the pattern is different for different groups (in this case, females dominating part-time work and males full-time work). Further increases in dimensionality can be achieved by changing the size, shape or colour intensity of data points or by adding a dynamic feature (such as time) across overlaid static graphs as used by Hans Rosling (2007) in his Gapminder graphs (www.gapminder.org).

.

Figure 1a: Weekly Income by Highest Educational Qualification and Hours Worked

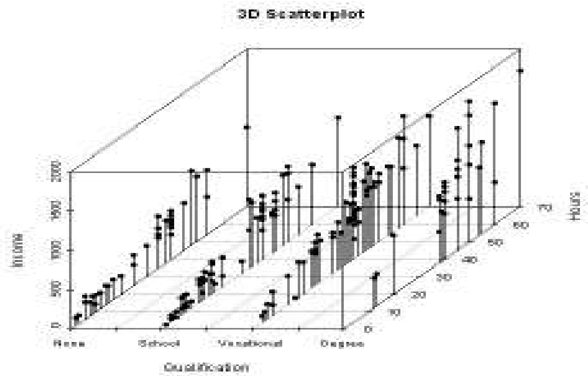
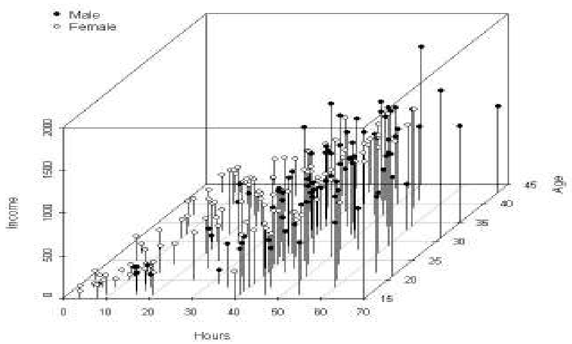


Figure 1b: Weekly Income by Sex, age and Hours Worked



Example 2: Interpreting Price Indices

Consumer's Price Indices (CPIs) measure the rate of price change of representative goods and services (the basket of goods) purchased by households. In New Zealand the CPI is often reported by the media and is used by the Reserve Bank to set monetary policy, by the government to adjust benefit rates and by employers and employees in wage negotiations. It is relatively easy to explain to students the calculation of a single item price index as

$$\frac{current\ price}{reference\ price} \times index\ reference$$

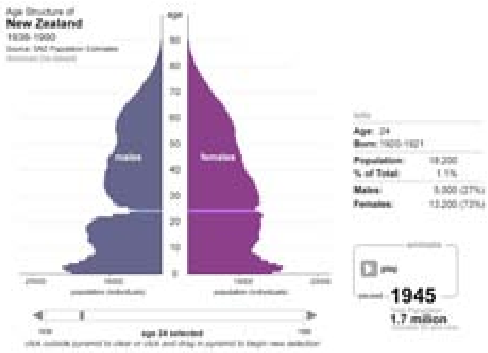
but it is more difficult for the whole CPI, even when the relatively simple Laspeyres formula[1] (Statistics New Zealand, 1999) is used as each good or service is now assigned an (expenditure) weight representing its relative importance in household spending patterns. Groups, subgroups and weights can be displayed visually with the Price Kaleidoscope produced by the Federal Statistical Office of Germany (http://www.destatis.de/Voronoi/PriceKaleidoscope.svg). The CPI is represented by a circle and within this each group (and subgroup) has an area proportional to its weight. Clicking on an area displays its weight and quarterly change allowing students to explore these and see that some subgroups have large positive or negative changes and the effect of each change on the total CPI is determined by its weight.

Example 3: Simplifying demographic concepts

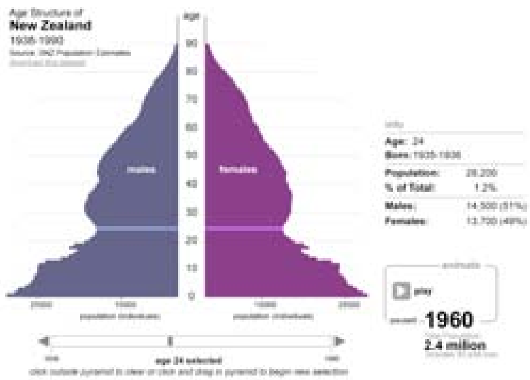
As stated above, changes over time can be viewed by overlaying graphs (or maps) and using an animation tool to 'play' them over time. An example is the dynamic population pyramids now in relatively common usage by national statistics agencies. These clearly indicate (Figures 2a – 2d) changes in the population structure as it 'ages' and can be used to demonstrate demographic effects such as momentum (population growth resulting from a youthful age structure or population decline resulting from an older age structure)

Figure 2: Snapshots of the Statistics New Zealand population pyramids

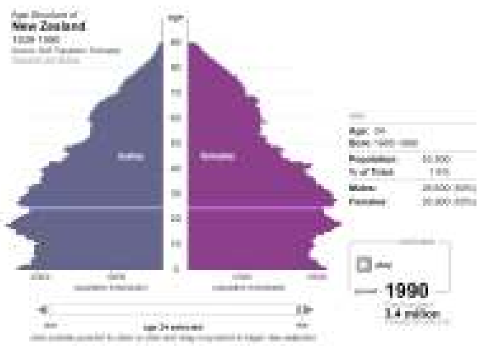
(a) ***At year 1945***



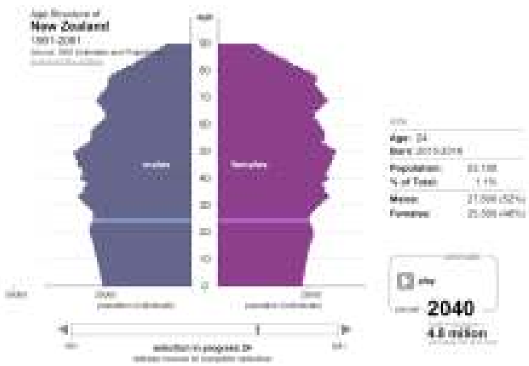
(b) ***Fifteen years later - 1960***



(b) Another thirty years on – 1990



(d) Another fifty years on – 2040?



The pyramids show the momentum process in action, leading up to and beyond the point (different for each country) where natural increase (growth) shifts to become natural decrease (decline). They can also be used to investigate 'what if' scenarios using population projections.

[1] The Laspeyres formula calculates the index for period t on base period o by: $Index = \frac{\sum P_t Q_o}{\sum P_o Q_o} \times 1000$ where Q is the quantity (weight) and P the price of the item.

Some policy uses of geo-visualisation

Example 5: Statistics New Zealand's Commuterview product

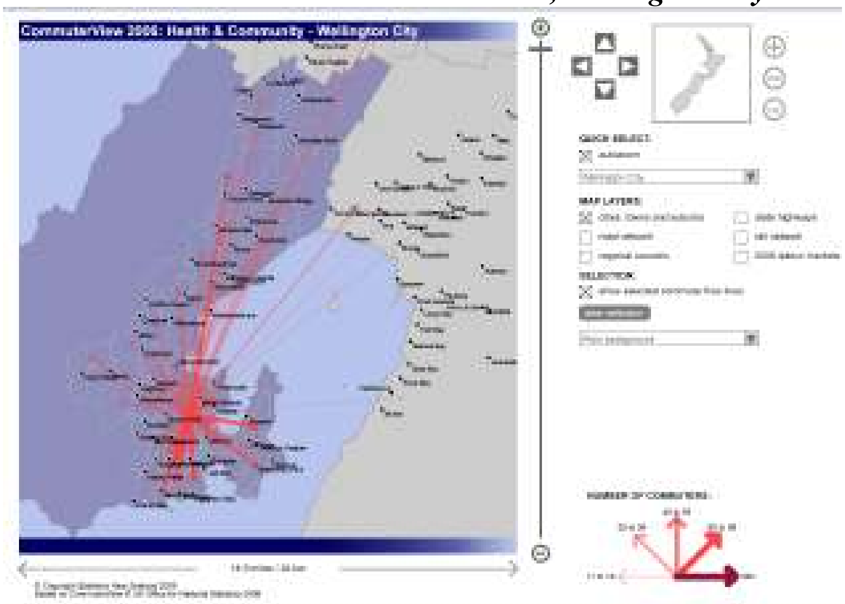
One example of data reduction of a large and complex official statistics dataset is the interactive mapping tool "Commuterview" released on DVD by Statistics New Zealand. It is based on a UK product (Office for National Statistics, 2008) and displays linked information from two 2006 New Zealand Census of Population and Dwellings questions: 'Where do you live?' and 'Where do you work?'. Even at a high level of geography, such as Territorial Authority of which there are 74 in New Zealand, this table would be very large (over 5,000 cells). A small section is given in Figure 3. There are 1,927 Area Units at the next level resulting in a table of more than 3 million cells and, at the lowest level there are 41,392 Mesh-Blocks resulting in a table of almost 2 billion cells, too large for the human mind to assimilate.

Figure 3: Commuting Patterns (Territorial Authority by Territorial Authority) in New Zealand

	Far North District	Whangarei District	Kaipara District	Rodney District	North Shore City	Waitakere City	Auckland City	Manukau City	Total Akld	Papakura District
Far North District	16860	396	36	30	36	12	108	42	201	9
Whangarei District	285	26379	276	75	57	36	171	54	321	12
Kaipara District	42	327	5931	315	33	12	69	27	138	0
Rodney District	30	48	126	21183	6822	1701	5706	627	14856	54
North Shore City	48	63	24	1755	58383	1905	28188	2604	91077	180
Waitakere City	48	48	12	1155	4332	31794	30957	3288	70371	258
Auckland City	201	141	39	738	7257	6183	140517	16023	169983	942
Manukau City	75	69	18	282	1824	1050	40881	66210	109962	3384
Total Akld	372	321	93	3930	71793	40932	240543	88122	441396	4764
Papakura District	9	15	0	33	177	84	3894	5079	9231	6567
Franklin District	12	15	3	48	171	99	3117	3720	7110	1869

But it is at these smaller levels of geography that this information is most useful to city planners so that they can provide public transport services, etc. By using maps and showing the data as 'spider' graphs (Figure 4) rather than tables this data can be viewed for selected ethnic groups, modes of transport, industries and occupations.

Figure 4: New Zealand Commuter Flows: Work to Home, Wellington City

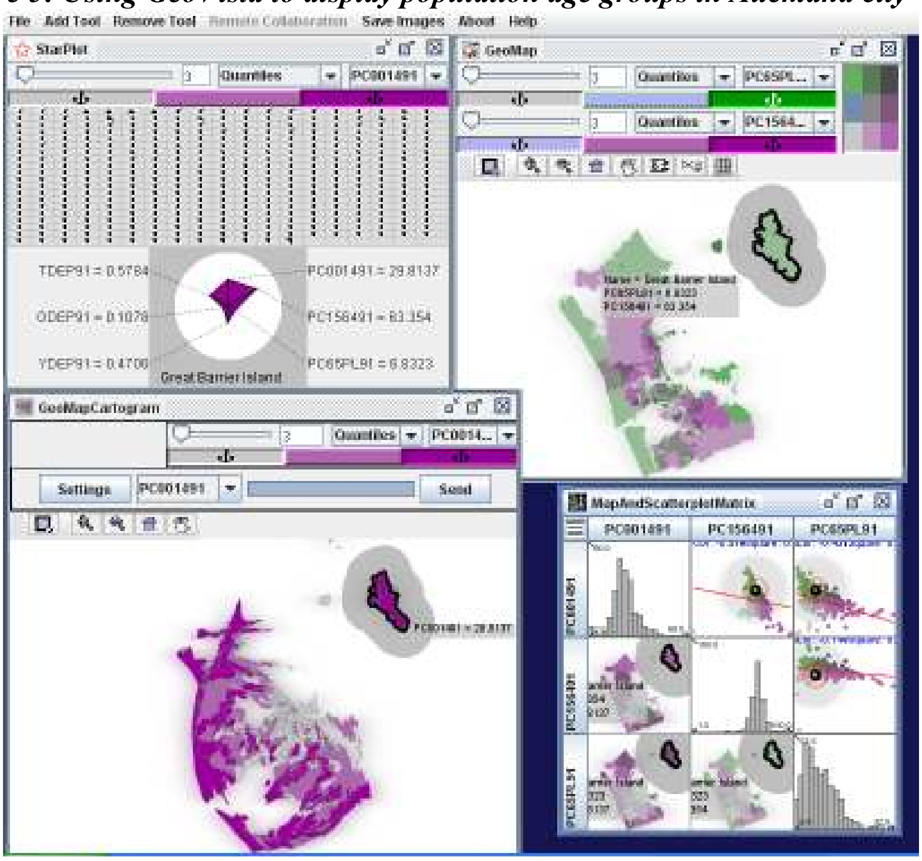


In Figure 4 above, the home locations of staff employed by the central public hospital in Wellington city (that lies on a major fault line) show that many staff live over hills or at some distance from the hospital which could be an issue in risk or disaster (such as earthquake) mitigation planning. These visualizations are useful for any type of flow data, including analyzing internal migration and the geography of the labour market (Ralphs and Goodyear, 2008) and demonstrate the multi-disciplinary nature of statistics.

Example 6: Using Geo-Vista with official data

Census data for a range of derived variables (such as proportions in various age or ethnic groups, with certain household, family or employment characteristics, median income and average population growth) for Auckland city was geo-coded (linked to its geographic area) at area unit level by Statistics New Zealand in a trial of the GeoVista software mentioned above. Figure 5 displays, using 2006 Census data, the proportions of this population that are children (aged 0-14), working age (15-64) and elderly (65 and over) together with the youth dependency (total children divided by total working age population), old age dependency (total elderly divided by total working age population) and total dependency (sum of children and elderly divided by total working age population) ratios in related star plots. The geographic location of the data may help explain the nature of the outliers. Many of the outliers in figure 4 are small islands with tiny populations (and therefore can be used to discuss the effect of small populations on population proportions). The univariate cartogram shows that it is the outer areas of Auckland city (the suburbs) that have the highest proportion of children, of interest to retailers establishing new premises.

Figure 5: Using GeoVista to display population age groups in Auckland city



Policy analysts with a conceptual understanding of statistics can use tools such as GeoVista to visually explore complex official statistics datasets. For example, in New Zealand, Goodyear (2010) examined the geographical distribution of (household) crowding, and the relationship between variables such as ethnicity and crowding at area unit level showing that crowding within Auckland is very skewed, with most areas experiencing little or no crowding and that areas with high levels of crowding are highly correlated with areas with high proportions of people of Pacific Island ethnicity. de Róiste et al (2009) used Geovista to investigate four case studies in health, crime, progress towards urban sustainability (looking at the covariation of dwelling occupancy rate and dwelling density) and labour market clearing (the degree to which people can get the employment they want and employers get the labour they need at a local level). The first two case studies used administrative data (cancer rates and reported crime statistics) together with statistics from the national statistics office. Dynamic geo-visualisations are already in use, for example the real-time geo-visualisation created by Paul Nicholls of Canterbury University (http://www.christchurchquakemap.co.nz/) of the 7.1M September 2010 Christchurch earthquake and aftershocks including the 6.3M February 22 2011 one that resulted in over 150 fatalities.

Conclusion

From even just the few examples given above it is clear that new data visualizations not only allow us to visually demonstrate statistical concepts but also enable us to look at data in new ways. Geo-visualisations, in particular, can be used to reinforce the multivariate and interdisciplinary nature of official statistics. The increased ability to link data with its underlying geography does raise the question of what will be the impact on both the teaching and the analysis of statistics when this linking is so easy that it is no longer appropriate to analyse official data separately from its geography.

Already, the new visualizations are enabling new explorations of government policy concerns with currently existing datasets. New policy questions may also arise as a result of explorations using these tools and it is likely that as users become more familiar with this software that there will be demand for a greater range of analytical tools to be available on, or linked to, official datasets. The benefit of having a dynamic facility added to enable the viewing of changes over time is already apparent. Both longitudinal and integrated datasets (containing data obtained from a number of sources) are becoming more common in official statistics. The potential use of new visualizations with these types of data has yet to be realized.

REFERENCES

de Róiste, M., Gahagen, M., Morrison, P., Ralphs, M., Bucknall, P. (2009). *Geovisualisation and policy: exploring the links.* Official Statistics Research Project Report. Statistics New Zealand.

Dominguez-Dominguez, J. and Dominguez-Lopez, J.A. (2010). A visual approach to the teaching of statistics and probability. Invited paper presented to the Eighth International Conference on Teaching Statistics (ICOTS 8). Available at www.stat.auckland.ac.nz/~iase/

Forbes, S., Ralphs, M., Goodyear, R., Pihama,N. (in press). *New ways of visualising official statistics.* Working paper series. Information and Decision Support Centre. Cairo.

Goodyear, R. (2010). *Using Geovis to determine which crowding index works best in New Zealand.* Internal report for Statistics New Zealand, Wellington, New Zealand.

Hidalgo, C. A. (2010). '*Graphical Statistical Methods for the Representation of the Human Development Index and its Components*' United Nations Development Programme, Human Development Reports, Research Paper 2010/39. United Nations, New York.

Paparistodemou, E. and Meletiou-Mavrotheris, M. (2008). Developing young students' informal inference skills in data analysis. *Statistics Education research Journal,* **7**(2), 83-106. IASE.

Ralphs, M. and Goodyear, R. (2008). *The daily commute: An analysis of the geography of the labour market using 2006 Census data*, Labour Employment and Work Conference Proceedings, Victoria University of Wellington. New Zealand.

Statistics New Zealand (1999). *A Layperson's guide. CPI. All about the Consumers Price Index.*.Statistics New Zealand. Wellington. New Zealand. www.stats.govt.nz.

Statistical Journal of the IAOS. (2008). *Statistical Journal of the IAOS: Journal of the International Association for Official Statistics.*Vol.25, No.3-4. IOS press

ten Bosch, O. and de Jonge, E. (2008). Visualising official statistics. *Statistical Journal of the IAOS: Journal of the International Association for Official Statistics.*Vol.25, No.3-4. IOS press.

Tufte, E. (1983). *The visual display of quantitative data.* Graphics Press, UK