International Journal of Secondary Education
2014; 2(4): 66-72
Published online August 20, 2014 (http://www.sciencepublishinggroup.com/j/ijsedu)
doi: 10.11648/j.ijsedu.20140204.12

Drawing inference from data visualisations

Theodosia Prodromou

School of Education, University of New England, Armidale NSW 2351, AUSTRALIA

Email address:
theodosia.prodromou@une.edu.au

To cite this article:
Theodosia Prodromou. Drawing Inference from Data Visualisations. *International Journal of Secondary Education.*
Vol. 2, No. 4, 2014, pp. 66-72. doi: 10.11648/j.ijsedu.20140204.12

Abstract: This article investigates how 14- to 16- year-old students interpret representations of multivariate data generated by data visualisation tools and how they then seek to construct their own meaningful data visualizations that highlight emerging important aspects of data. Students were asked a single question—about where they would like to live—that involved reasoning about a complex data set with many different variables that they were able to explore using a dynamic visualization tool that allowed them to easily generate multiple visualizations of the relevant data set. Findings show the diverse inferences that students articulated to reason about covariation between multiple variables while using the cycle of inquiry and visual analysis. Students revisited their specific kinds of inferences while using complex data visualisation tools, inventing and revising their visual representations of data. Once they obtained some necessary insight, they readily made an informed decision.

Keywords: Inference, Big Data, Multivariate Data, Covariation, Data Visualisations, Visualisation Tools

1. Introduction

Our world is increasingly data-rich and data-dependent. According to International Business Machines (IBM) [1], every day, 2.5 quintillion bytes of new data are created, so much that 90% of the data in the world today has been created in the last two years alone. The largest leaps forward in the next several decades - in business, and society at large - will come from insights gained through understanding large volumes of data. The new revolutions in the volume of data available to inform decision-making and in the data visualization tools for effectively handling data, have dramatically changed what is possible. Such a change requires an analogous transformation of the way the learner or the citizen thinks, learns, and acquires skills. The Principles and Standards for School Mathematics (National Council of Teachers of Mathematics [2]; Australian Curriculum and Reporting Authority [ACARA], [3]) calls for the development of skills for “data representation and interpretation,” especially with identifying and investigating “issues involving continuous or large count data collected from primary and secondary sources” (p. 42) and describing and interpreting data sets in terms of location (centre) and spread.

The importance of statistical education, coupled with the emergence of powerful visualisation tools, has led to some reconceptualization of the teaching of statistics [4, 5]. There is a need to place less emphasis on simple linear models and more on multivariate descriptions of data, multivariate data visualizations, and a wider variety of models. Students need to become familiar with reasoning about multiple variables, taking account of covariation between multiple variables, and the use of complex visualizations to represent quantities in new ways, using intuitive visual artefacts. However, we know little from empirical research about how students interpret the variety of ways to view any data set within data visualization tools and even less about students’ ability to meaningfully construct data visualisations that highlight important aspects of data. This paper presents data from a study, focusing on how 14- to 16-year-old students interpret the multivariate nature of data and make informed decisions about how to visually represent a data set as they engage within the Gapminder [6] data visualization tool.

2. Theoretical Framework

There are two main theoretical issues that guide the inquiry in this study. The first of these is the concern for how people reason about multivariate data, and common misconceptions in such reasoning. The second is the cycle of inquiry and visual analysis so important in learning and which should be encouraged in education.

Psychological research has shown that reasoning about associations between variables, in both bidirectional (correlations) and unidirectional (regression) analyses, is difficult. Nisbett and Ross (1980) [7] showed that even people with statistical background sometimes struggle with assessing and interpreting associations between variables. In particular, sometimes people notice associations between variables, and covariance where there is none [8]. In studies, where no prior expectations exist, small correlations between two variables tended to be perceived as significant if the base rate of the dependent variable was high. For example, in research studies of artificial diseases and potential symptoms, small correlations between a disease and a symptom are seen as considerable if the base rate of the disease was substantially high [9].

People are heavily influenced by previous beliefs and have a tendency to base their reasoning about associations of data on their previous beliefs about the observed associations that ought to exist between the variables instead of the empirical possibilities presented in the data [8].

Implicit theories about co-variation seem to strongly influence the way people recognize correlation in bidirectional analyses of data when previous conditions about associations are verified. Nevertheless, to identify associations between variables when previous theories hold true, requires a strong correlation between data. In such cases, correlation is likely to be underestimated when reasoning about associations between variables [10]. Lane, Anderson, and Kellam (1985) [11] showed that the format of data presentation seemed to affect the observed co-variation since graphical representations of data tend to encourage judgments of stronger correlation [11]. Erlick and Mills (1967) [12] research has indicated that positively correlated variables are more likely to be understood than negatively correlated variables.

The aforementioned common misconceptions that people hold about multiple data and associations between variables can be substantially eradicated if statistics education programs provide secondary school students with opportunities to work (at an intuitive level) with multivariate data sets and their graphical representations. In fact, it is a crucial necessity to educate contemporary citizens to reason about complex data representations and large global datasets.

Univariate and bivariate data analysis using the typical graphical representations used in schools for the past 40 or 50 years (e.g., frequency and two-way contingency tables, bar charts, box plots, histograms, pie charts, scatterplot) can no longer be a sufficient standard goal for quantitative and statistical literacy. The increasing use of new tools of data visualization aid in the development of new skills that are needed for data interpretation.

Students need to become familiar with the use of visualisation tools as an aid to reasoning about multivariate data. In order to understand students' changing experiences of how to best represent quantitative data and reason about covariation between multiple variables using the power of digital tools to represent quantities and measures in new ways, we focused on the cycle of visual analysis [13] with these new tools (Fig. 1). This diagrammatic representation of the cycle of visual analysis shows both the structure that encompasses the task orientation, and the need to capture what is going on in the minds of individuals as they work, often idiosyncratically, using visualisation tools to examine multivariate data while reasoning about a complex decision.

Morton et al. [13] describe the cycle of visual analysis as the process that starts with some task or question about which a knowledge worker (shown at the centre) seeks to gain understanding.

In Figure 1, the heavy arrows signify transitions between the stages. The total complex decision process is described by following the heavy arrows clockwise around the *knowledge worker* or *problem-solving individual* [13] from the top left.

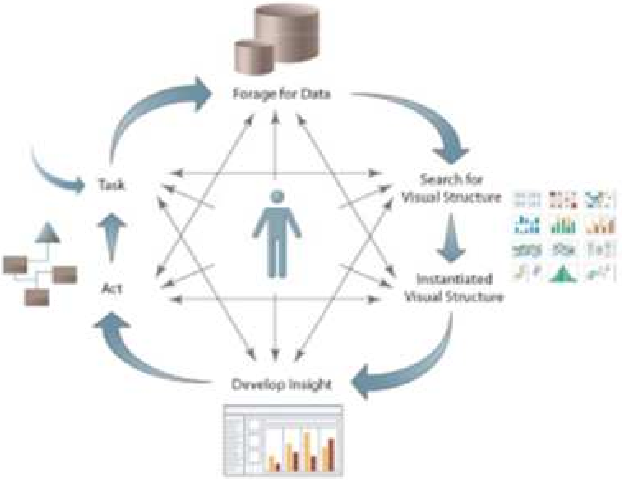

Figure 1. Cycle of visual analysis (Morton et al., 2012, p. 807)

According to Morton et al., [13], the cycle of visual analysis begins with a question Fig. 1. The purpose of the cycle of visual analysis is to provide a scaffolding infrastructure to help the knowledge worker through stages of what can appear as a challenging (or opaque) question. In the second stage, the knowledge worker forages for data from the context of the task that may contain relevant information for constructing a feasible statement and partially answering to the task. Next, they search for an appropriate data visualisation structure of specified variables that highlights important aspects of data. At this point, the knowledge worker instantiates that structure. At the next stage, the knowledge worker interacts with the resulting data visualizations to develop insights about important aspects of data. Once the necessary insight from the data representations is obtained, the knowledge worker can then act. The *knowledge worker* can accept the outputs and either end the cycle of visual analysis; or check for further insight from the data in terms of output expected, and monitor the correctness of output by referring back and forth to the task and looking for other data sets that better capture necessary features of the question.

The remaining structure included in Fig.1 consists of light arrows that are positioned around the *knowledge worker* to show the continuous interactions of the *knowledge worker* with the different stages of the cycle of visual analysis.

This cycle is centred around and driven by the knowledge worker(s)' use of visualisation tools and requires that the visualisation system be flexible enough to support students' feedback and allow students to investigate a variety of exploratory tasks in order to develop stronger understandings of the possible relationships between specified variables.

In this article we will refer to the cycle of visual analysis as the cycle of inquiry and visual analysis, in order to highlight the important role of questions in motivating the cycle.

Because our interest in teaching and learning is central, we need to consider this cycle of inquiry and visual analysis to be more oriented towards the *knowledge worker* [13] or the *problem-solving individual*, to provide not only a better understanding of what students do when using visualisation tools to interpret the multivariate nature of data and reason in their attempt to answer a complex question, but also a better basis for teachers' diagnoses and interventions.

In this research study, we aim to learn more about critical aspects during the process of having students follow the cycle of inquiry and visual analysis.

Ultimately, the research reported here illustrates the processes that students go through when using visualisation tools to examine data while reasoning about a complex decision.

3. The Gapminder World Map

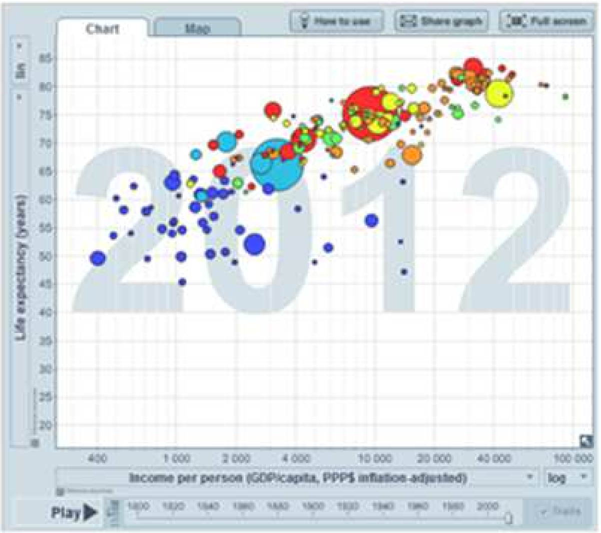

Figure 2. *Default graph on Gapminder*

The "Gapminder World Map" graph, chosen as the visualization tool used in this study, can display several attributes to support analysis of the relations between them. Visualisation tools such as Gapminder have built-in access to large global datasets regarding the economy, education, energy, environment, health, infrastructure, population, society, and work. In this example (Figure 2), Gapminder uses the horizontal axis (income per person), vertical axis (life expectancy), colour (geographic region), and relative size of a bubble (population of country), to display four different attributes of the countries. A fifth attribute (year, from 1800 to present; added by a slider) allows the Gapminder graph to use animation to show change over time.

4. Methodology

4.1. Study Design

This article presents the results of a collective case study [14], in which twelve cases are examined to provide an in-depth insight into students' emerging reasoning about a complex decision based on multiple data, while using Gapminder, a powerful visualisation tool. While the individual cases were of interest, the collective case study allowed us to gain a better insight into the actions, reasoning, and processes across many cases. Although the study was exploratory in nature, the researcher aimed at gaining usable knowledge that would inform students' interactions with data visualization tools. Moreover, the researcher wanted to capture the similarities and differences between various ways of reasoning about multivariate data generated by data visualisation tools and how they then seek to construct their own meaningful data visualizations that highlight emerging important aspects of data.

The descriptions of students' interactions with Gapminder world map, although of somewhat limited generalizability due to the specifics of the technology involved, may still be used to develop preliminary understandings of how students engage with similar data visualisation tools [15] informing therefore future research studies.

4.2. Context and Participants

A total of twenty-four students in Grade 9-10 participated in this research study. The students ranged in age from 14 to 16 years. The author worked with twelve pairs of students. Each pair consisted of a boy and a girl who came from the same class, so interpersonal relationships had been established prior to the research. The participating students were chosen by their teacher, who was asked to select articulate students who would have no difficulty in discussing their ideas with the researcher and who would collaborate well with each other.

Students were familiarised with the Gapminder World Map in two lessons (40-45 minutes each) through a number of introductory activities asking students to make sense of data using the Gapminder World Map to help reason about the relevant data. After the two introductory lessons, each pair of students worked with the researcher. Because the intent was to observe the students as they tried to make sense of a large, complex data set, and the development of their reasoning about big data sets when using the Gapminder world map, the activity was based on a single, simple question that called for reasoning that incorporated many different factors or variables.

It is noteworthy to point out that the activity wasn't structured other than by the guiding question, and that it wasn't an activity about observing "informed decisions," but rather about observing the use of visualization tools in supporting the decision-making processes, whether or not those decisions were correct.

The students were asked to use the Gapminder visualization tool to select appropriate quantities represented in the data to construct visual representations that would help them answer the following question: "Preparing to live in an unknown country in the future: Which is the best country in which to live and work abroad? Consider different variables that impact your decision."

Students were paired to work together to explore the data and discuss their choices, the various quantities represented in the data set, and the distinct variables that influenced their decision. The students' discussions were orchestrated by the participant─researcher.

The pair was then able to use the Gapminder tool while exploring the alternatives. The researcher was present while they worked, and asked questions or made suggestions. The exercise was otherwise unstructured, in order to give the students the opportunity to shape the inquiry, thus allowing observation of the choices they made to organize and understand the data and to come up with an answer to the question they had been posed.

4.3. Instruments and Data Collection Procedures

The author acted as a participant observer who probed for the reasons or intuitions that lie behind students' actions, and frequently intervened in order to tease out the motivation for particular actions that were not so transparent. The author wrote extensive field notes during and immediately after each session. When students' body language or facial expression appeared to be indicative of their conceptual evolution, brief notes were kept to minimize loss of potentially significant data. The researcher found it essential to keep reflective notes that dealt mainly with problems, impressions, feelings about the processes and procedures associated with the study. When issues or themes emerged from students' activities with the Gapminder, the researcher wrote analytic memos.

The data collected included audio recordings of each pair's voices and video recordings of the screen output on the computer activity using Camtasia software. The students were asked to use the mouse systematically to point to objects on the screen when they reasoned about the various quantities presented in the Gapminder World Map graph. Having students to point on the screen helped to supplement the recording of the students' voices and explained their actions and interactions with the quantities illustrated by the Gapminder visualization tool that may otherwise have been subject to many interpretations.

4.4. Data Analysis Procedures

The recordings were transcribed and analysed qualitatively. By drawing on the notes and memos, plain accounts for each pair of students were created. The plain accounts avoided as far as possible interpretations of the transcript. Afterwards, interpretative analyses were developed based on the plain case accounts. These case analyses became the main focus for subsequent analysis and triggered further phases of progressive focusing [16] to identify the key foci and the various characteristics of the phases within the cycle of inquiry and visual analysis. Important similarities and differences between the interpretative case analyses were then identified by constant comparisons [17] of the twelve interpretative case analyses. Once the first set of characteristics has been established, the raw data was revisited to explore the validity of the descriptions, and to evaluate whether all the characteristics of each phase had been adequately captured in the qualitative analysis. When this impression of completeness was not confirmed, at least one new characteristic was incorporated and validated against the data. This iterative cycle was repeated a number of times.

5. Results

For the rest of this article we will focus on the work of one pair of students, George (Ge) and Penny (Pe), while they are looking for larger inferences using multiple representations. Although the same insights as reported below were evident in the analysis of the sessions of other pairs of students, George and Penny's reasoning (in the researcher's view) was representative of the data collected for this research study and provided a clear illustration of how students search for, organise, and generate new knowledge from data. The process starts with the question "which is the best place to live," about which students seek to gain an understanding from their analysis of multiple data resources by foraging for data that may contain relevant information for their analysis task. Next, they sought to construct meaningful data visualizations that highlight important aspects of data and instantiate that visual structure.

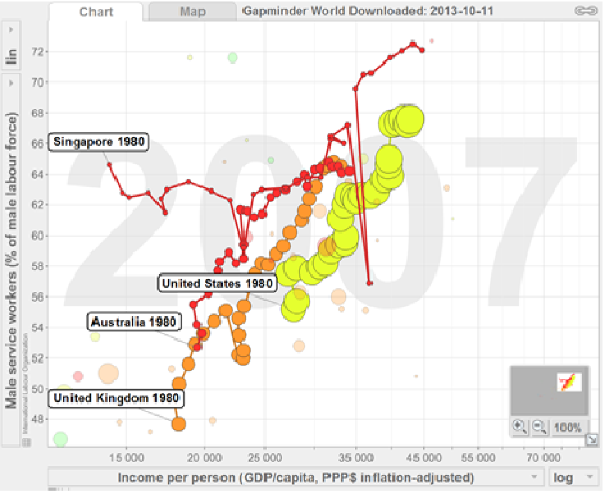

Figure 3. *Graph of male service workers (% MLF) vs. income per person (GDP/capita)*

While looking for the complex inference about which country is best to live in, George and Penny created a series of different visualisation structures. In their first attempt, they created a display that showed GDP/capita (PPP$, inflation adjusted, on the horizontal axis) and "male service workers - % of male labour force (MLF)" (on the vertical axis) for the United Kingdom, Singapore, United States, and Australia from 1980 until 2007 (see Figure 3).

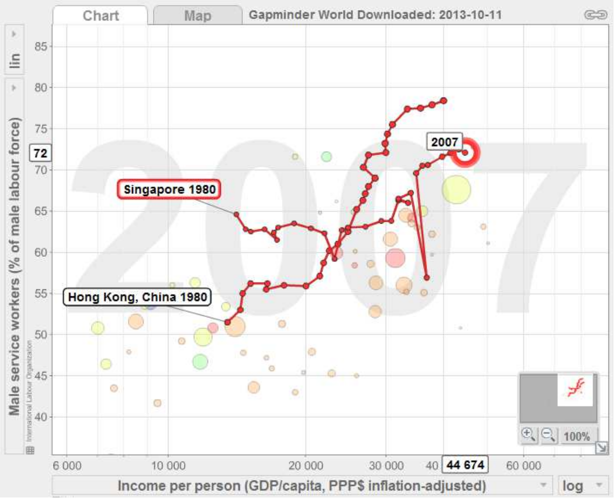

Figure 4. *Graph of male service workers (%MLF) vs. income per person (GDP/capita)*

1. Ge: That's Singapore [pointing within Figure 3], so Singapore paid a lot, and then it …it sort of went down in the 1990 and then income per person increased at the same rate as us, Australia I think, then it drops in 2000 which could have been, I don't know. It took a huge drop there. And male service workers' income sort of dropped from 70% (in 2001) to 57% in 2000.
2. Researcher (Re): Why do you think?
3. Ge: I have no idea. Did anything happen then? Then it shot back up to 72% from 2001 until 2007.
4. Pe: Singapore is the highest.
5. Ge: And it is pretty much steady since 2001 for the last 6 years.
6. Re: What about Australia compared to United Kingdom.
7. Pe: Australia is pretty much the same from, up around here (pointing to 2003) from 2001.
8. Ge: Yeah, Australia and United Kingdom are pretty similar in the amount they pay.
9. Re: Ok, from this data can you decide where you would like to live and work in the future?
10. Ge: It would probably, if you were moving it would probably be best to go to Australia, cause, just out of these four, in Singapore the income per person drops a lot…, if it does drop it drops a lot.
11. Pe: The income per person (GDP) is not stable I guess.

At this point Penny and George constructed a data visualisation and drilled down to details from the resulting data visualisation to develop a better insight that "the income per person in Singapore is not stable" (lines 1-11), which is closely aligned to the statistical insight that one might get by calculating variance or standard deviation. Hence, such an insight is the kind of statistical inference that would be of interest. Nevertheless, once the necessary insight was obtained, students decided to gain further insight from the data and followed alternative paths asking their analytical questions over multiple data sources that allow a context switch. Such a context switch that allowed students to seamlessly combine and visualise data from different data sources that better capture necessary features of the question, thus students naturally would extract the minimum information necessary to accomplish the visual analysis task.

The students selected Hong Kong in the visual representation of "male service workers (% MLF)" versus "GDP/ capita (PPP$ inflation-adjusted)," and engaged with making inferences from the visual representation (Figure 4):

12. Pe: Hong Kong is better.
13. Re: What do you mean?
14. Ge: Singapore pays more because its income per person is greater
15. Pe: But you have a better chance of getting work in Hong Kong, 'cause. . .
16. Re: How do you know that you have a better chance?
17. Ge: We don't know about the unemployment rate
18. Pe: Yes, you are right.

George and Penny created the visual representations of "Male service workers (% of male labour force MLF)" vs. "males aged 25-54, unemployment rate (%)," for Hong Kong, Singapore, United Kingdom, Australia and United States (Figure 5). They then looked at the resulting data visualisations to develop insights from the data.

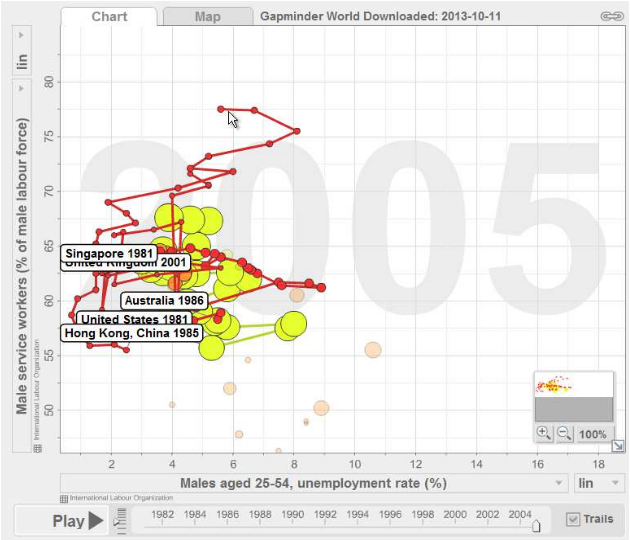

Figure 5. *Graph of male service workers vs. males aged 25-54, unemployment rate (%)*

19. Ge: So Hong Kong it, it has low unemployment rate, well not really no.
20. Pe: So it's Hong Kong that is going up.
21. Ge: Yeah, Hong Kong is this one up here and it has... its unemployment rate jumps and goes down
22. Pe: Australia is steady enough, steady.
23. Ge: Yeah, and America's isn't very steady, because it, oh it just jumps back and forth like this, so, I think you might not have a secure job in United States.
24. Re: So, what will you decide by looking at these graphs?
25. Ge: That, Hong Kong will probably be the best place to work, I guess. Yeah, Hong Kong would be the best.
26. Pe: It has a lower unemployment rate… In Hong Kong unemployment rate is about 6%.
27. Ge: 5.6%.
28. Re: How this will impact on your decision.
29. Pe: Hong Kong has some more problems.
30. Ge: United Kingdom has the lowest unemployment rate out of the five countries. Oh no, sorry, Singapore has a really, really low unemployment rate but its, oh work male service workers are lower. . .
31. Pe: than most other, or, actually Singapore would actually probably be the best country to work at out of the five.
32. Ge: Well, I'm not sure, on the one hand Hong Kong has a higher unemployment rate than Singapore so, maybe to get a job at Singapore is better. Umm, on the other hand I would say Hong Kong over Singapore
33. Pe: Let's explore other graphs that will illustrate only results of Hong Kong and Singapore.

The students created the visual structure that was appropriate for the data and instantiated that structure. They then looked at the data visualisation of different variables from different datasets, and the complexity of data led students to a trade-off (line 32). Students in their attempt to develop insight from the instantiated visual structures, they interacted with the resulting visualissations rolling up to summarise and "explore other graphs that will illustrate only results of Hong Kong and Singapore" (line 33).

The students struggled to visualise the data for the two countries of their interest, so they decided to implement a few principles that would make their exploration easier: (1) reduce the number of (countries) variables from the data visualisations, (2) explore more data visualisations of different attributes of the countries of interest.

6. Discussion

Findings show the diverse inferences that students articulated to reason about data and covariation between multiple variables while using the cycle of inquiry and visual analysis. The results suggest that the students were able to adaptively use the visualization tool to construct visual structures that highlight information relevant to their analysis task. Students constructed visual structures that were appropriate for the data and instantiated those structures. They then interacted with the appropriate visualisations that incorporate important aspects of the data. Using the interactive visualisations, they investigated the co-variation over time of several variables in data sets in order to find meaning in the data. In their justification efforts, students revisited their specific kinds of inferences while using complex data visualization tools, built on them and took new actions based on the needs of their exploratory tasks. The result is an iterative process so that the cycle of inquiry and visual analysis features cycles of invention and revision of visual representations of data. Of course, to look for larger inferences using appropriate visual structures of data demands systematic attention to multiple elements of evidence about the context of the exploratory task. This attention often involves the parallel exploration of multiple variables, and quantities and measures represented in new ways using intuitive visual interfaces.

Such an iterative process is centred around and driven exclusively by the students and requires the visualisation system to be flexible enough to support students' feedback and allow a variety of paths based on the varying student inquiries.

When the students deal with exploratory tasks that require making sense of large collections of data, the students often achieved insight into big data by implementing a few principles that would make their data exploration easier:

(1) explore the trends over time of fewer variables on the instantiated data structures,
(2) When faced with a trade-off in which there was clear best choice, they would create a new visualisation to further explore the question.

In summary, we observed that the students' use of the Gapminder visualization tool allowed students to create a variety of displays of different representations of multivariate data in response to their search for insights into a complex question, and this activity helped students develop mental models of possible relationships between multiple variables that could give them a stronger conceptual basis for considering the formal statistical analysis. In this context of working with complex visualization tools such as the Gapminder, an insight is an inference related to one that could be statistically tested.

The results point to the potential of using complex visualisation digital tools to represent quantities and measures in new ways using intuitive visual interfaces that foster the development of students' mental models of possible relationships between multiple variables.

These novel ways of data visualization have the potential to give rise to different ways of reasoning with quantitative data. The integration of statistical data with geographic data can help students to study how geographic and social aspects of our world may be associated with several trends observed in data. However, new practices need to provide students with appropriate knowledge about quantitative attributes and measures concerning categorical characteristics such as geographic location to reason about complex multivariate data displays.

These new visualisation tools give rise to new practices that aid students' learning, should inform the development of novel learning theories, which, in turn, should inform the future standards and curriculum efforts in mathematics education.

References

[1] IBM. (2002). What is big data? — Bringing big data to the enterprise. Retrieved August 26, 2013, from http://www.ibm.com

[2] National Council of Teachers of Mathematics. (2000). Principles and standards for school mathematics. Reston, VA: National Council of teachers of Mathematics.

[3] Australian Curriculum, Assessment and Reporting Authority. (2011). Australian Curriculum: Mathematics. Version 1.2. Retrieved March 15, 2011, from http://www.acara.edu.au

[4] Ridgway, J., Nicholson, J., & McCusker, S. (2013). Reasoning with Multivariate Evidence. Technology innovations in Statistics, 7 (2), 1933-4214.

[5] Prodromou, T. (2013). Data Visualisation and Statistics from the Future. Proceedings of the 59th ISI World Statistics Congress (Data visualization for youth appeal Sponsoring Association(s)) (Paper 3), p. 1-6. Hong Kong, China: International Statistical Institute (ISI). Online: http://www.statistics.gov.hk/wsc/IPS049-P3-S.pdf

[6] Gapminder Online: http://www.Gapminder.org/downloads/

[7] Nisbett R., & Ross, L. (1980). Human inference: Strategies and shortcomings of social judgment. New Jersey: prentice Hall.

[8] Engel, J., & Sedlmeier, P. (2011). Correlation and Regression in the training of teachers. In C. Batanero, G. Burrill, & C. Reading (Eds.), Teaching statistics in school mathematics-challenges for teaching and teacher education (pp. 97-107). New York: Springer Science+Business Media B.V. 2011.

[9] Vallee-Tourangeau, F., Hollingsworth, L., Murphy , R. (1998). Attentional bias in correlation judgments? Smedslund (1963) revisited. Scandinavian Journal of Psychology 39, 221-233.

[10] Jennings, D. L., Ammabile, T. M., & Ross, L., (1982). Informal covariation assessment: Data-based versus theory-based judgments. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), Judgments under uncertainty: Heuristics and biases (pp. 221-230). New York: Cambridge University Press.

[11] Lane, D.M., Anderson, C. A., & Kellam, K. L. (1985). Judging the relatedness of variables: The psychophysics or cavariation detection. Journal of experimental Psychology, 11 (5), 640-649.

[12] Erlick, D. E., & Mills, R. G. (1967). Perceptual quantification of conditional dependency. Journal of experimental Psychology, 73 (1), 9-14.

[13] Morton, K., Bunker, R., Mackinlay, J., Morton, R., & Stolte, C. (2012). Dynamic workload driven data integration in Tableau. In proceedings of the Special interest Group on Management of Data Conference (pp. 807−816).

[14] Stake, R. E. (2000). Case studies. In N. K. Denzin, & Y. S. Lincoln (Eds.). Handbook of qualitative research (2nd ed., pp. 435-354). Thousand Oaks, CA: Sage.

[15] Stake, R. E. (1995). The Art of case Study Research. Thousands Oaks, CA: Cage Publications.

[16] Robson, C. (1993). Real World Research. Oxford: Blackwell.

[17] Glaser, B. G. (1978). Theoretical Sensitivity: Advances in the Methodology of Grounded Theory. Mill Valley, CA: Sociology Press.