COMPETENCIES AND DISPOSITIONS FOR EXPLORING MICRO DATA WITH DIGITAL TOOLS

Daniel Frischemeier[1], Rolf Biehler[1] and Joachim Engel[2]
[1]University of Paderborn, Germany
[2]Ludwigsburg University of Education, Germany
dafr@math.upb.de

Exploring micro data requires the ability to use digital tools for managing large multivariate data. Digital tools allow changing easily between different uni- and multivariate displays and summary statistics for a deeper insight into the data. In this paper we examine the suitability of several digital data analysis tools for exploring a large, multivariate socio-economic dataset, ranging from educational tools (TinkerPlots, Fathom) to professional software (R). Based on German income structure data, we will point out benefits and limitations of TinkerPlots, Fathom and R for comparing groups, investigating subgroups, analyzing relationships between variables and for exploring multivariate phenomena.

INTRODUCTION

Informed participation in public decision processes requires from the concerned citizen statistical skills and knowledge to make sense of complex multivariate data. Relevant open datasets are nowadays freely and in abundance available through National Statistics Offices, Eurostat, UN, NGOs etc. and can be used for teaching at school and university level. Advantages and challenges of using authentic multivariate large data in the classroom are well recognized by statistics instructors (e.g. Engel, 2007; Hall 2011, Gould 2014). Engel (2016) presents a framework for statistical issues involved in understanding multivariate data including exploration of the data, operationalization of variables, data quality and provenance, comparison of distributions, modeling functional dependencies (including nonlinearity issues), conditional probabilities and frequencies (e.g., investigating subgroups), Simpson´s paradox, critical thinking, drawing conclusions from data and beyond. Due to their high number of cases and correlated variables managing large multivariate data sets require the use of digital tools capable of changing interactively between displays and summary statistics for a deeper insight into the data (see Pratt, Davies and Connor, 2011, p. 100). Additional "implications for the data revolution in statistics education" are discussed in Ridgway (2015) where also a distinction between open and big data is made and implications for a statistics curriculum are considered.

In this paper we examine the suitability of several digital data analysis tools for exploring large, multivariate socio-economic micro data, ranging from educational tools (TinkerPlots, Fathom) to the professional software R. In particular we focus our digital explorations on comparing distributions, investigating relationships between variables and exploring multivariate phenomena. The dataset we refer to in this paper is the 2006 German Income Structure data (short: GIS data), provided by the German Statistics Office as campus file[1] which covers micro data from a random selection of 59,504 adult employees and 16 variables (like monthly income, gender, type of job, region, etc.). These data are suitable for exploring income discrepancies among German employees and in particular to investigate the so-called gender pay gap, i.e., the question why women earn less than men do.

EXPLORING GIS DATA WITH DIGITAL TOOLS

Educational software is designed to facilitate and deepen students' learning of statistics, so TinkerPlots and Fathom as educational software may facilitate data analysis processes and may help learners to follow individual approaches when exploring data. For a description of both tools see Biehler, Ben-Zvi, Bakker and Makar (2013, pp. 653-671). As a recent development, the Common Online Data Analysis Platform (short: CODAP, see http://codap.concord.org/) is closely related to Fathom and TinkerPlots. But unlike Fathom and TinkerPlots, CODAP is free and runs in a web browser. It is open source which means that any developer can modify and extend it. The

[1] http://www.forschungsdatenzentrum.de/campus-file.asp (retrieved on 22 Jan 2016)

professional software R, on the other hand, offers a powerful environment for statistical computing and graphics. R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering…) and graphical techniques, and is highly extensible."[2]

Regarding specific "data analysis dispositions" when exploring data with software one may distinguish (see for example Makar & Confrey, 2014) between a wanderer perspective implying a structured, goal-oriented exploration of data and a wonderer perspective having an exploratory "EDA" approach when walking through the data with digital tools. The dispositions are dependent on the digital tool used for the exploration. Whence, for example, educational software like TinkerPlots rather allows an exploratory wonderer perspective, professional software like R also affords a wanderer perspective, a goal-oriented exploration of data. In the following we outline and discuss capabilities and limitations of educational software when exploring multivariate micro data. Contrasting the potential of TinkerPlots and Fathom, we also illustrate how multivariate micro data can be explored with professional software like R.

EXPLORING GIS DATA WITH EDUCATIONAL SOFTWARE (TINKERPLOTS & FATHOM)

At first we concentrate on the exploration of GIS data with the educational software TinkerPlots, and Fathom.

Comparing distributions with TinkerPlots & Fathom

Amongst the possibility of producing conventional and individual plots via the three main operations "stack", "separate" and "order", TinkerPlots offers the possibility to calculate summary statistics like median and mean for the distribution of numerical variables without the need to use formulas or specific commands. Also, further "conventional" graphs for group comparisons like histograms or boxplots can easily be produced in TinkerPlots. In Figure 1 we see for example the distributions of the variable "hourly_pay" in regard to male and female employees displayed as boxplots in TinkerPlots.

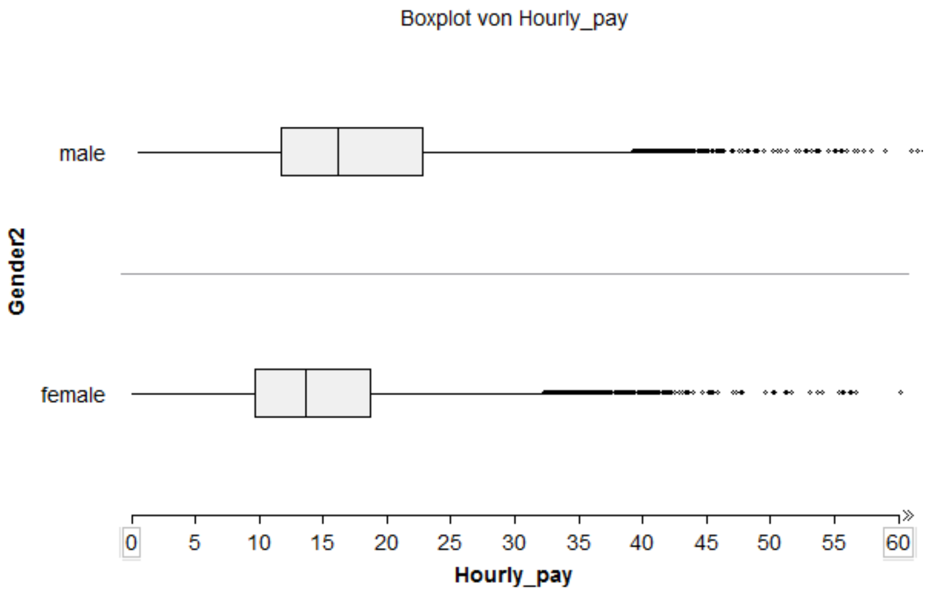


Figure 1. Boxplots of the variable "hourly_pay" (in Euros) distinguished by gender in TinkerPlots.

Due to these features, learners are enabled to focus on the interpretation of graphs, the comparison of summary statistics and on the data analysis itself rather than on programming commands and formulas. In addition TinkerPlots offers the "divider" tool, which calculates the relative frequency of cases in a chosen interval. This allows learners to easily do individual

[2] https://www.r-project.org/about.html (retrieved on 22 Jan 2016)

comparisons like p-based or q-based comparisons (for details on p- and q-based comparisons see Biehler, 2001, p. 110). For example, Figure 2 shows a p-based comparison of the income distribution for male and female employees by highlighting different proportions of employees with hourly income above a certain threshold value and how this proportion differs between both groups (men and women).

We can see that nearly 2.6% of the women earn 30€ or more per hour, the proportion of men earning 30€ or more is higher, approx. 10.6% (Fig. 2).

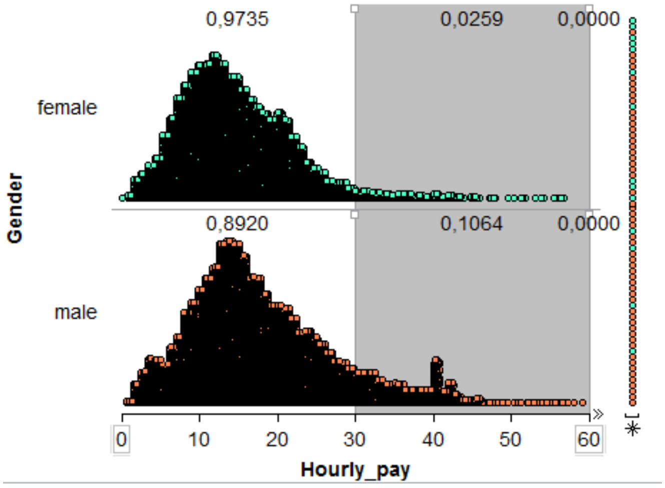
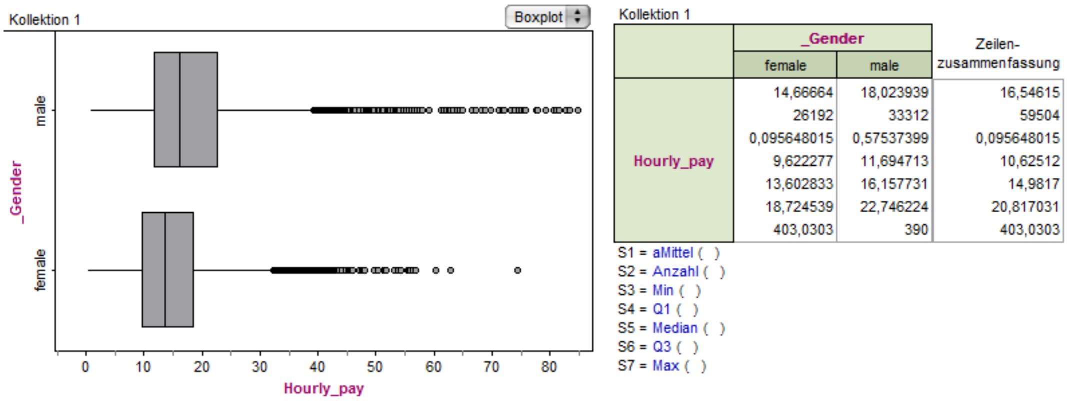


Figure 2. Stacked dot plots of the distribution of variable "hourly_pay" (in Euros) distinguished by gender with dividers in TinkerPlots.

Whereas graphs like dot plots, histograms, etc. in TinkerPlots have to be produced by a sequence of operations ("stack", "separate", and "order"), Fathom offers "readymade" plots and directly produces the desired plot of a distribution of a numerical variable by choosing options like "histogram", "dot plot" or "box plot" (see Figure 3, left, for boxplots of the distribution of the variable "hourly_pay", with outliers identified by dots). Furthermore, Fathom offers the feature to calculate summary statistics (by default mean, median, Min, Max, 1st quartile and 3rd quartile) which are displayed in a table (see Figure 3, right). Here learners can use the exact summary statistics to identify a shift between two distributions on a numerical basis.



Kollektion 1

	_Gender		Zeilen-zusammenfassung
	female	male	
Hourly_pay	14,66664	18,023939	16,54615
	26192	33312	59504
	0,095648015	0,57537399	0,095648015
	9,622277	11,694713	10,62512
	13,602833	16,157731	14,9817
	18,724539	22,746224	20,817031
	403,0303	390	403,0303

S1 = aMittel ( )
S2 = Anzahl ( )
S3 = Min ( )
S4 = Q1 ( )
S5 = Median ( )
S6 = Q3 ( )
S7 = Max ( )

Figure 3. Boxplots of the distributions of the variable "hourly_pay" (in Euros) distinguished by gender (left) and table with summary statistics of the distributions of "hourly_pay" (right) in Fathom.

The capacity of a dynamic and interactive change of the bin widths when using histograms (see Figure 4) opens an explorative, wonderer perspective for learners by investigating additional features of the data distribution such as skewness or multimodality.

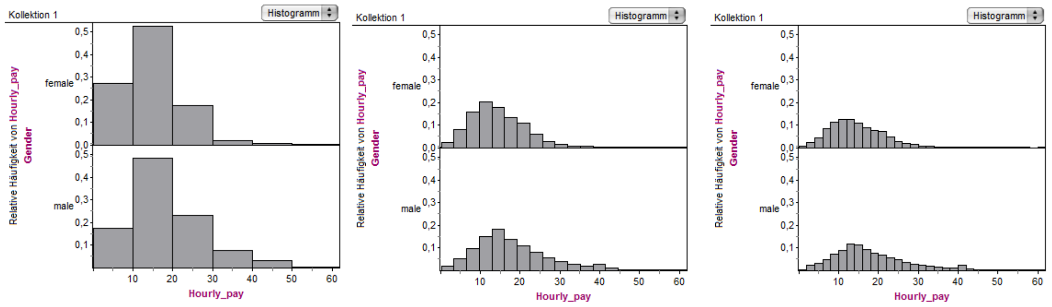


Figure 4. Histograms of the distributions of the variable “hourly_pay” (in Euros) distinguished by male and female employees with different bin widths in Fathom.

To summarize, TinkerPlots and Fathom allow learners to compare distributions of a metric variable with regard to center, spread, skewness, shift, p-based and q-based comparisons. Further possibilities and concepts how educational software like TinkerPlots can enhance learners to compare distributions are discussed in detail in Frischemeier (2017). The flexible and interactive switching of graphical representations and the calculation of summary statistics without needing specific commands helps in particular novice learners to concentrate on the explorative data analysis process instead of being absorbed by handling software commands.

Investigating subgroups with TinkerPlots & Fathom

A deeper insight into understanding the pay gap between the hourly income of men and women may be provided by studying subgroups and by displaying data separated by occupation or position. TinkerPlots offers an easy way of dropping the variables on the axes and separating the “hourly pay” axis completely to take into account the variables hourly pay, gender and profession. In Figure 5 we see the distributions of the variable “hourly_pay” for female employees separated by subgroups according to the type of employment (civil servant, fulltime/part-time; clerk, worker, trainee etc.). Here learners can concentrate on the “center and spread-interpretation” (for details see Biehler, 2001, p. 108) and compare the boxplots to identify differences in “hourly income” between the different subgroups of professions.

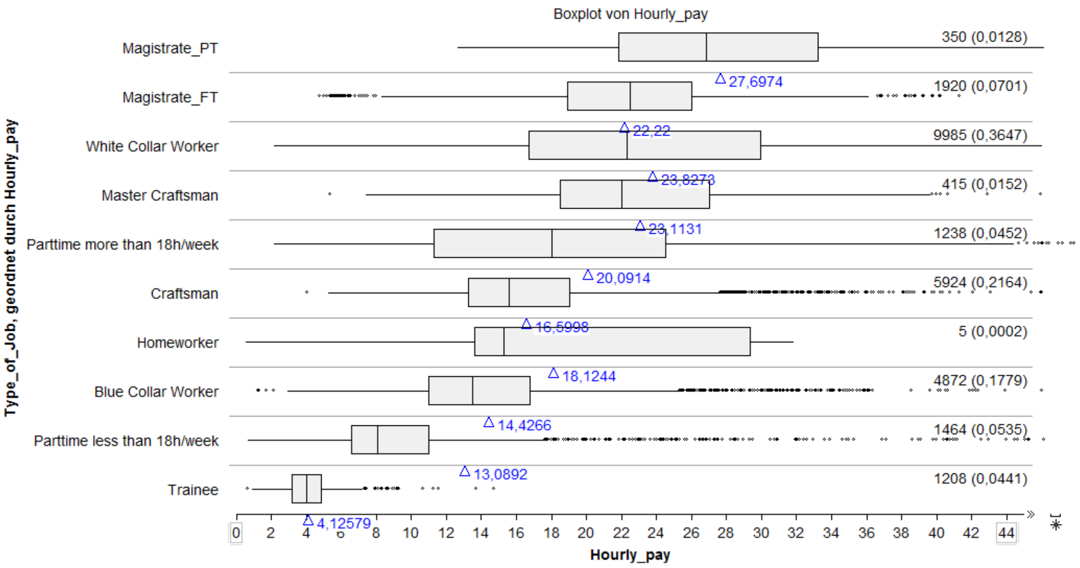


Figure 5. Boxplots of the distributions of „hourly_pay“ of German female employees distinguished by profession in TinkerPlots

However, it is challenging to take into account as many characteristics in the comparison process as possible. So one might wish a further simplification and just want to compare the medians or means of the several distributions of the variable "hourly_pay". This reduction has then to be made by hand, e.g. in Excel. Similar and further explorations in comparing groups and investigating subgroups in the GIS data with TinkerPlots are described in detail in Biehler and Frischemeier (2015).

Investigating relationships between two and more variables with TinkerPlots & Fathom

When investigating the relationship between "hourly pay" and "age", one might produce a scatterplot for investigating a possible trend between these two variables. Although TinkerPlots enables to produce a scatterplot of two numerical variables, one limitation is that TinkerPlots does not offer the modeling of functional relationships in a formal way (e.g., calculating r^2 or displaying a line of best fit). Nevertheless TinkerPlots can be very useful for leaners to identify trends in the data and provides means to get a sense of a relationship between two variables by so called "scatterplots slices" (Konold, 2002). The TinkerPlots display in Figure 6 reorganized the continuous variable "age" into the categories ages 16-25, 26-35, 36-45, 46-55, and 56-65. These "scatterplot slices" can be seen as a more elementary representation of a common scatterplot. The intention is that the students see "each vertical slice of data in this plot as a distribution of a discrete group, [and that] students can apply skills they have learned in comparing two distributions to visually compare the centers of the distributions in the "sliced" scatterplot" (Konold, 2002, p. 3). In this case a first observation might be that the median of the distributions of "hourly pay" tends to increase with increasing "age." Educational software tools like TinkerPlots and Fathom offer pre-stages (like "scatterplot slices") and readymade plots (scatterplots) for learners for a first investigation of relationships between two numerical variables. However, in TinkerPlots no formal methods of modeling (e.g., modeling a linear relationship) are given.

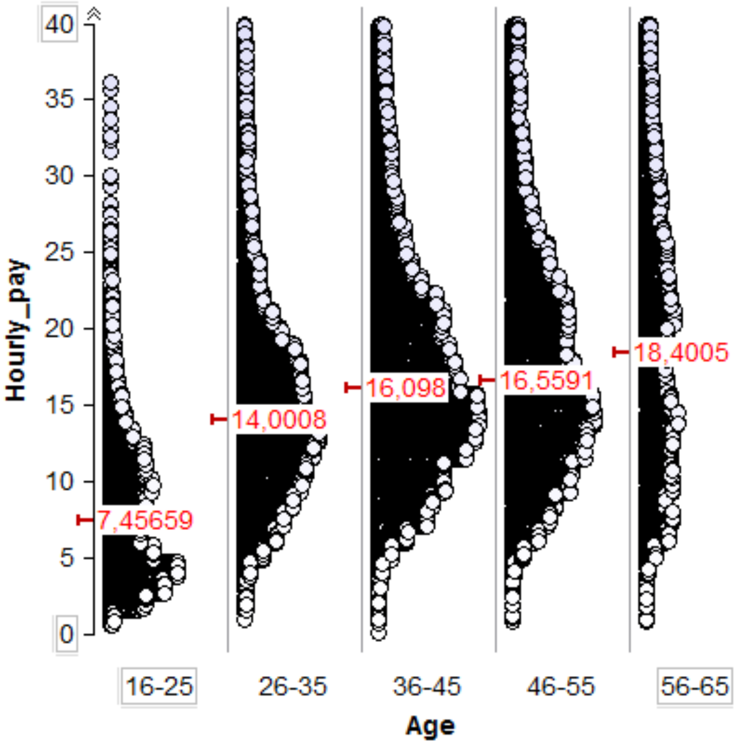


Figure 6. "Scatterplot slices" in TinkerPlots.

In contrast to TinkerPlots, Fathom offers the opportunity to discover the relation between two numerical variables and to model a functional relationship by fitting curves to the data.

Learners can easily drop both numerical variables on x- or y-axis to produce a common scatterplot and can implement regression lines. Thus, Fathom offers amongst sliced scatterplots also the opportunity to discover the relation between two numerical variables in a formal way (see Figure 7, taken from Engel (2014)). Here we can see that "the slopes (18 for men versus 14.7 for women) confirm the sizeable differences in income. The data cloud, however, suggests that the assumption of a linear relationship between the two variables under consideration and the arithmetic mean as measure of comparison is more than questionable. More robust regression techniques including nonlinear and nonparametric regression may be more appropriate." (Engel, 2014)

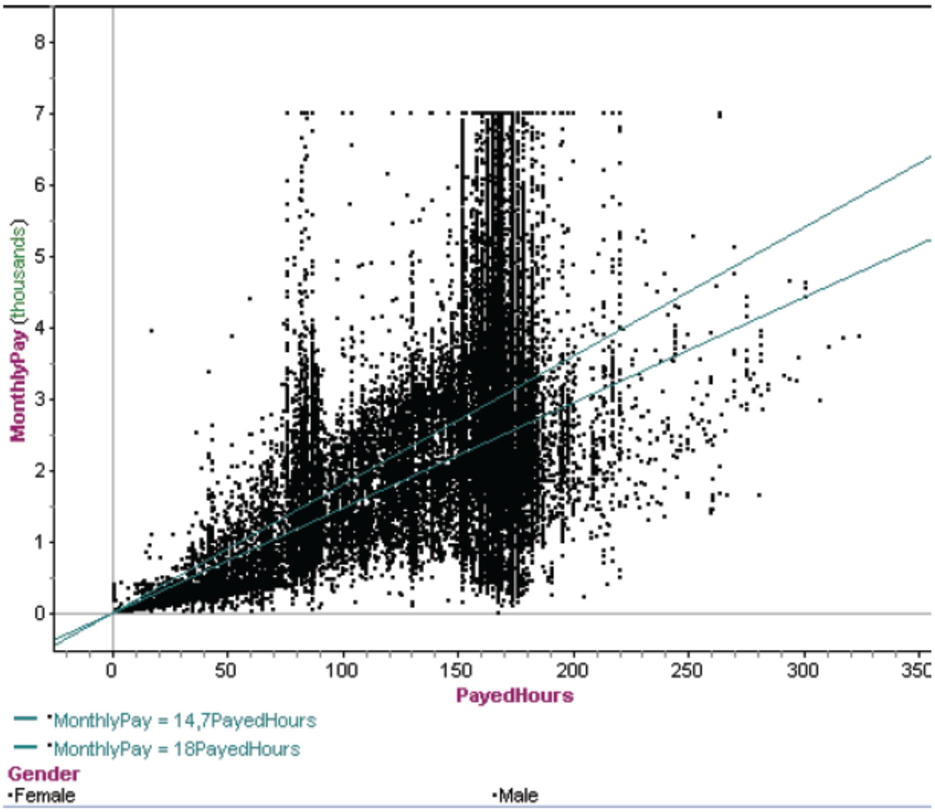


Figure 7. Scatterplot of variables „MonthlyPay“ and „PayedHours“ with regression lines for the subgroups of male and female employees in Fathom (figure taken from Engel, 2014)

To summarize, Fathom offers more formal procedures of analyzing data than TinkerPlots and also enhances an exploratory data analysis style: "Readymade" graphs can be varied by one click, formulas are pre-defined and allow a calculation of different summary statistics. Furthermore Fathom offers a formal exploration of the relationship of two numerical variables and to fit curves in given plots.

Exploring multivariate phenomena with TinkerPlots & Fathom

For exploring multivariate phenomena or non-linear trends in large multivariate micro data it will be useful to consider the interpretation of matrix plots, trellis plots, or regression trees. These types of graphs cannot be produced with TinkerPlots or Fathom.

EXPLORING GIS DATA WITH R

While TinkerPlots and Fathom are designed as educational tools for learning about data exploration in school, R is a language and powerful environment for statistical computing and graphics that represents the state of the art in statistical analysis and graphics. Needless to say that each graphical representation that is possible with Fathom or TinkerPlots can also be created with R, however, for the price of handling the commands of the R language. In this section about exploring GIS data with R, we emphasize two graphical representations of importance when investigating social data with R: estimating distributions without any a priori assumptions regarding shape and exploring functional relationships between a plethora of variables.

Comparing groups with R

Figure 8 shows kernel density estimates for the hourly pay for men and women, scaled such that the sum of the two density curves adds up to the density of the total population. One notices that both distributions (for men as well as for women) are skewed to the right, with a clearly observable shift of male income to the right.

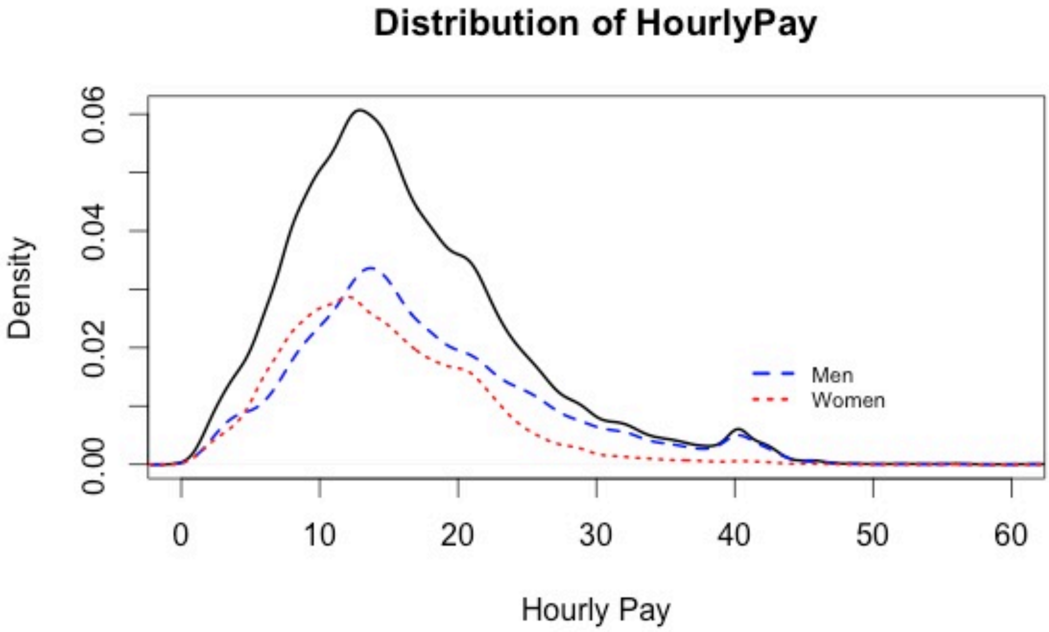


Figure 8. Kernel density estimates for hourly pay rates (in Euros), separate for men (n=33430) and women (n=26382), as well as for the entire sample.

R further provides a whole range of tools for investigating subgroups and for exploring relationships between numerical variables. In the following we concentrate on exploring multivariate phenomena with R.

Exploring multivariate phenomena with R

R offers many different approaches to investigate and display mutual relationships between several variables. The Trellis package of R (Sarkar, 2008) is particularly suited for the visualization of multivariate data. As an example, Figure 9 shows two matrix plots, separated by region (1 = former West Germany, 2 = former East Germany) displaying relationship between the variables, Age, Tenure (years working at the same company), and Hourly Pay. All bivariate combinations of these three variables are organized into a matrix, making it easy to look at the pairwise covariation of these variables. To avoid overloading the graph, the option to suppress printing the large number of data points was chosen. Instead a scatterplot smoother (LOESS, locally weighted scatterplot smoothing, see Cleveland, 1979) for the male and female data was added to visualize correlational dependencies between the three variables, with separate lines for males and females. One striking insight from the upper central panel is the observation that in the West and in the East men and women increase their earnings at roughly equal pace during their twenties, but only in the west at around age 30 the female income stagnates while the male average income keeps on increasing - see also Figure 10. Here it is noticeable, that the average age of a German woman at birth of her first child is at 28.9. The capacity to investigate a matrix of bivariate scatterplots simultaneously, conditioned on other variables, is a very helpful feature when exploring multivariate data.

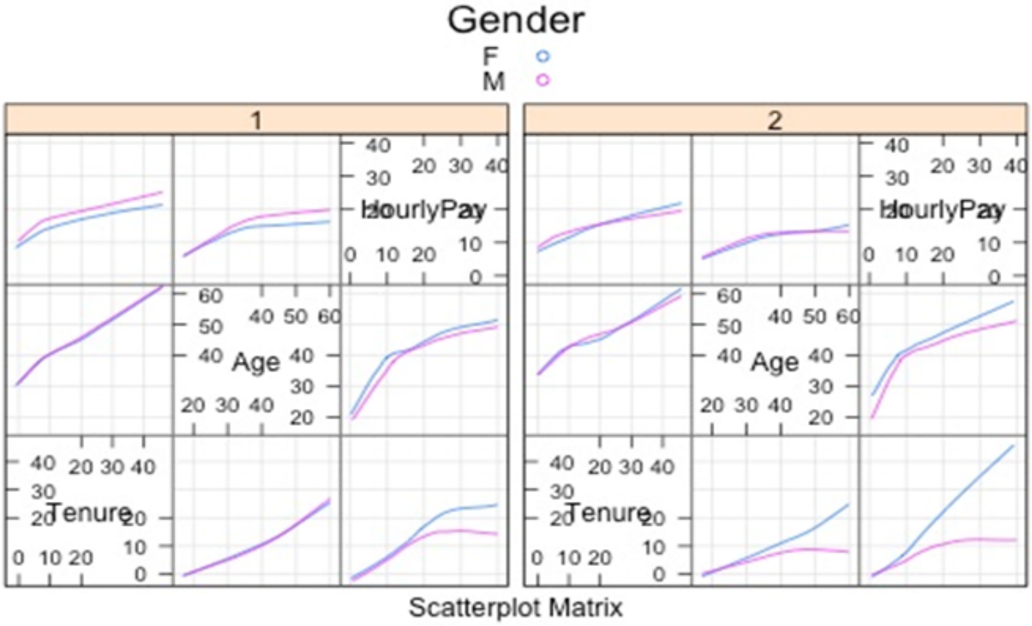


Figure 9. Scatterplot matrix for the variables Hourly Pay (in Euros), Age, and Tenure (years employed by the same company) separated by Region in Germany (1=West, 2=East) produced by R. The curves are scatterplot smoothers for men (red) and women (blue).

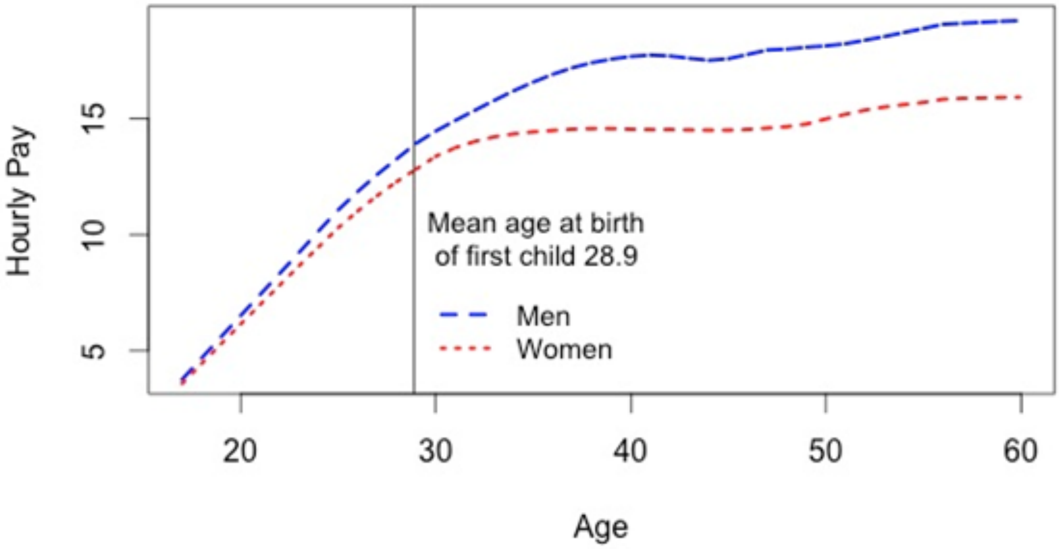


Figure 10: Hourly Wages for men and women in dependence of age. The dashed line corresponds to the average female age at first child birth

A very different approach to explore the structure of multivariate data and dependencies among several variables are regression and classification trees (Breiman, Friedman, Olshen & Stone, 1984) which are easily obtained with R (see Figure 11). Trees are intuitive, conceptually easy to comprehend and result in nice representations of highly complex datasets. While being computationally very intensive, the algorithms proceeds by producing binary splits dividing the original sample into increasingly more homogeneous subsamples. Figure 11 shows a regression tree with averages and sample sizes in each node and the splitting criteria along the edges. Figure

11 reveals that for the GIS data the position held in the company is the most influential criteria determining income. Other relevant variables are age, tenure, and subordinate education, gender and type of job.

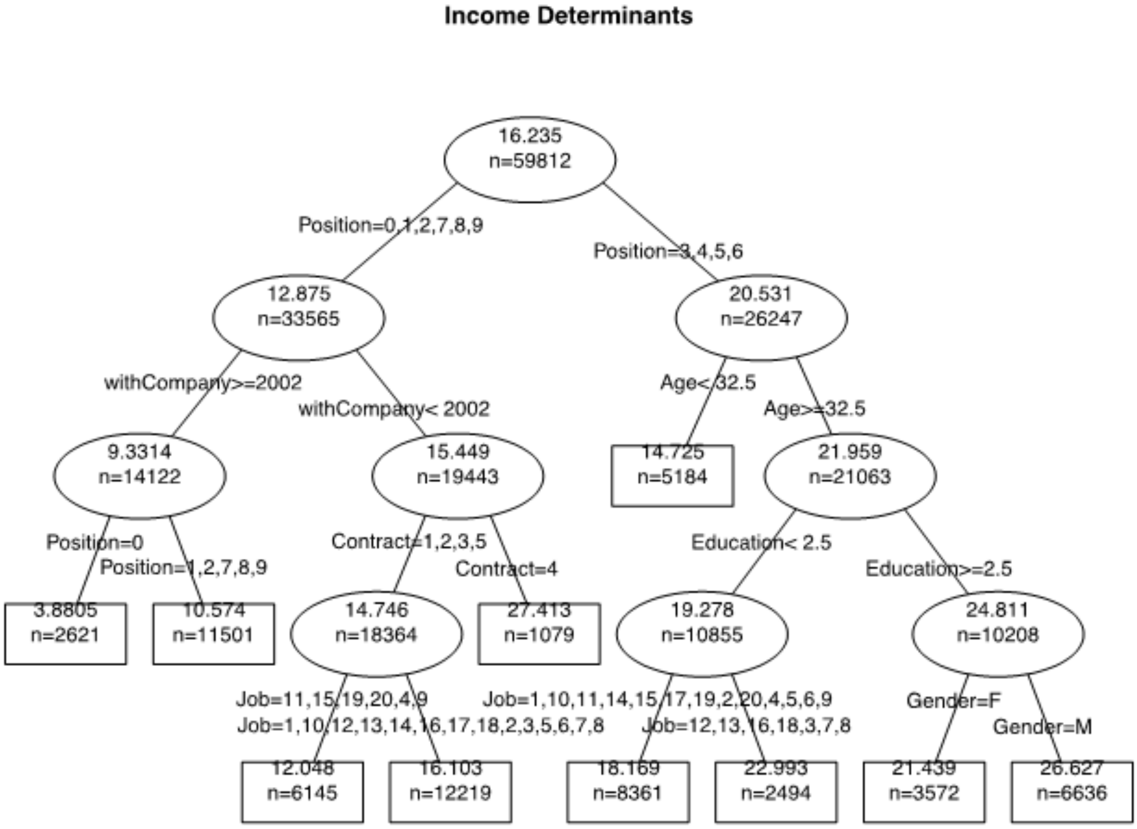


Figure 11. Regression Tree for the GIS data produced by R.

CONCLUSION

Each digital tool has its own advantages and pitfalls. TinkerPlots as a tool for learners (basically for primary and secondary level) allows the construction of conventional graphs and also of self-invented graphs by the three operations “stack”, “separate” and “order”. Also TinkerPlots enhances an exploratory working style since displays can be switched in an easy way. Fathom offers more formal procedures of analyzing data than TinkerPlots and also enhances an exploratory data analysis style: “Readymade” graphs can be varied by one “click”, formulas are pre-defined and allow a calculation of different summary statistics. Thus, comparing distributions in regard to center, spread, shift, skewness or other aspects (p-based and q-based) can be done easily via switching between representations and identifying crucial summary statistics in TinkerPlots and Fathom. Furthermore, learners can also follow individual approaches. “Scatterplot slices”, e.g., are a simplified representation of the relationship among two numerical variables and offer learners an easy access to identify a trend in the data. In addition Fathom offers a formal exploration of the relationship of two numerical variables and to fit curves in given plots. Limitations arise when exploring multivariate data. Here, R is a powerful tool for advanced learners and professional users and offers a whole landscape of powerful displays to explore multivariate data with matrix plots, Trellis plots, regression trees, etc. The price for this variety is that learners have to use a programming language and learn specific commands, a demanding challenge for novice users.

ACKNOWLEDGMENT:

The work reported in this paper was supported in part by ProCivicStat project, a strategic partnership of the Universities of Durham, Haifa, Ludwigsburg, Paderborn, Porto and Szeged, funded by the ERASMUS+ program of the European Commission. However the views and opinions expressed in this paper are those of the authors and do not necessarily reflect those of the funding agency.

REFERENCES

Biehler, R. (2001). Statistische Kompetenz von Schülerinnen und Schülern - Konzepte und Ergebnisse empirischer Studien am Beispiel des Vergleichens empirischer Verteilungen. In: M. Borovcnik, J. Engel and D. Wickmann (Eds.) *Anregungen zum Stochastikunterricht* (pp. 97-114). Hildesheim: Franz Becker.

Biehler, R., Ben-Zvi, D., Bakker, A. , & Makar, K. (2013). Technology for Enhancing Statistical Reasoning at the School Level. In: M. A. Clements, A. J. Bishop, C. Keitel-Kreidt, J. Kilpatrick and F. K.-S. Leung (Eds.) *Third International Handbook of Mathematics Education* (pp. 643-689). New York, Springer Science + Business Media.

Biehler, R. & Frischemeier, D. (2015). "Verdienen Männer mehr als Frauen?" - Reale Daten im Stochastikunterricht mit der Software TinkerPlots erforschen. *Stochastik in der Schule*, 35(1), 7-18.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. I. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, *74*(368), 829-836.

Engel, J. (2007). Daten im Mathematikunterricht: Wozu? Welche? Woher? *Der Mathematik-unterricht*, *53*(3), 12-22.

Engel, J. (2014). Open data, civil society and monitoring progress: challenges for statistics education. In: K. Makar, B. de Sousa and R. Gould (Eds.), *Sustainability in statistics education*. *Proceedings of the Ninth International Conference on Teaching Statistics*, Flagstaff, Arizona, USA. Voorburg, The Netherlands: International Statistical Institute.

Engel, J. (2016). Open data at the interface of mathematics and civics education: Challenges of the data revolution for the statistics curriculum. *Journal of Mathematics and Statistical Science, Vol. 2 (5)*, 264-273, http:// www.ss-pub.org/wp-content/uploads/2016/05/JMSS16010601.pdf

Frischemeier, D. (2017). *Statistisch denken und forschen lernen mit der Software TinkerPlots*. Wiesbaden: Springer Spektrum.

Gould, R. (2014). Datafest: Celebrating data in the data deluge. In K. Makar, B. de Sousa, and R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics* (ICOTS9, July, 2014), Flagstaff, Arizona, USA. Voorburg, The Netherlands: International Statistical Institute

Hall, J. (2011). Engaging teachers and students with real data: Benefits and challenges. In C. Batanero, G. Burrill and C. Reading (Eds.), *Teaching statistics in school mathematics – Challenges for teaching and for teacher education. A joint ICMI/ IASE study: The 18th ICMI Study* (pp. 335-346). Dordrecht, the Netherlands: Springer.

Konold, C. (2002). Alternatives to scatterplots. In: *Proceedings of the Sixth International Conference on Teaching Statistics*. Cape Town, South Africa.

Makar, K., & Confrey, J. (2014). Wondering, wandering or unwavering? Learners' statistical investigations with Fathom. In: T. Wassong, D. Frischemeier, P. R. Fischer, R. Hochmuth and P. Bender (Eds.), *Mit Werkzeugen Mathematik und Stochastik lernen - Using tools for learning mathematics and statistics* (pp. 351-362). Wiesbaden: Springer Spektrum.

Pratt, D., Davies, N. &, Connor, D. (2011). The role of technology in teaching and learning statistics. In: C. Batanero, G. Burrill and C. Reading (Eds.) *Teaching statistics in school mathematics-challenges for teaching and teacher education*. Dordrecht/Heidelberg/London/New York, Springer: 97-107.

Ridgway, J. (2015). Implications of the Data Revolution for Statistics Education. *International Statistical Review*. doi: 10.1111/insr.12110

Sakar, D. (2008). Lattice. *Multivariate Data Visualization with R*. Springer: New York.