THE USE OF MICRODATA VERSUS AGGREGATED DATA IN TEACHING AND LEARNING MIGRATION STATISTICS

Pedro Campos
[1]LIAAD INESC TEC, [2-] FEP - Faceulty of Economics at University of Porto and [3]Statistics Portugal
FEP - R. Dr. Roberto Frias, 4000 Porto Portugal
pcampos@fep.up.pt

Data analysis and visualization of real data is important when it comes to teach and learn social phenomena, such as migration, unemployment, poverty, etc. The interdependency of the variables allows for the use of multivariate techniques along with rich and authentic data from official agencies or other data providers. However, certain phenomena are confidential at individual level, or are not easy to capture, so raw data is not always available for certain types of data. The purpose of this paper is to share a teaching experience at the tertiary level using individual data of migration, where students explore several socioeconomics aspects related to migration in the overall population dynamics using micro and macrodata.

INTRODUCTION: THE ADVANTAGES OF REAL DATA

Data collection and analysis is the heart of statistical thinking, since it promotes learning by experience and links the learning process to reality (Snee, 1993). In order to develop statistical reasoning, one has to incorporate active learning strategies in the curriculum, so that students complement what they have heard and read on statistics. In recent years, the teaching and learning of statistics has become more practice-oriented, and interactive. Statistics is more than a branch of mathematics supported by data analysis: it involves experiences with real-life data and problem-based matters that need careful thinking and reasoning (Garfield and Gal, 1999; Moore, 1992, 1998). The use of real data has been increasingly recommended in the field of statistics education. The GAISE report (Aliaga, et. al, 2005), as well as other national initiatives in Australia, UK, South Africa, and New Zealand all emphasize the need to use real data in teaching. The need for real data is particularly important in the social sciences programs, since many students have negative attitudes towards statistics (Neumann et al., 2013). Both cognitive and affective/motivational factors are associated with using real life data to teach statistics. In particular, there is some association with student specific learning experiences such as relevance, understanding, motivation and engagement. Some good examples of these experiences are described by Neumann et al. (2013): datasets comprised the Forbes list of billionaires, Eruption times of Old Faithful Geyser at Yellowstone Park (USA), Salary of players from the Boston Celtics (USA) and Men's swims times from the Sydney 2000 Olympics. The first three datasets were used for descriptive statistics purposes, while the Sydney 2000 Olympics data was used for teaching outlier detection. In parallel, there has been a growing trend to teach statistics using interactive and dynamic visualization tools. A well-known example is the Hans Rosling's Gapminder 'bubble' graphs (Hans Rosling, 2007), showing the multivariate properties of real data collected from the World Bank, International Labour Organization, the FAO Aquastat database, WHO, etc. More recently, Forbes et al. (2014) explored the use of data visualization tools to assist learning concepts in official statistics. They used several tools, such as iNZight (iNZight, 2016), a free open source software that facilitates the rapid exploration of multivariate data and prioritizes looking at the data. Ridgway et al. (2015) created The Constituency Explorer, containing a tool to explore data visualizations, many with a specific theme, such as Health, Demographics, Ethnicity, or the 2010 Election results in the UK.

MICRODATA AND PUBLIC USE FILES

Microdata is data at the level of the individual respondents also known as raw data. Required essentially by the academic community, users can do more in-depth analyses using microdata, such as regression, correlation, factor analysis, as well as other types of multivariate data analysis. On the other hand, macrodata is aggregated or summarized data commonly used by government planners. In macrodata, analyses can only be based on the summarized or aggregated information available. One specific type of microdata is Public Use Files (PUF). PUF contain individual records that preserve the privacy of the respondents, since their direct and indirect identifiers as well as sensible information are removed. Records for dissemination are edited by suppressing information from direct and indirect identifiers to protect the anonymity of respondents. PUFs are mostly disseminated for free and are often available on-line. Several methods have been appearing in the literature for producing Public Use Files, such as the synthetic data production methods, sampling methods, top and bottom coding, variable suppression, and others (Viana and Campos, 2015). Many National Statistical offices create PUF as a strategy to disseminate their information and give the users a first view of their databases.

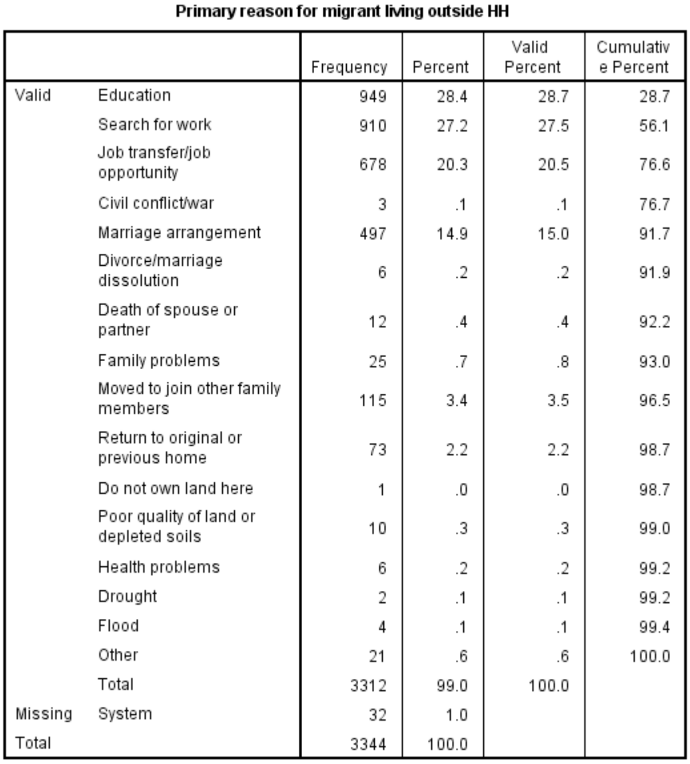
PUF are not exclusive of National Statistical Offices. Other institutions, such as the IPUMS-International integrate, and disseminate samples of census microdata with a major funding by the National Science Foundation and the National Institutes of Health (USA) having become the largest repository of census microdata in the world. With the official statistical authorities of more than 85 countries, encompassing over half of the world’s population, entrusting a total of 249 censuses to the Minnesota Population Center, the IPUMS-International project offers a uniform solution to providing access to Census microdata for policy analysts, researchers, and students while protecting statistical confidentiality (Meier and Lam, 2011). Meier and Lam (2011) describe the use of IPUMS microdata in economic theory to guide empirical analysis of issues such as fertility, marriage, investments in children, and household bargaining and analyze interactions between demographic change and economic change, including the effects of age structure on government programs such as education and old-age support.

DATA AND METHODS

In this paper, we aim at sharing a teaching experience at the tertiary level using migration microdata. A PUF has been accessed containing data from the Migration Household Survey 2009 of Nigeria, delivered in 2013 (World Bank, 2013). The scope of the Migration Household Survey includes household demographic characteristics, housing conditions, household assets and expenditures, household use of financial services, internal and international migration and remittances from former household members, remittances from non-former household members and return migrants. Individual characteristics (sex, age, region, level of education, etc.) are also available in the dataset. We performed a mixed quantitative/qualitative research in one Master Course of the Faculty of Economics at University of Porto during December 2015. The target population comprised 32 students having a good knowledge of data analysis/data mining techniques. All the participants were able to use SPSS software. An electronic questionnaire has been created, through Google Forms, and a (convenience) sample of 20 students that were attending one class answered the questionnaire. The respondents used macro data, such as tables and graphs, to answer the first five questions, with no statistical analysis. These five initial questions were multiple choice. The last question was open, where students were requested to explore variable Age together with two more variables of their choice, and draw conclusions related to the emigration phenomena in Nigeria. For that purpose, students accessed the PUF, available in the SPSS microdata file where they were free to use and apply the methods they wanted.

Two initial questions included a table and a graph displayed to students: Table 1 is a frequency table of the primary reasons for emigration. Here, students had to explain what they had to do to understand those reasons - possible answers were (i) doing a cross tabulating with age or

(ii) sex, (iii) doing cluster analysis and (iv) factor analysis. Students should avoid answers (i) and (iv) due to the type of the variables involved.



Primary reason for migrant living outside HH

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Education	949	28.4	28.7	28.7
	Search for work	910	27.2	27.5	56.1
	Job transfer/job opportunity	678	20.3	20.5	76.6
	Civil conflict/war	3	.1	.1	76.7
	Marriage arrangement	497	14.9	15.0	91.7
	Divorce/marriage dissolution	6	.2	.2	91.9
	Death of spouse or partner	12	.4	.4	92.2
	Family problems	25	.7	.8	93.0
	Moved to join other family members	115	3.4	3.5	96.5
	Return to original or previous home	73	2.2	2.2	98.7
	Do not own land here	1	.0	.0	98.7
	Poor quality of land or depleted soils	10	.3	.3	99.0
	Health problems	6	.2	.2	99.2
	Drought	2	.1	.1	99.2
	Flood	4	.1	.1	99.4
	Other	21	.6	.6	100.0
	Total	3312	99.0	100.0	
Missing	System	32	1.0		
Total		3344	100.0		

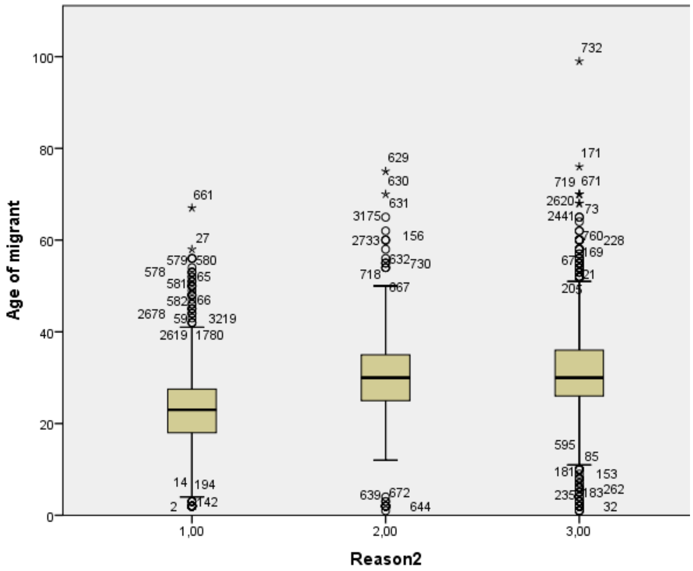


Table 1 – Frequency table of primary reasons for emigration; Fig. 1 – Boxplots of the age of migrants according to their primary reasons for emigration (1=Education, 2=Marriage; 3= Other)

In Fig. 1, students were shown boxplots of the age of migrants according to their primary reasons for emigration (1=Education, 2=Marriage; 3= Other). In this question, students were asked to choose adequate tests of hypothesis (from among different alternatives in the multiple choice questions), to verify the existence of significant differences between groups. They could choose more than one alternative. Correct alternatives included Mann-Whitney tests between any pair of groups and ANOVA/Kruskall-Wallis for testing the difference among groups 1, 2, and 3. Most part of the answers were correct in these five initial questions. In the final (open) question, we could see that most part of students used correlation measures, hypothesis testing and descriptive statistics to explore variable Age with two other variables. Only a relatively small percentage of students of 10% used multivariate techniques, such as cluster analysis, Kruskall-Wallis, and decision trees. Students were able to discover new aspects with microdata that were not accessible with macrodata: they found out that the majority of emigrants are young and males, and that females are younger than males. They also found out that the average education level of the emigrants was high (12 years). Students also explored other variables such as Region, and Number of persons in the household before emigration.

CONCLUSIONS, LIMITATIONS, AND FUTURE RESEARCH

One of the main outcomes of this study is the importance of experiential learning (a learning by doing) that must be supplemented by the written and oral presentation of results (United Nations, 2012). In this paper, students worked with macrodata and microdata to study the emigration in Nigeria. Students showed some ability to understand tables and graphs. In the free question using microdata, students explored different variables. However, we think that there were some obstacles when working with microdata. Students were not able to suggest appropriate statistical techniques to expand the study (such as Factor and Regression Analysis), that could help to better understand the phenomenon. An example of this limitation is that students came to the conclusion, just by using descriptive statistics, that individuals born in the urban centers of Nigeria emigrate at a younger age, but students could not further explore the reasons of that association. Maybe this is due to the fact that there is not enough metadata available for students to understand the relationship between variables. As a consequence, we are aware that this requires a high cognitive demand. Maybe if students could, themselves, formulate a problem, based on data, it would facilitate the analysis. In the future, we aim at doing a qualitative research on this type of experiences, by observing and registering the behavior of students when exploring the microdata.

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

Aliaga, M., Cobb, G., Cuff , C., Garfield, J., (chair), Gould, R., Lock, R., Moore, T., Rossman, A., Stephenson, B., Utts , J., Velleman, P., Witmer, J., (2005). Guidelines for Assessment and Instruction in Statistics Education - college report, American Statistical Association.

iNZight, (2016), Wild, C., and others, A free, easy to use software for statistical data analysis, available at: https://www.stat.auckland.ac.nz/~wild/iNZight/index.php

Forbes, S., Chapman, J., Harraway, J., Stirling, D., Wild, C., (2014). Use of Data Visualisation in the Teaching of Statistics: a New Zealand Perspective, *Statistics Education Research Journal*, 13(2), 187-201

Garfield, J., and Gal, I., (1999). Teaching and Assessing Statistical Reasoning, in L. Stiff (Ed), *Developing Mathematical Reasoning in Grades K-12*, pp. 207-219, National Council Teachers of Mathematics 1999 Yearbook

Meier, A.,& Lam, D., (2011). Creating Statistically Literate Global Citizens: The Use of IPUMS-International Integrated Census Microdata in Teaching, *Statistical Journal of the IAOS*, 27(3-4), 145–156.

Moore, D. S. (1992). Teaching Statistics as a Respectable Subject, in . F. Gordon and S. Gordon (Eds), S*tatistics for the Twenty-First Century*, 14-25, Washington, DC: The Mathematical Association of American.

Moore,D. S. (1998). Statistics Among the Liberal Arts, *Journal of the American Statistical Association*, 93, 1253-1259.

Neumann, D., Hood, M.,& Neumann, M. (2013). Using real-life data when teaching statistics: Student perceptions of this strategy in an introductory statistics course, *Statistics Education Research Journal,* 12 (2), 59-70

Ridgway, J, Nicholson, J., Sutherland, S.,& Hedg, S. (2015) Strategies for Public Engagement with Official Statistics, in: M.A. Sorto (Ed.), *Advances in statistics education: developments, experiences and assessments. Proceedings of the Satellite conference of the International Association for Statistical Education (IASE), July 2015*, Rio de Janeiro, Brazil.

Rosling, H. (2007). Gapminder [Computer software]. GapMinder Foundation. [Online: http://www.gapminder.org]

Snee, R., (1993). What's Missing in Statistical Education?, *The American Statistician*, 47, 149-154.

United Nations (2012). *Making Data Meaningful Part 4: A guide to improving statistical literacy*, Geneva

Viana, I., Campos, P. (2015). A New Method of Generating Synthetic Data for Public Use Files, in *Proceedings of ISI 2015, World Statistics Congress*, 26-31 July 2015, Rio de Janeiro

World Bank (2013). Household Surveys for the African Migration Project in Nigeria, Zibah Consults Limited.