101. Data Quality (Dr. Albert Esteve)

European Doctoral School of Demography (EDSD)

Gonzalo Garcia

7/11/2020

A1. Write in a concise way about the types of data that you have used so far in your career and the topics you wanted to study with these data. Discuss their limitations.

I have worked with both quantitative and qualitative data, from data sources as diverse as Surveys, Longitudinal Data and Digital Trails (Big Data), covering topics conditioned by the workplace I was in that moment: I have used log registries for generating customer journeys inside an e-commerce webpage to understand possible improvement points of the web design; I have used Surveys to understand the customers of an insurance company and plan better marketing campaigns. I used qualitative data for my master's thesis when I scraped the World Bank's blog web page to understand the discourses' shifts in the last 10 years regarding the recommended education policies for the Global South. As for Longitudinal Data, I analyzed the European Social Survey to understand if the level of immigration in European countries had any effect on the equality and social justice perceptions on European citizens.

The limitations of each type of data are different: digital trail information could be biased since the individuals that use the apps could be actually a sub-sample of the population that wants to be analyzed (for example: political perceptions through Twitter data is biased). The Longitudinal data has attrition problems, meaning that some individuals will be lost across time and the final sample will be smaller than the initial one, hence the project becomes more expensive. In the case of Surveys, the initial design both of the sample and the questionnaire are extremely important to overcome statistical biases and missing data problems.

A2. Identify two data sources of demographic interest that provide longitudinal data at the individual level and provide a short description about how the micro and metadata can be accessed.

1. European Social Survey: cross-national longitudinal data at individual level on socio-demographic features. Information on well-being and health & care available for different waves. The micro data is accessible through registration (more straightforward than IPUMS: no need to declare what kind of research will be conducted with the data) and is downloadable in SPSS, SAS and Stata format. The metadata access is (sorry for the non-academic vocabulary) just beautiful: both in online and pdf format there is a list with the name of the variable in the data file, the exact question posed by the interviewer, the possible results and a summary statistics of missing values and the format of the variable. Access to documentation on methodology is easily available through the web page.

2. EPH (Encuesta Permanente de Hogares – Argentina): Longitudinal data at individual level about socio-demographic and economic features from people living in urban areas in Argentina (>500k inhabitants). Microdata downloadable without any registration, in xls or txt file format. The metadata is accessible through a unique pdf file comprising the codebook; not as fancy as the ESS but a development from the R community in Buenos Aires (package "*eph*" in CRAN) allows us to have a better access to the data (download), a second data dictionary that complements the official one and functions to create most used features in socio-economic research.

A3. As part of the Global Burden of Disease project, the Lancet published a paper last July with population scenarios for 195 countries in the world. Explore the paper and its appendixes to identify the data sources used in their forecasts. Select only three countries. At least one of these countries should be your country of birth or nationality.

I have chosen Argentina, Colombia and Spain as countries to follow and search for the data sources for the forecasts. The data sources used by the GDB 2017 can be found in the following link: http://ghdx.healthdata.org/gbd-2017/data-input-sources

Mortality:

- Argentina: Census data (IPUMS 1970, 1980, 1991) and Vital Registrations taken from UN Demographic Yearbook and WHO Mortality Database.
- Colombia: Colombia Demographic and Health Survey, Vital Registrations from WHO Mortality Database and UN Demographic Yearbook.
- Spain: Spain Vital Registrations from 1975 to 2015 and Human Mortality Database (HMD).

Fertility:

- Argentina: Census data (IPUMS 1970, 1980, 1991, 2001/2) and UN Demographic Yearbook from 2000 onwards.
- Colombia: Colombia Demographic and Health Survey (1986 to 2015), Census data (IPUMS 1973, 1985), Vital Statistics (2000 onwards) and Surveys conducted by Greece (2000) and Cyprus (2000-2010).
- Spain: UN Demographic Yearbook from 1997 onwards, and the Human Fertility Database.

Migration:

- Argentina, Colombia and Spain: United Nations Population Division Migration Report 2017.

Education attainment (fertility covariate):

- Argentina: Census data (IPUMS 1970, 1980, 1991, 2001/2, 2010), Argentina International Social Survey Programme (2007, 2010, 2012) and Argentina Permanent Household Survey (2016).
- Colombia: Census data (IPUMS 1964, 1973, 1985, 1993, 2005), Colombia Demographic and Health Survey and Colombia World Fertility Survey (1976).
- Spain: Eurobarometer Surveys, Spain International Social Survey Program, Census data (IPUMS 1981, 1990, 2001, 2011).

Contraceptive met need (covariate fertility):

- Argentina: Argentina Multiple Indicator Cluster Survey 2011-2012.
- Colombia: Colombia Demographic and Health Survey.
- Spain: Spain Fertility Survey (1977, 1994, 1998, 2006) and Spain Knowledge, Attitudes, and Practices Survey (1985)

I know I reached the 250 words, but I would like to point out some things that does not look good in the data:

1. As can be easily noticed, the data comes from different sources and both Latin American countries have more sources for each feature compared to Spain. The harmonization of this many data sources cannot be an easy task and Global South countries (with worse data quality) suffer more from this approach.
2. In Colombia fertility's data source, there are two surveys conducted by Greece and Cyprus. This could be a mistake when retrieving the information sources or this could be real and, in that case, how can we be sure that fertility is being measured in a similar way between two different statistical offices? For example, in the case of fertility, the contraception methods could be different (tagged as traditional and modern in GBD and not both available for every country).
3. Given that the code for each measure is available (https://github.com/ihmeuw/ihme-modeling/tree/master/gbd_2017), it seems that they are using raw data from the cited data sources without any consideration for under registration. Again, those countries with worse data quality and sub-registration will be problematic for comparison.

A4. IPUMS international and the Generations and Gender Survey have different approaches for granting access to their microdata. Please, compare both approaches, make a critical assessment and discuss advantages and disadvantages.

IPUMS has a one-time registration and signing of usage license while GGP is also a one-time registration (for creating user) but every time you request new data you have to sign a new usage license and send it be mail to GGP administration, while in IPUMS the license is sent automatically. In this case IPUMS is faster and less bureaucratic.

I have not downloaded data from GGP but it seems that no one asks what you are planning to research with the data, while for IPUMS that is a required information for granting access to the data base. If I were the owner of the data, GGP flow has less control to grant access in relation to the actual use of the data for research, which could be a weak point.

The possible advantage I find in GGP is when requesting the information: complete survey is available just by clicking, while in IPUMS the navigation tool bar goes by variables. Of course, you should know beforehand the kind of data available in each dataset for this feature to be an advantage. For IPUMS it took me a while to understand that I had to select each variable and at the end request the data extraction. Maybe if whole survey/census data were available for user in case was required, that could save time for the researcher.

A5. Longitudinal and panel household data are very popular in Demography nowadays. Understanding Society - the UK household longitudinal study – and the German Socio-Economic Panel (DIW) are two of the most important panels. Please, explore their metadata and find and comment relevant information about attrition problems.

The UK survey has an attrition problem of around 38% no response. This can be examined by seeing the variable "person number" (*a_pno*) in the metadata. At the same time, Lynn (Lynn et al. 2012) analyzes the reasons for the attrition: in the overall sample, for the first wave 36% of the attrition had as a cause the refusal of the person/household to continue being surveyed, and 2% was because the person/household couldn't been found. The proportions have stayed stable across time, even though the University of Essex has different approaches to lower the attrition (Lynn 2020). Benzeval (Benzeval et al. 2020) compares the attrition problem in this survey with other UK surveys and shows that the level of non-response is similar across all studies.

As for the DWI, Kroh (Kroh et al. 2018) analyzes the attrition problem for the survey for both causes: refusal and unable to follow up. In this case the overall attrition proportion is lower than UK's: spans from 0.5% and 3.2% for unsuccessful follow up and from 10.7% to 27.6% (conditioned to the sample) for the last wave of the survey. The metadata of the DWI does not give information about missing data for any variable. Unlike the UK survey, there could not be found any methodology document explaining the approach to reduce the attrition problem, but there is a better understanding (for each sample) of the main reasons for the attrition, differentiated between survey-related (refusal and no contact) and survey-unrelated (deceased, moved-abroad, under 16).

A6. Explore and select data from IPUMS (at least one sample). Download and read the data in R and make a simple frequency of educational attainment by sex among men and women aged 25-34. Copy and paste the resulting table and the syntax in R.

Data for France, Greece, Italy, Portugal and Spain was downloaded from IPUMS for year 2011 regarding educational attainment and sex of the respondent. The resulting data set comprises 3,753,647 individuals and 18 variables.

In the following table there is a quick summary of percentage of population by educational attainment for each country:

	France	Greece	Italy	Portugal	Spain
Less than primary	6.68	2.17	3.28	6.00	1.83
Primary (first stage of basic education)	4.94	7.76	0.00	12.44	7.10
Lower secondary (second stage of basic education)	3.36	10.28	26.99	24.36	26.51
Upper secondary	39.56	43.75	47.03	25.53	34.43
Post-secondary non-tertiary education	0.00	15.25	0.60	2.90	11.37
University completed	45.46	20.79	22.09	28.76	18.76

As can be seen, there is no registry for Primary for Italy and almost no Post-Secondary for Italy and France. That is why it was decided to create new groups for educational, as follows:

- *Less than primary* and *Primary (first stage of basic education)* in *"01 - Primary Education or less"*
- *Lower secondary (second stage of basic education)* and *Upper secondary* in *"02 - Secondary Education"*
- *Post-secondary non-tertiary education* and *University completed* in *"03 - Post-secondary + University"*

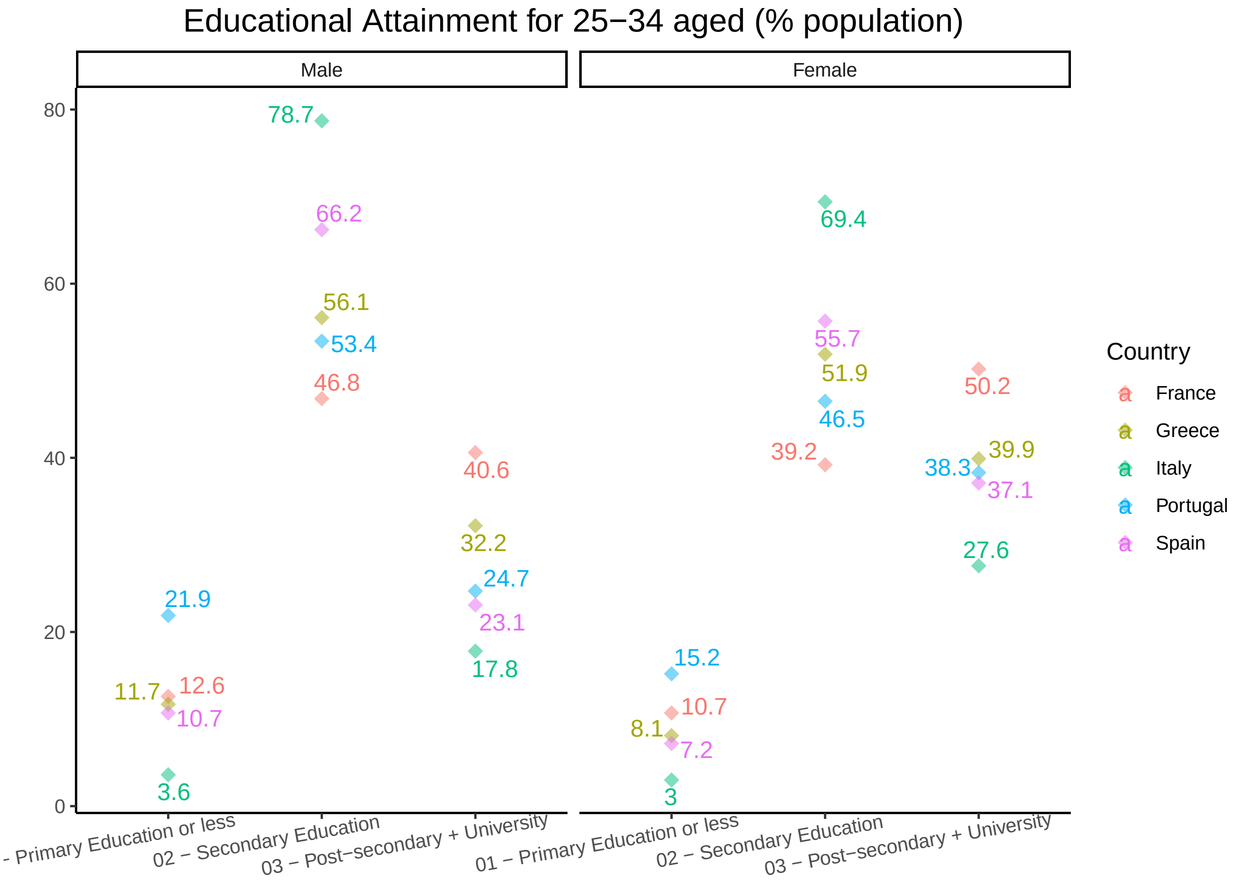
Giving the following results:

	France	Greece	Italy	Portugal	Spain
01 - Primary Education or less	11.62	9.93	3.28	18.44	8.93
02 - Secondary Education	42.92	54.03	74.03	49.90	60.94
03 - Post-secondary + University	45.46	36.04	22.69	31.66	30.13

And by analyzing the differences by sex:

	Male	Female
01 - Primary Education or less	11.63	9.50
02 - Secondary Education	52.86	44.73
03 - Post-secondary + University	35.51	45.77

It is also possible to represent the tables in a plot format:



We can see that women attain University degrees (as proportion of population) more than men do, for each country. France has the higher proportion of its 25-39 population with University degree compared to the other countries, while Italy and Spain are the two countries where Secondary Education attainment seems to be preferable, with men leading the ranking.

A7. What would you like to study in your EDSD Master's thesis? Which data are you planning to use? Which are their limitations? How would your ideal data look like?

I am interested in understanding the attributes that Barcelona's neighborhoods have to attract high skilled migrants. For doing so, I have registration data (*empadronamiento*) from Barcelona's town hall, provided by Dr. Antonio López Gay, and I am planning to use other variables accessible through government sources (middle income, political preferences, age, marital status and ethnic composition) and use Big Data sources to provide soft factors by scraping different digital sources (Idealistas.com, Twitter, Instagram, Google Maps).

I already know a problem that I will have with registration data: in case of Argentinians that entered Spain with a European passport, they will be registered as Europeans instead of Argentinians. The same will happen to any migrant that satisfy that condition; I know that in case of Argentinians the proportion is much higher.

As for the Big Data sources, in some cases I could have a problem with under-representation since not everyone has an Instagram or Twitter account, but I think that the population group under study should have a lesser bias. I could also have a problem for accessing the data: companies are very jealous of data; I know I can scrap Google Maps to look for bars per neighborhood but I do not know the “rules of engagement” for scraping Idealistas. Finally, since I am going to work with 2016-2018 registration data, I have to find out in which web pages can I find a cache of the digital trails I am interested in for that specific point in time.

Reproducibility

All necessary code for reproducing this assignment can be found in the following link:

https://github.com/gonzalofichero/EDSD_101_Data_Quality

References

Benzeval, M., C. Bollinger, J. Burton, T. Crossley, and P. Lynn. 2020. “The Representativeness of Understanding Society.” Understanding Society at the Institute for Social Economic Research; Working Paper No 2020-08.

Kroh, Martin, S. Kühne, R. Siegers, and V. Belcheva. 2018. “SOEP-Core-Documentation of Sample Sizes and Panel Attrition (1984 Until 2016).” German Institute for Economic Research (DIW Berlin); SOEP Survey Papers No 480.

Lynn, P., J. Burton, O. Kaminska, G. Knies, and A. Nandi. 2012. “An Initial Look at Non-Response and Attrition in Understanding Society.” Understanding Society at the Institute for Social Economic Research; Working Paper No 2012-02.

Lynn, Peter. 2020. “Methods for Recruitment and Retention.” Understanding Society at the Institute for Social Economic Research; Working Paper No 2020-07.