GETTING AND VISUALIZING WORLDWIDE CENSUS AND SURVEY DATA

An introduction to IPUMS
Instructor: Victoria Prieto Rosas
Department of Sociology and Anthropology
vprieto@clemson.edu

1. Script

    1.1. Which are the main IPUMS projects?
    1.2. What does IPUMS?
    1.3. Why is it useful for research on Health, Sociology, Economics, Demography, Geography and Environmental Studies?
    1.4. What kind of research questions might be addressed from IPUMS?
    1.5. How to use it?
    1.6. Download data from IPUMS / SPSS for aggregating data / Tableau for visualization

2. Which are the main IPUMS projects?

Ipums, a non-profit scientific initiative from the Minnesota Center for Population (University of Minnesota) has nine ongoing projects, all of them available at
www.ipums.org

IPUMS USA and IPUMS CPS

Projects for harmonization across time of American Community Surveys and Current Population Surveys

Time span CPS: 1962-2016
Time span US Census: 1850-2000 (every 10 years)
Time span ACS: 2001-2015 (annual)
Geo span: country/state/metro/city

What can I find at IPUMS CPS/ACS that isn't available at US Census Bureau? Harmonization and variables that might be comparable to other countries census waves

Census data (IPUMS International)

Project for the harmonization across time and countries of census data.
Time span CPS: 1960-2010 (not available for all countries)
Countries: 82
Census: 277
Person records: 614
Data: household/individuals/
Topics: housing, living arrangements, demographics, socioeconomics
Geography: national/1st subnational /2nd subnational

Time Use Surveys Harmonized (IPUMS TUS)

This project includes individual-level time use data extracted from diaries. The data extract systems make it easy to create data sets containing time use and other variables a user needs, i.e. you can build durations and time allocation on line and add them to the sociodemographic variables of your extracts. This is one of the newest, and therefore least exploded projects.

Clemson Center for Geospatial Technologies

Demographic and Health Surveys (DHS IPUMS) Harmonization across time and countries of DHS conducted in low- and middle-income countries for over 30 years. Include harmonized thousands of variables on health, well-being of women, children, and births.

You can decide if unit of analysis is women, children of births, and build de extract for those specific groups.

Time span: 1988-2014 (irregular timing by country, with some only having a few samples)
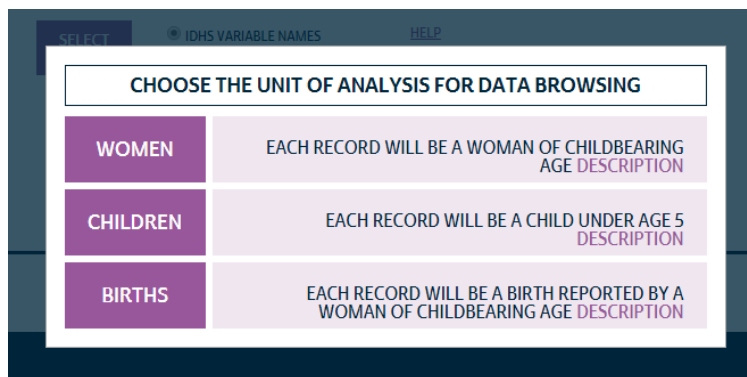Countries: 21

1. Benin
2. Burkina Faso
3. Cameroon
4. Cote d'Ivoire
5. Egypt
6. Ethiopia
7. Ghana
8. Guinea
9. India
10. Kenya
11. Madagascar
12. Malawi
13. Mali
14. Mozambique
15. Niger
16. Nigeria
17. Rwanda
18. Tanzania
19. Uganda
20. Zambia
21. Zimbabwe

IPUMS NHGIS

The National Historical Geographic Information System (NHGIS) provides population, housing, agricultural, and economic data, along with GIS-compatible boundary files, for geographic units in the United States from 1790 to the present (for ACS this covers the shorter period 2011-15). This project is mainly devoted to providing GIS files where several layers of georeferenced data were harmonized across time using a time comparable shapes for the country, state, county, zip code and metro area.

However, IPUMS International also include a session for downloading the country, 1st and 2nd sub-national level shape data files (**Figure 1**).

Figure 1 – GIS data at IPUMS International

Select this label in case you want to download all shape files for world/country/1st and 2nd subnational level data.

Let´s do it because we will use soon.

3. What does IPUMS?

Ipums harmonizes data; this means they work on creating categories and codes that are interchangeable among different country or time samples. For example, if one country codifies education in a way which is unique and distinct from the way other countries do, IPUMS translate that unique coding into a code that is translatable to other countries. Also, it enables cross-time comparison within the same country creating a code that is valid for all time periods despite the changes occurred in the coding methods at certain time periods.

Let´s see for example what happens with education if we want to compare it between Albania and Argentina, where we found at EDATTAIN a harmonized version of educational attainment for both countries and the rest of the 80 countries at IPUMS.

Clemson Center for Geospatial Technologies

4. Why is it useful for research on Health, Sociology, Economics, Demography, Geography and Environmental Studies?

Several research questions demanding international comparison on a broad diversity of socio-economic indicators could be addressed using IPUMS.

Let's take for example scholars at the Department of Parks, Recreation and Tourism Management, which might be interested in studying sedentary behaviors. Now just have a look at some of the papers published under the search "SEDENTARY" that have used IPUMS data:

- Boehmer,Tegan K.; Brownson,Ross C.; Luke,Douglas A.
  2005. "**Declining Rates of Physical Activity in the United States**: What are the Contributors?" *Annual Review of Public Health* 26: 421-443

- Brownson,Ross C.; Boehmer,Tegan K.
  "**Patterns and Trends in Physical Activity, Occupation, Transportation, Land Use, and Sedentary Behaviors**." Report for the Transportation Research Board, 2004

- Cockburn,Myles; Mack,Thomas; Hamilton,Ann; Hawkins,Steve
  2004. "**Estimate of Physical Activity Prevalence in a Large Population-based Cohort.**" *Medicine & Science in Sports & Exercise* 36: 253-260

- Mathew Joseph, Nitha,; Bishop, Sheryl, L
  2014. "**Self-Reported Physical Activity, Sedentary Behavior and Body Mass Index Among US Asian Indian Women**." *Journal of Medical Research and Practice* 3: 63-70

- Pepin, Joanna, R; Sayer, Liana, C; Casper, Lynne, M
  2015. "Marital Status and **Mothers Time Use: Child Care, Housework, Leisure, and Sleep**." Presented at Population Association of America, San Diego, CA

- Vargas, Andres J.
  2011. "BMI, **Physical Activity and Sedentary Behaviors Among Mexican Immigrants** to the US: A Time Use Perspective." Texas Tech University Working Paper Series

Figure 2 . IPUMS Bibliography – papers based on IPUMS Data

Clemson Center for Geospatial Technologies

These are just some examples taken from the IPUMS repository of papers published using IPUMS. There you may find an extensive collection of papers addressing the most diverse research questions. The way to use this Bibliographic section is very easy.

The users constantly feed this section of IPUMS since one of the terms of agreements of using IPUMS is to cite the source and to report IPUMS once you have published. This has become a powerful literature browser, such as scholar google of others, so do not hesitate in including it in your research toolbox.

Take the example of someone interested in finding publications on "migration and immigration" published on "2014" as "Journal Article" using data from "any IPUMS projects" (Figure 2).[1]

---

[1] You can select a specific project used for the data of the bibliography you are looking for (**Figure 2**).

Clemson Center for Geospatial Technologies

5. How to use IPUMS?

There are two ways of using IPUMS INTERNATIONAL:

- use the online tabulator
- build your own extract of microdata

However, before going any further…

1.1. Log in at www.ipums.org
1.2. Select the IPUMS project you will be working-on. For today we will pick up IPUMS International

A - Exercise – Using IPUMS on-line tabulator

Estimate the share of the foreign-born population by sex, living in every Latin American country 2010 census waves.

> **1- Let's have a quick consider the online tabulator at**
> **https://international.ipums.org/international/sda.shtml**

You can analyze Single-Sample Datasets or Multi-sample/ Select ALL LATIN AMERICAN SAMPLES.

2. **Tabulate on-line** in response to exercise just described. For this exercise, please, select **nativity** from Demographics into a ROW, country sample from GLOBAL GEOGRAPHY in COLUMNS, **year** 2010-11 as FILTER, CONTROL by **sex**, select if you want PERCENTAGES by color row, and check for WEIGHTS. Run table.

Figure 3. IPUMS Tabulator

Figure 4. Outcomes by sex for Latin American census 2010-2011

Now, why don't you suggest a question we may answer using the on-line tab?

B - Exercise – Using IPUMS for extract creation

Estimate the share of women 20-29 that are not in a union for the whole population and those most educated in the **Dominican Republic, Costa Rica, Puerto Rico, and Panama at 1st sub – national level in 2010 wave**

1. **Select the option CREATE AND AN EXTRACT,** and the following dashboard would be displayed (Figure 5).

   Figure 5.Getting started loading your cart

From here you will start loading first, the samples variables and second, the variables.

Figure 6. Samples selection

NOTE: This and the following images do not necessarily reflect the ones for the specific exercise of PA, CR, RD, PR. Take it as an example for how to select samples.

Once you have selected the country-period samples from the Americas shown above, check the option INTEGRATED VARIABLES (Figure 7) is marked, and proceed to select three variables from the group DEMOGRAPHICS: age5, sex, and marital status. After you selected, you will see that your ¨shopping cart¨ starts getting loaded with variables and samples.

Clemson Center for Geospatial Technologies

Figure 7. Select variables from INTEGRATED variables (IPUMS)

Figure 8. You can also select from UNHARMONIZED variables (original)

Don´t miss the distinction between harmonized and un-harmonized variables. In the future if you are working with a single country -not oriented to international comparison- you might be interested in using the original country variables. Harmonization has the pro of comparison but the cons of losing precious information.

For example, Brazil is one of the few countries that include income at the census, or Uruguay asks for same – sex couple. Every country has its uniqueness, and IPUMS takes that into account.

Today we will use **INTEGRATED VARIABLES as in Figure 7**

Clemson Center for Geospatial Technologies

If you click on the variable, you may see the codes description, full definition and even a view of the cases count (following image).

Figure 9. Selecting the variable for EDATTAIN

Since we are interested in mapping our results, we need to add geographical data, by selecting HOUSEHOLD/GEOGRAPHY GLOBAL/ and the GEOLEVEL 1 VARIABLES FOR 1ST SUB-NATIONAL LEVEL for selected countries and years.

Figure 10. Add geo variables – recall we have already downloaded GIS shape files in Figure 1

Clemson Center for Geospatial Technologies

We make sure to have the ID variables and the WEIGHTS (Figure 11 AND Figure 12) for person and household (recall these are 1-10% samples of the original census). Moreover, once that is added we check out by clicking VIEW CART and click on CREATE DATA EXTRACT.

Figure 11. FROM SECTION HOUSEHOLD ADD TECHNICAL Variables

Figure 12. FROM SECTION PERSON ADD TECHNICAL Variables

2. Submit extract and prepare to download the data

Finally, we name the extract and customize or filter cases in case we are interested in, and we SUBMIT EXTRACT Figure 13. However, here you can select a particular group of cases, attach additional characteristics and reduce the size of the file.

Figure 13. Finish and submit extract

Clemson Center for Geospatial Technologies

Once the extract is sent, you will immediately have access to the syntax for SAS STATA SPSS that enables opening the extract (txt FORMAT). <u>Please select the syntax for SPSS for today´s workshop. A few minutes later an e-mail will let you know that the zip data file is ready to be download and you will find it right below where it states ¨Data¨</u>

Figure 14 – Download the data

Also check you´ll always be able to revise and resubmit your extract from the past.

3. **Now open SPSS to visualize the microdata for your extract**

Go to File/Open/Syntax, and get the SPSS syntax downloaded from your IPUMS extract **"COURSE_12APRIL2017".** You will have to add the path to the data file into the first line of the code letting know SPPS that the data should be open from that directory and that it should open it by using all the labels and values coded below (Figure 15). Remember to unzip the data files downloaded from IPUMS. **In this case, the .txt data is actually in file 'ipumsi_00064.dat'.**

Figure 15.a – Open the data in Spss

This will take a while, while data is getting extracted.
Later check your e-mail, and you will see the data you have just ordered.

Clemson Center for Geospatial Technologies

Figure 15.b - Once the data is ready the microdata file would look like this

Now we will proceed to the construction of aggregated variables to map and visualize IPUMS data. Estimate and plot the share of female 20-29 living in a union by subnational level 1 in the Americas, first for 2000, second for 2010

4. Create a new syntax file to preliminary steps for aggregation of data

There we will be coding to create the set of new dummy variables that will be useful to work with rates to be plotted in maps

Figure 16 –Create a new syntax file

On this new file, we will be copy-paste the following code (only sections within orange border lines).

Clemson Center for Geospatial Technologies

5. Creation of the first set of dummy variables

*Creation of variable female where women will take value 1 (former 2), and males value 0 (previous 1). The SYSMIS in SPSS stands for coding as missing values ".".
This and the following variables will be used in exercise 1.

```
RECODE SEX (1=0) (2=1) (9=SYSMIS) INTO fem.
VARIABLE LABELS fem 'female'.
EXECUTE.
```

*Creation of an age group 20-29 as dummy variable. Since we have selected AGE in groups we have to use the labels for the ages while re-codifying into this new dummy AGE2029 variable.

```
RECODE AGE2 (0 thru 7=0) (8 thru 9=1) (10 thru 21=0) (99=sysmis) INTO age2029.
VARIABLE LABELS age2029 'age2029'.
EXECUTE.
```

*Creation of a dummy for identifying the women that are or haven't ever been in any union (marriage or cohabitation). This is the key variable for Exercise 1: Have cohabitation been expanding from the 2000s to 2010s? The increase of cohabitation among female has been interpreted as one of the signs for the Second Demographic
Transition, which involves: more divorces, more cohabitation, fewer marriages, lower and later fertility, among other indicators of family change.

```
RECODE MARST (9=SYSMIS) (0=SYSMIS) (1=1) (2 thru 4=0) INTO nonunion.
VARIABLE LABELS nonunion 'never union'.
EXECUTE.
```

*Creation of a dummy for identifying the highly-educated people
We are creating this variable to cross it in the future with non-in-union women, to explore if the speed of the cohabitation expansion in the Americas was larger among the most educated women than for the total female population aged 20-29.

```
RECODE EDATTAIN (9=SYSMIS) (0=SYSMIS) (4=1) (1 thru 3=0) INTO university.
VARIABLE LABELS university 'university completed'.
EXECUTE.
```

6. The importance of weights

Although IPUMS International provides data on CENSUS, we are working with 1%, 5% or 10% samples. Therefore, we need to weight the data. Weights in IPUMS are identified with suffix WT and work for both weighting and expanding. The command for weighting in SPSS follows. Recall using it for every frequency tabulation and check weights are on before aggregating data.

```
*weight the dataset.
WEIGHT BY PERWT.
```

Clemson Center for Geospatial Technologies

*Asking for a crosstab.

```
CROSSTABS
 /TABLES=university BY COUNTRY
 /FORMAT=AVALUE TABLES
 /CELLS=COUNT
 /COUNT ROUND CELL.
```

Compare outcomes after weight is on and with weight off. Recall that SPPS shows if the data is weighted or not at the bottom right of the data view browser.

university completed * Country Crosstabulation – UNWEIGHTED
Count

			Country			
		Costa Rica	Dominican Republic	Panama	Puerto Rico	Total
university completed	.00	342361	833053	279992	29233	1484639
	1.00	53831	55497	31270	5760	146358
Total		396192	888550	311262	34993	1630997

university completed * Country Crosstabulation – WEIGHTED

Count

			Country			
		Costa Rica	Dominican Republic	Panama	Puerto Rico	Total
university completed	.00	3423610	8330530	2799920	3026041	17580101
	1.00	538310	554970	312700	566779	1972759
Total		3961920	8885500	3112620	3592820	19552860

Clemson Center for Geospatial Technologies

7. Creation of the second set of variables by combining the recently created dummy variables

In this way, we will be preparing the numerator and denominator of our future shares of singlehood and immigration.

---

*numerator.
COMPUTE nonunion_female_2029=fem = 1 & age2029 = 1 & nonunion = 1.
VARIABLE LABELS nonunion_female_2029 'Female2029_NonUnion'.
EXECUTE.

---

*denominator.
COMPUTE female_2029=fem = 1 & age2029 = 1.
VARIABLE LABELS female_2029 'Female2029'.
EXECUTE.

---

*numerator by education.
COMPUTE nonunion_female_2029_univ=fem = 1 & age2029 = 1 & university = 1 & nonunion = 1.
VARIABLE LABELS nonunion_female_2029_univ 'Female2029university'.
EXECUTE.

---

*denominator by education.
COMPUTE female_2029_univ=fem = 1 & age2029 = 1 & university = 1.
VARIABLE LABELS female_2029_univ 'Female2029university'.
EXECUTE.

---

8. Transform data from long to wide / Aggregate command in SPSS

This will be the most time-consuming part of your work if you are working with SPSS.
At this point, we have data in long format, but we would need to transform it into wide to respond our research questions (share of female with certain characteristics and proportion of immigrants by country, sub-national level, and year). Also, most of the GIS analysis is based on this format.

Figure 17. Just an example. From long microdata to wide aggregated by date

Therefore, the transformation from microdata to aggregate data (long to wide) is necessary. The idea is that you have several rows for each subject (country/county/state) and you want to transform these observations into a single line. In SPSS this process is called aggregate.

Clemson Center for Geospatial Technologies

5.1. Let's go to DATA / AGGREGATE

Once this window appears, enter the variables that will be your new units of analysis into the **"Break Variable"** box, where you are expected to add the variables you will aggregate by. For the purpose of this exercise include YEAR, COUNTRY and GEOLEVEL1.

Then define your **"Summaries of Variable"** Our summary variables are in this case we will use This mean we will end up with a file with as many rows as years x country x geolevel1, where the variables would be the SUM of our selected indicators created from microdata. Remember you can always change the default summary operation on "Function" (MIN, MAX, MEAN, DEV, SUM, COUNT, etc.) according to your needs and the characteristics of the variables.

Figure 18. These are the steps to aggregate in SPSS using the menu

Figure 19. And this is the way your wide-data will look like after aggregating

Compare these rows to the ones we had at the beginning in Figure 16.a.

In case you rather to work on syntax, here is what you should be typing.

Clemson Center for Geospatial Technologies

```
*SYNTAX FOR aggregate by year, country, 1st subnational level.
        DATASET DECLARE aggregate_wkshopIPUMS.
        SORT CASES BY COUNTRY GEOLEV1 YEAR.
        AGGREGATE
         /OUTFILE='aggregate_wkshopIPUMS'
         /PRESORTED
         /BREAK=COUNTRY GEOLEV1 YEAR
         /fem_sum=SUM(fem)
         /age2029_sum=SUM(age2029)
         /nonunion_sum=SUM(nonunion)
         /university_sum=SUM(university)
         /foreign_sum=SUM(foreign)
         /female_2029_univ_sum=SUM(female_2029_univ)
         /nonunion_female_2029_univ_sum=SUM(nonunion_female_2029_univ)
         /female_2029_sum=SUM(female_2029)
         /nonunion_female_2029_sum=SUM(nonunion_female_2029)
         /native_2029_sum_1=SUM(native_2029)
 /foreigborn2029_sum=SUM(foreigborn2029).
```
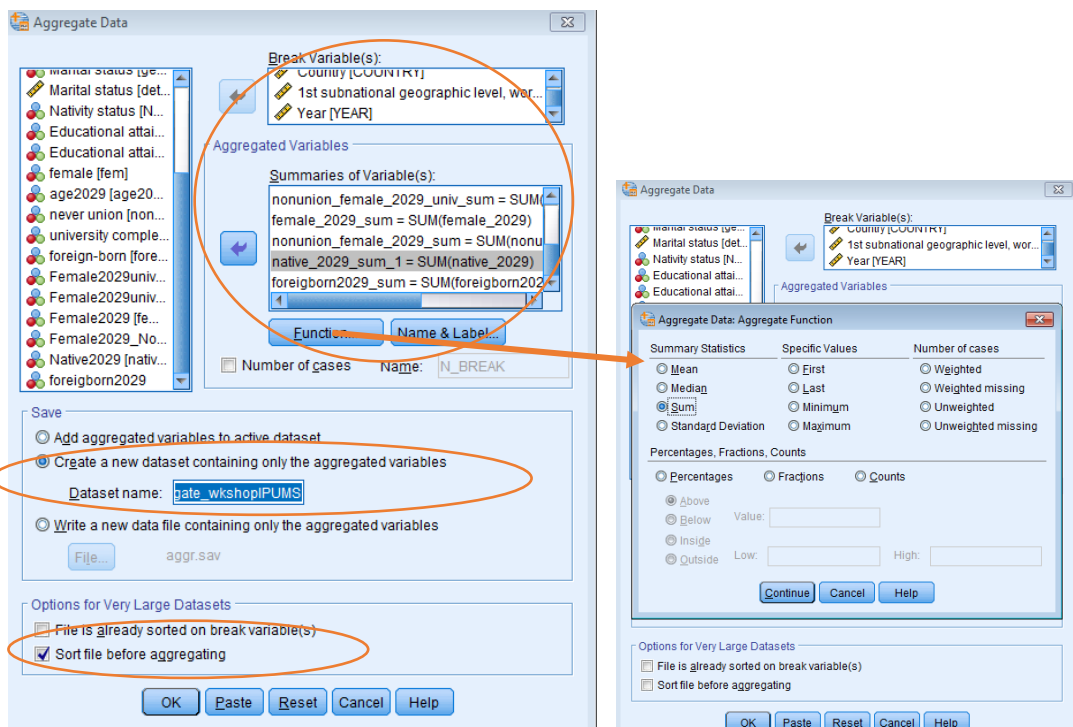
9. Create the rates we will be using for visualization later

*share of women 20-29 that are not in a union

```
        DATASET ACTIVATE aggregate_wkshopIPUMS.
        COMPUTE nonunionrate=nonunion_female_2029_sum / female_2029_sum.
        EXECUTE.
```
*share of highly educated women 20-29 that are not in a union

```
        COMPUTE          nonunionrate_univ=nonunion_female_2029_univ_sum          /
        female_2029_univ_sum.
        EXECUTE.
```
*share of immigration in total population

```
        COMPUTE   nonunionrate=foreigborn2029_sum    /    (native_2029_sum_1    +
foreigborn2029_sum).
        EXECUTE.
```

*For some playground on SPSS using aggregated data.

*some playground on tabulating visualizing aggregated data in STATA

```
        SORT CASES  BY COUNTRY.
        SPLIT FILE LAYERED BY COUNTRY.
        MEANS TABLES=nonunionrate_univ BY YEAR
         /CELLS=MEAN COUNT STDDEV.
SPLIT FILE OFF.
```

Just a quick comparison of means by year and country for the indicator non-union rate created

for female 20-29.

Report

nonunionrate

Country	Year	Mean	N	Std. Deviation
Costa Rica	2011	46.4812	7	9.29016
	Total	46.4812	7	9.29016
Dominican Republic	2010	22.3959	25	5.02065
	Total	22.3959	25	5.02065
Panama	2010	29.6261	7	7.52232
	Total	29.6261	7	7.52232
Puerto Rico	2010	77.5835	6	5.11122
	Total	77.5835	6	5.11122

10. **You can visualize sub-national level data in many GIS software. Today we will try Tableu**

First, save the data in SPSS to keep the work we have done. Second, go to File/Export and save the data into EXCEL file. In this way, we will have an Excel version of the data aggregated in SPSS. That data will work as our Excel (statistical information) for Tableau.

Now, open Tableau.

Then, in ADD A CONNECTION click on SPATIAL FILE (Figure 21). We will be bringing the shape file containing GEOLEVEL1 variable (IPUMS identificatory for the 1st sub-national level). Remember that in Figure 1 we got the ¨world1.geolevel1.shp¨, when we downloaded from IPUMS GIS the ¨world_geolevel1¨ zipped folder. Now is the time to use that data: the .shp file contained in that folder.

Then you will see at the right top of the recently added file the label ¨Add¨, click on it and insert the Excel we exported from SPSS.

Figure 20. Loading geo and stats data in Tableau

Clemson Center for Geospatial Technologies
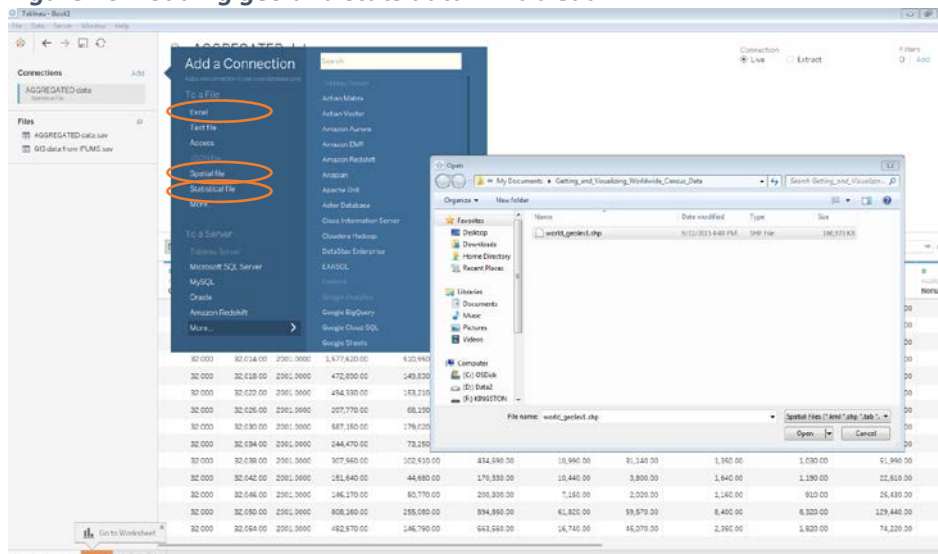
Then you will have in blue the display of geographical data and in orange a display of the statistical variables. However, there is one more step before getting both files matched. You have to select the geolevel1 variable in both files (that is the variable that will enable the match since it has the same ID for the 1st sub-national units). To get the joint done you need to see that the red exclamation mark on the top of the options ¨interior/left/right/full¨ is gone. To remove this, it will be necessary that you check that both geolevel variables, one in EXCEL FILE IMPORTED, other in GEOGRAPHY IMPORTED are in string format.

**Figure 21. Selecting the matching variable to do the joint**
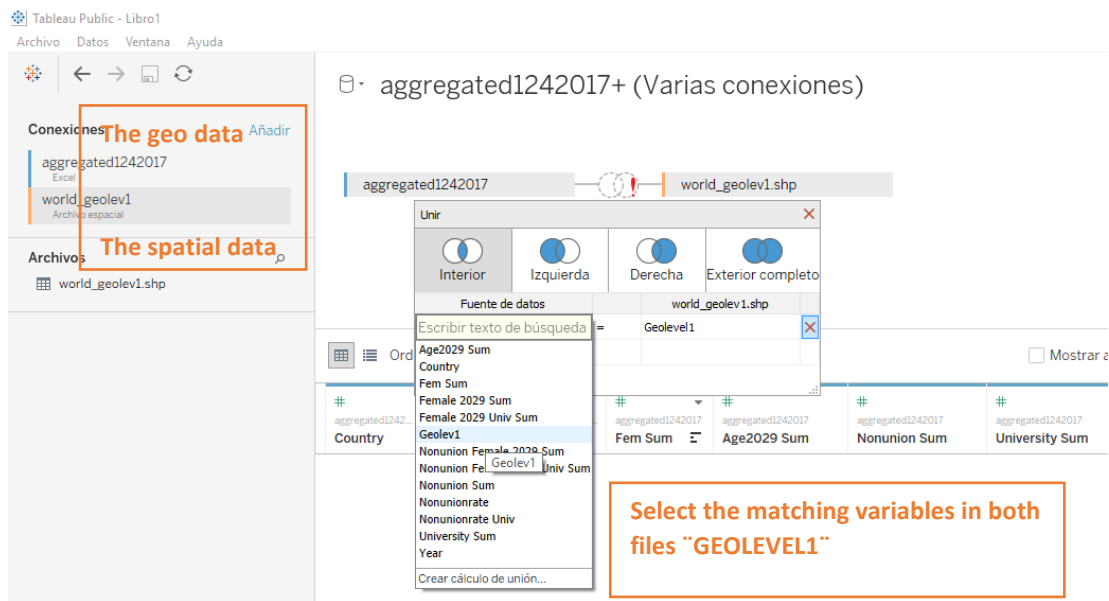
**Figure 22. Loading geo and stats data in Tableau**
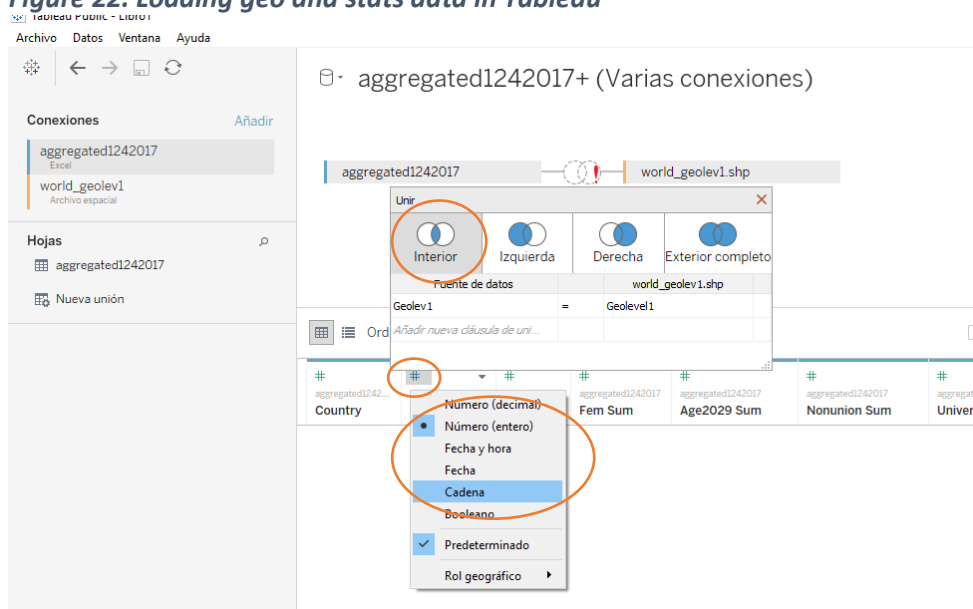
Click on the numeral sign # on left top of the label for "geolevel1" at the section for data ¨aggregate12042017¨, and change numeric into a string. Immediately after, you will get the match done. Select interior or inner match.

Clemson Center for Geospatial Technologies

Figure 23. Final appearance of the data
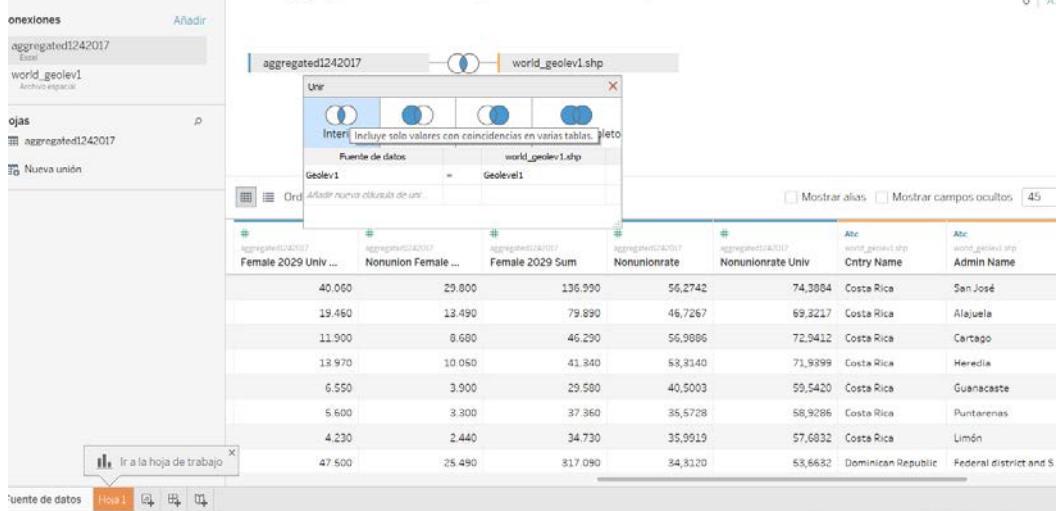
In case you want to see a map, select a new worksheet, and grab the geography towards the central area of the worksheet

*Figure 24. Step 1 to create a map*
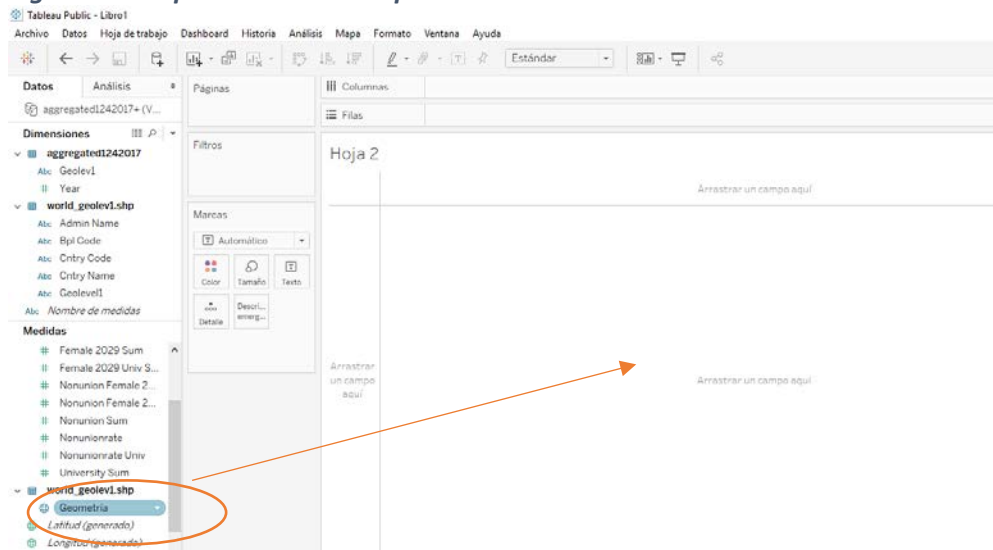
This is what you will get, but to build the map representing your data you will need to add the selected indicator into the area where a set of colorful circles are shown.

Clemson Center for Geospatial Technologies
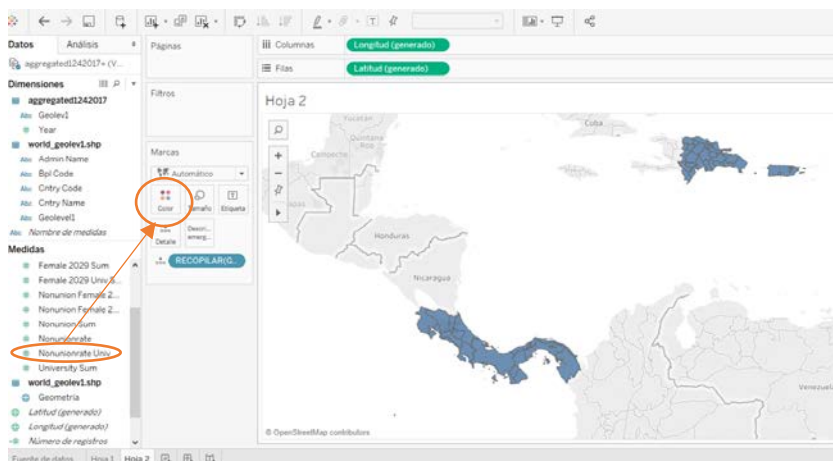
Finally, you just need to select the function of the selected variable and switch it from SUM to DIMENSION, and a hatch map with gradient colors according to the rate intensity will appear The legend will be at the top right side of the map.

Figure 25. Visualization of non-union rate among 20-29 female any educational attainment
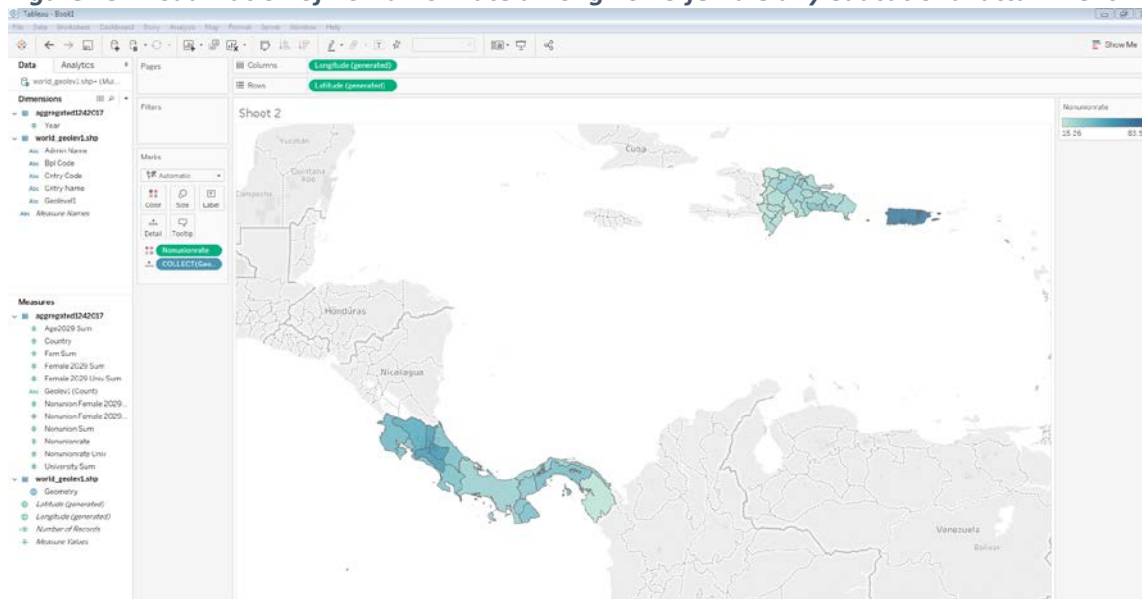
Figure 26. Visualization of non-union rate among 20-29 female highly educated

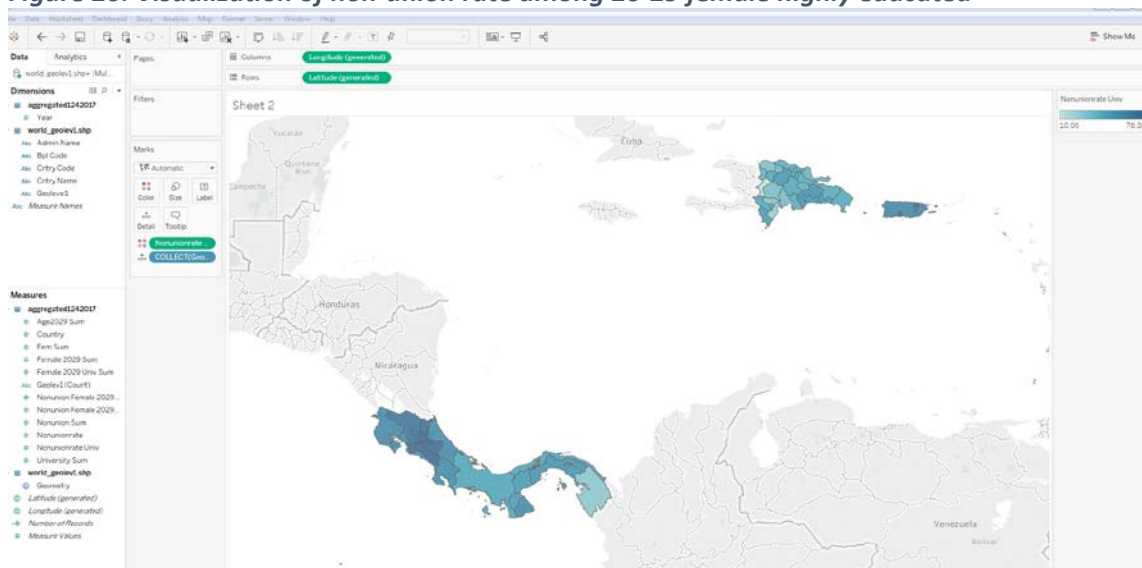Clemson Center for Geospatial Technologies

Figure 27. Tableau also enables visualization by bars. In this case the median by country of the non-union rate among 20-29 female. This median is estimated from the 1st country sub-national level values

Clemson Center for Geospatial Technologies