Populations of concern in Europe. The European refugee crisis in numbers

Debora Ferreira, Nadina Iacob

December 14, 2017

1. Purpose

This paper aims to use data available on the populations of concern (refugees, asylum seekers, stateless persons, etc.) around the word in order to better understand the refugee crisis experienced in Europe in 2015. We are particularly interested in analyzing the situation before and during this period, to better comprehend the magnitude of the crisis and the characteristics of the populations of concern spread among the European countries. We will also perform comparisons between Europe and other continents to evaluate if this is a regional phenomenon or a global experience. Lastly, we intend to assess some popular statements on refugees, checking if they hold against the analyzed data or not.

2. Descriptive Statistics

2.1. The data frames

This project was developed using data from the UNHCR, the UN Refugee Agency, on migration trends. The population statistics made available by this office include 6 datasets, which are described as following by the institution[1]:

Data frame	Description
pers_concern	Each row of data represents the information about UNHCR's populations of concern for a given year and country of residence and/or origin. In the 2016 data, figures between 1 and 4 have been replaced with an asterisk (*). These represent situations where the figures are being kept confidential to protect the anonymity of individuals. Such figures are not included in any totals.
time_series	Each row of data represents the information about UNHCR's populations of concern for a given year and country of residence and/or origin. In the 2016 data, figures between 1 and 4 have been replaced with an asterisk (*). These represent situations where the figures are being kept confidential to protect the anonymity of individuals. Such figures are not included in any totals.
demographics	Information about persons of concern broken down by sex and age, as well as by location within the country of residence (where such information is available). Note that data broken down in this way is not always available, so it may not be possible to reconcile the figures on this page with those on the Persons of Concern and Time Series pages. Such data is available since 2000.
asylum_all	Information about asylum applications in a given year and the progress of asylum-seekers through the refugee status determination process. Such data is available since 2000.

[1] information obtanined at http://popstats.unhcr.org/en/overview

Data frame	Description
asylum_month	Information about asylum applications lodged in 38 European and 6 non-European countries. Data are broken down by month and origin. Where possible, figures exclude repeat/re-opened asylum applications and applications lodged on appeal or with courts. For some countries, the monthly data are available since 1999 while for others at a later period. In the 2016 and 2017 data, figures between 1 and 4 have been replaced with an asterisk (*). These represent situations where figures are being kept confidential to protect the anonymity of individuals. Such figures are not included in any totals. Due to retroactive adjustments implemented by States, totals in this dataset may differ from annual totals published by the competent national authorities.
ressettlement	Information on resettlement arrivals of refugees, with or without UNHCR assistance. This dataset is based on Government statistics and, in principle, excludes humanitarian admissions. In the 2016 data, figures between 1 and 4 have been replaced with an asterisk (*). These represent situations where the figures are being kept confidential to protect the anonymity of individuals. Such figures are not included in any totals.

These datasets were downloaded and opened in the same database for the beginning of the project, already excluding the missing values informed by the UNCHR (*). It was also necessary to tidy the data from "demographics", as the age categories per gender were distributed in different columns and, according to tidy data principles, each variable must form a column. The tidy data frame was named "demo_tidy".[2]

After analyzing the datasets, it was possible to see that "pers_concern" and "time_series" contained equal information; thus, only "time_series" was kept for the present analysis, as it was the tidiest dataset. It was also possible to observe that the "resettlement" dataset could be merged with "time_series", by adding the resettled persons as a new category of populations of concern (variable "Population.type") in the latter dataset. The new tidy data frame was named "pop_time". Once the datasets were merged, the variables' names were unified to easy the future work and understanding.[3]

A similar operation could not, however, be carried out with the "demo_tidy" data frame. The data frames "pop_time" and the "demo_tidy" do not have identification numbers for the respondents/observations and include different key variables. More specifically, "pop_time" contains the types of populations of concern by country of origin and country of residence, whereas "demo_tidy" does not contain a breakdown of its values by country of origin or type of population of concern, only providing information about the sex and age for the sum of the populations of concern located in a given country of residence, regardless of type or origin. Therefore, merging these data frames would result in losing information.

As the present analysis intends to investigate the evolution of populations of concern over time, also considering the relevance of the demographic information for it, this work will focus on the data frame containing the tidy demographic data ("demo_Tidy") and on the merged time series data frame ("pop_time"). Firstly, "pop_time" will be used to draw conclusions about the populations of concern and their preferred destinations over time. Secondly, "demo_tidy" will help to establish the demographic composition of the overall populations of concern in their different places of residence. At principle, the status of the asylum requests in total and per months were not considered relevant, hence the disregard of the other datasets available.

2.2. Key variables of interest created/modified

After the data frames of interest were treated, having their missing values recoded, it was necessary to create a new variable in each of them that would reflect the continents in which the populations of concern were living (i.e. the continents of destiny of the migrants). This new variable, called "continent_res", is fundamentallly

[2]For a more detailed explanation on the tidying process, please see Apendix I.
[3]For a more detailed explanation on the variables used, please see Apendix II.

related to the purpose of the paper of analyzing an event (migrant crisis) that was reportedly more expressive in Europe, and also extensively debated in the media and politics of this continent.

Lastly, it was noticed that the UNHCR provides inconsistent information over time with regards to the coding of the age groups ("demo_tidy" data frame). From 2001 to 2005, the age group "5-17" was used, whereas from 2006 onwards, this group was split into two: "5-11" and "12-17" years old. As we deemed that, for the purposes of this paper, it would not be necessary to have demographic information on preteens and teenagers separately, the two split categories were unified into "5-17" years old for the entire the data frame. This way, a time series analysis could be conducted for a longer period o time, without discontinuity, which is relevant considering that the time range of this data frame (2001-2016) is already much smaller than the one in "pop_time" (1951-2016). There are also cases where no age was recorded.

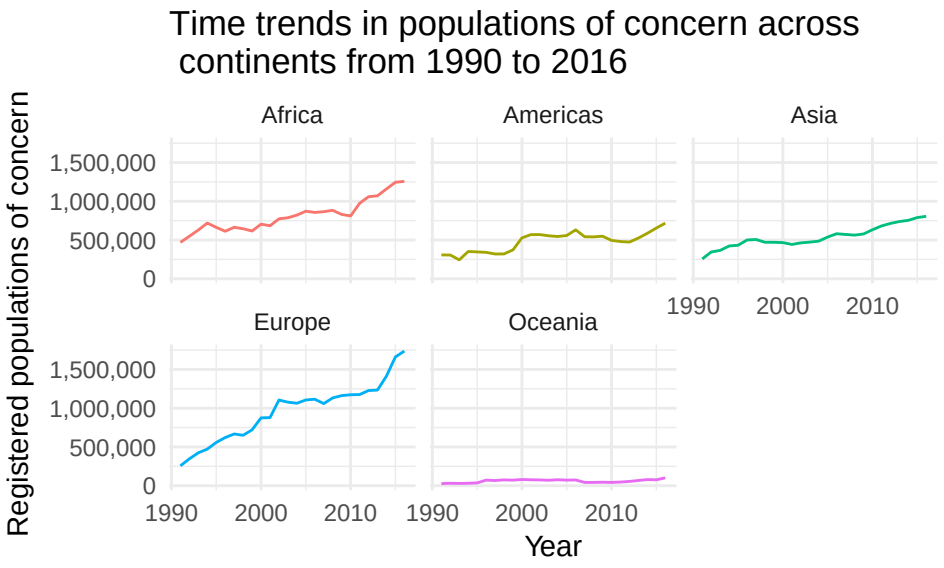
3. Methods

As observed, the UNHCR datasets are very limited, there are not enough controlling variables to run a plausible regression model with statistically significant results. Besides, there is relevant information distributed across two data frames which cannot be merged and, despite the amount of observations collected, they are not organized as panel data (there is not even an identification number for each observation). In this context, the only option left was conducting a time series analyses of the UNHCR data.

On a brief explanation, *"Time series analysis is the analysis of a series of data-points over time, allowing one to answer question such as what is the causal effect on a variable Y of a change in variable X over time?"* This will be explored in the following session of the paper.[4]

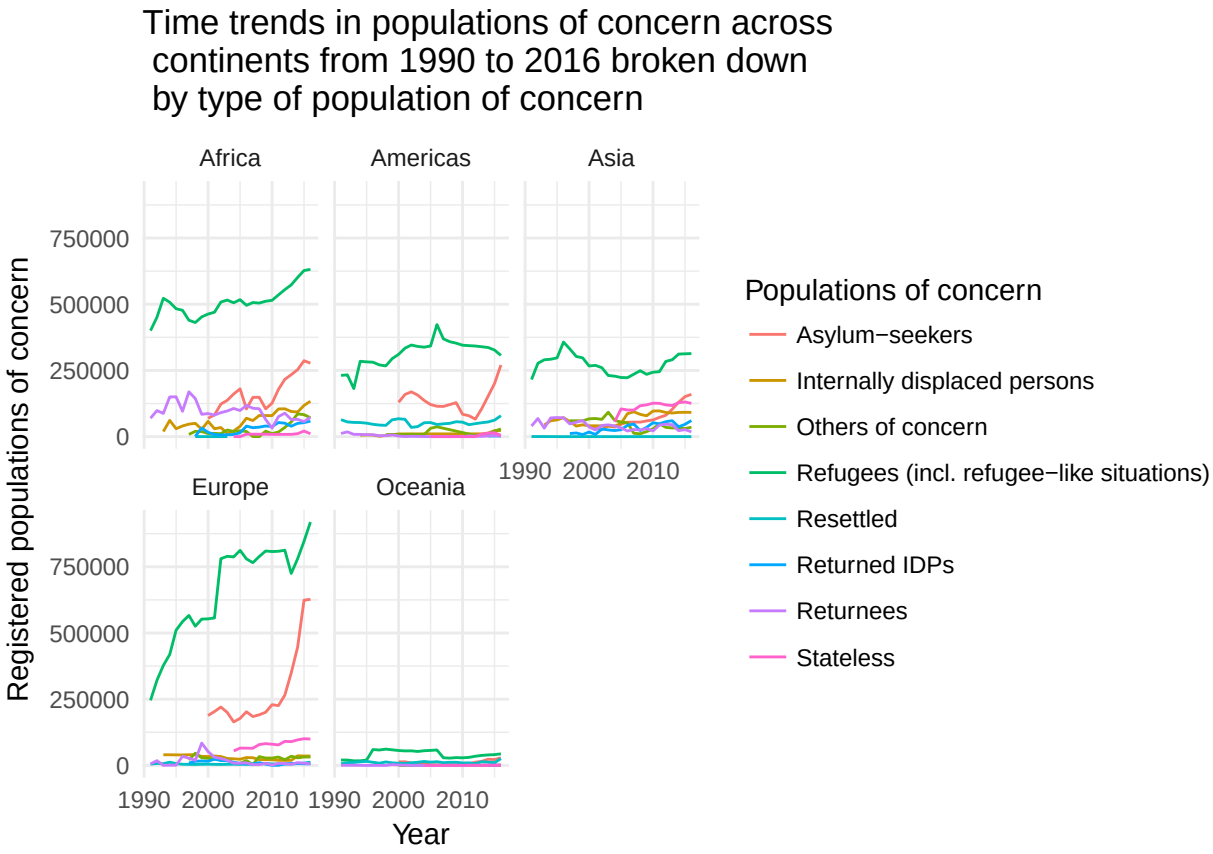
4. Analysis

We started the analysis by taking a global view at the time trends in populations of concern across continents. This resulted in a comparative view of the overall populations of concern situation over the past 30 years. The following graph shows that, in all continents, there has been a steady increase in the number of populations of concern. However, Europe experienced by far the steepest increase, especially during the 2000-2015 period, with the populations of concern reaching over 20 million people at aggregate level in the continent.

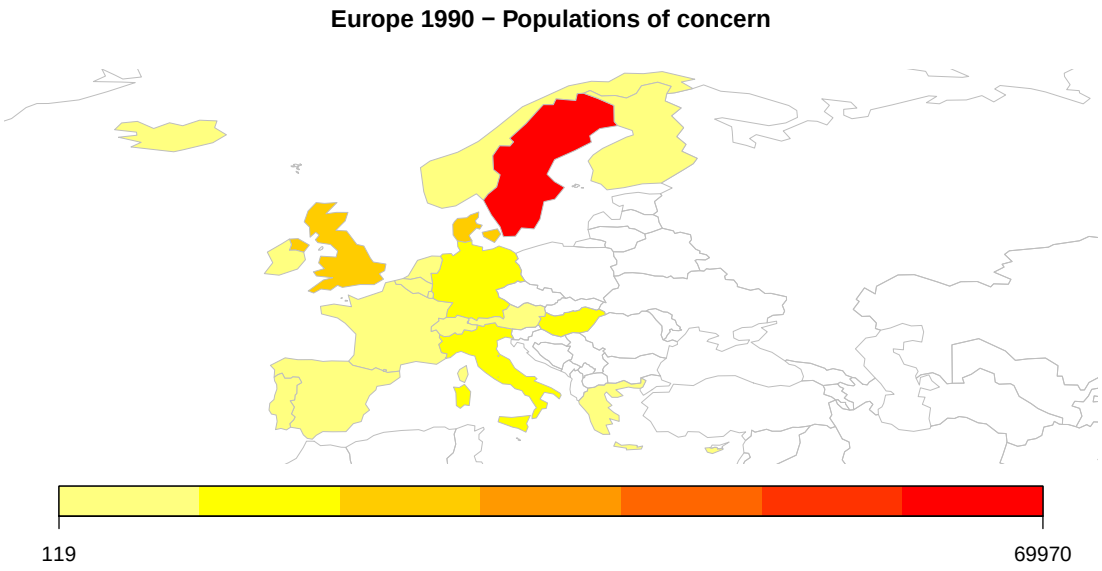


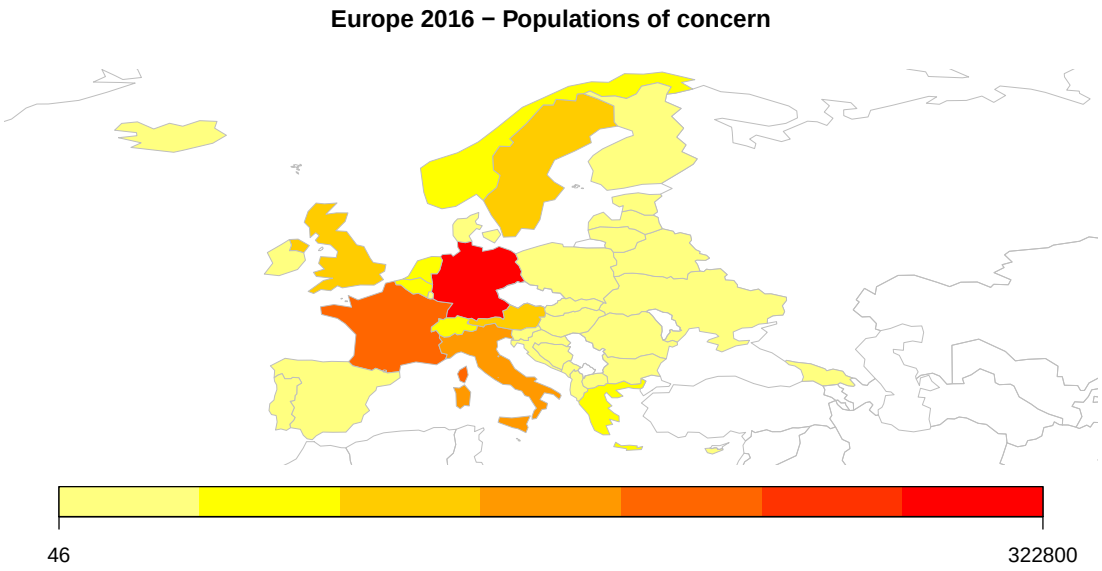
[4]Imdadullah (2013). Time Series Analysis and Forecasting. *In: Basic Statistics and Data Analysis.* Available at: http://itfeature.com/time-series-analysis-and-forecasting/time-series-analysis-forecasting. Retrived 14.12.2017.

In a next step, this increase was disentangled by the types of populations of concern in the last three decades; the idea was to see which type(s) of population drove this increase the most. We saw that refugees and asylum_seekers were the major populations of concern across all continents and much more preponderant over the other populations of concern when it is about Europe. While the other population types remained quite constant over time, we can affirm that the number of refugees and asylum_seekers was the cause of the increase in the number of persons of concern.



We then proceed narrowing down the analysis to Europe, now seeking to relate the distribution of populations of concern across countries. The maps below not only reaffirm the refugee crisis in the Continent (whose populations of concern doubled in the 2000's and from then on steadily increased), but also shows a spread in the destination of these migrants.





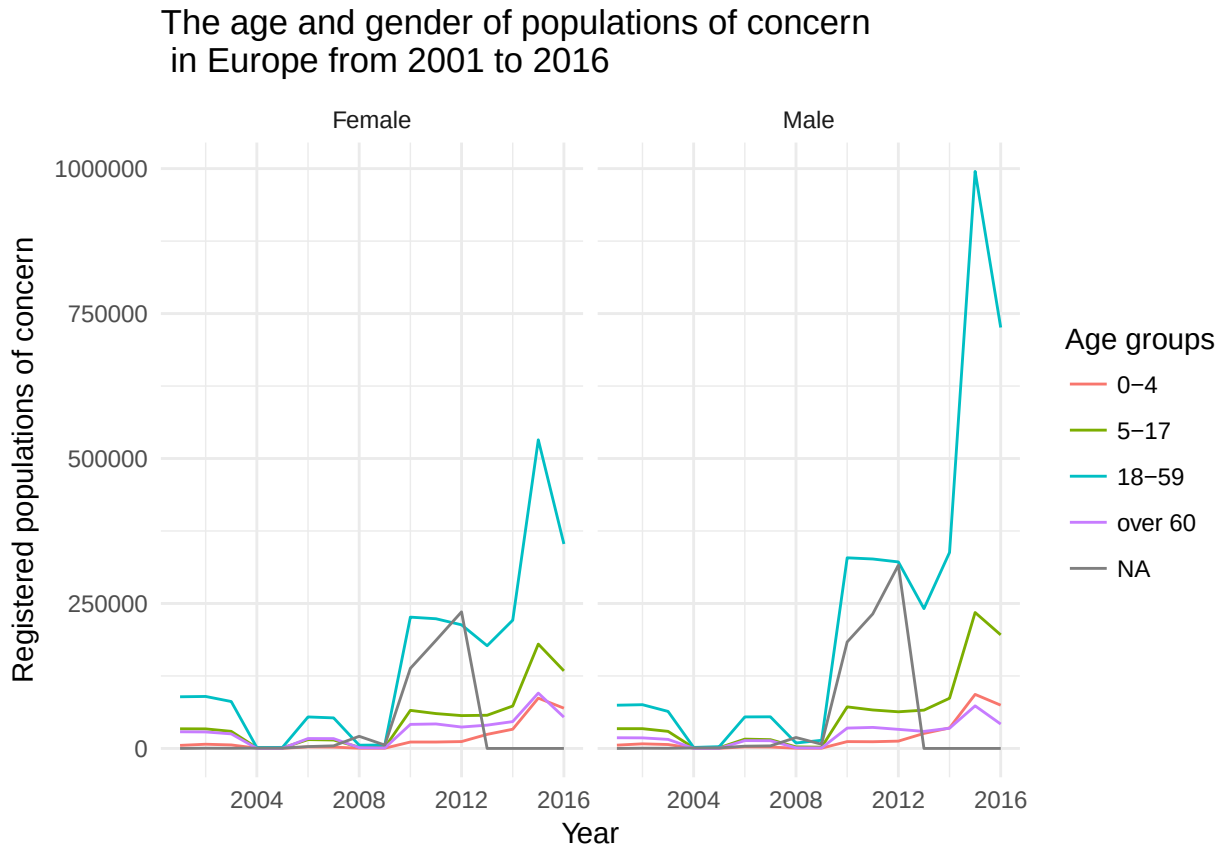
About 15 years ago, Italy, Sweden and the UK were outstanding as the top migrants' receivers. Recently, France, Germany and the Netherlands were added to the countries that most receive refugees; with many others, such as Belgium, Ireland, Switzerland and Norway being added to the list of European countries that have received over 1 million persons of concern during the last 15 years. Austria is approaching this numbers and Spain, Ireland and Norway were also noticeably affected by this influx of people.

The time trend perceived here is that the countries which were the biggest receivers of persons of concern 15 years ago keep being so, the reasons for that are not clear by the data though (they could very among countries' policies in relation to refugees, countries' economic situation, etc.). The geographical position, however, is clearly not the (only) explainer, since Italy and UK, for example, are far away from each other and keep being the top receivers of persons of concern.

Our last analysis consisted in verifying if there was any demographic difference among the populations of

concern in Europe. Graphing the populations of concern over time per gender and age, it was possible to see that, aside the groups over 60 years old, the persons of concern are more male than female, with the proportions keeping quite constant over time. Among the different age groups, the 18-59 years old is the one that outstandingly included the highest number of persons of concern in all the years analyzed.

Although we can say that refugees are more likely to be between 18 and 59 years old and male, we cannot affirm that being male in this age group leads to higher chances of being a refugee (that would depend on analyzing other variables that are not available in the UNHCR datasets). Nevertheless, as this is a time-trend that stayed constant, there is hint for similar future.



5. In conclusion, which "refugee myths" can we tackle down?

"This is a migrant crisis, not a refugee crisis" (also in the version "They're not really refugees, but just economic migrants seeking a better life")

Myth. For all the continents analyzed, the refugees were the biggest type of populations of concern - populations which, as shown above, began to grow considerably in the 2000's and had a steady peak in 2015 - meaning that there is indeed a refugee crisis.

Nevertheless, the second biggest group of populations of concern in all continents were the asylum seekers, those persons who are still waiting to be granted with a refugee status. As these persons might not be considered refugees by the UNHCR after the waiting process, it is unsure whether they will be labeled as refugees or not and, if not, on which grounds (it is possible that they are seen economic migrants, but there can be other reasons, like not observing the bureaucracy involved in the process). As an official measure of accepted requests, the European Commission reported that ca. 60% of the asylum requests were granted

in 2016[5]. Keeping in mind the other 40% and the lack of information on the datasets about the reasons of denial, parsimony is needed in assessing this statement. However, these numbers vary tremendously, and this can be seen in the case of Germany, in the table below tracking the evolution of the accepted applications as a share of the total decisions taken in a given year.

year	accepted_ratio
2011	0.1636917
2012	0.1417527
2013	0.1347897
2014	0.2583953
2015	0.4015603
2016	0.3438528

"Europe is hostile with refugees"

Myth. At least in number of refugees received, the continent is by far the one that receive most refugees over time, achieving 2 billion populations of concern in 2015, which is more than the populations of entire small countries, such as Estonia and Kosovo. Besides, the low numbers of returnees in Europe (lower than in the Americas and Asia), demonstrates that, in fact, a large proportion of refugees that seek the continent are allowed to stay there permanently.

"The refugees will steal our jobs" (also in the version "They're mostly men in working age, who would be capable to survive in their home countries")

Partially true. The data analyzed states undeniably that the refugees are mostly men in working age (between 18 and 59 years old). However, this does not automatically imply that the reason why they are displaced is to seek work abroad, neither that the conditions in their origin's country would allow surviving and minimum human life stands. Furthermore, there are many other factors to be considered in judging this statement, as for example the capacitation and education levels of the refugees (data which we do not have); the existence of laws in the country of residence that favor or allow their ingress in the job market (and the contrary); etc. Regarding the data analyzed, affirmations can be done solely about the demographic composition of the refugee population (as seen in the section above), no further assessments can be done.

Final note: For more detailed time series and demographics trends, explore out interactive dashboard available in the same repository.

[5]Asylum statistics. Eurostat. Available at: http://ec.europa.eu/eurostat/statistics-explained/index.php/Asylum_statistics. Retrieved 14.12.2017.

References

Asylum statistics. Eurostat. Available at: http://ec.europa.eu/eurostat/statistics-explained/index.php/Asylum_statistics. Retrieved 14.12.2017.

Imdadullah (2013). Time Series Analysis and Forecasting. *In: Basic Statistics and Data Analysis.* Available at: http://itfeature.com/time-series-analysis-and-forecasting/time-series-analysis-forecasting. Retrived 14.12.2017.

Wickham (2014). Tidy Data. *In: Journal of Statistical Software.* Vol. 59, issue 10.

Thomson (2015). 4 myths about refugees - debunked. World Economic Forum. Available at: https://www.weforum.org/agenda/2015/11/4-myths-about-refugees-debunked/. Retrived 14.12.2017.

UNHCR. *Population Statistics*. Data and description available at: http://popstats.unhcr.org/en/overview. Retrived 14.12.2017.

RStudio Blog. Website: https://blog.rstudio.com. Retrived 28.11.2017 - 14.12.2017.

Apendix I - Explanation on tyding process

The demographics dataset as downloaded from the UNHCR website does not follow the principles of a tidy dataset. More specifically, two variables, age and gender, are combined into multiple columns (e.g. Female.0.4, Female.5.11, Male.0.4 etc.), with values being headers. Therefore, we rearranged the information into a column for gender (with the variable values coded as "F" and "M", which we than replaced by "Female" and "Male" to match the rest of the data frame), and one for the age groups (variable name "age"), modifying at the same time the format of the values for better understanding ("0-4", "5-11", "5-17", "12-17", "18-59", "over 60", categories that were later unified and reduced to 5). The tidy dataframe is named "demo_tidy".

The "time_series" dataset provides information about the number of a specific type of population of concerned from a given country of origin located in a country of residence in a given year (from 1951 to 2016). The "resettlement" dataset provides similar information about the resettled populations. Given that the two datasets have the same variables (year, country of residence, country of origin, population type, and value), they can be joined vertically. The joined dataset is "pop_time", with "resettlement" being essentially transformed into additional observations in the "time_series dataset. This operation helps us have information about all populations of concern in one dataset.

Apendix II - Explanation on the variables

Variables in "pop_time"

Variable	Type	Description
year	Categorical/ continuous	Year when data was recorded, from 1951 to 2016
continent_res	Categorical	Continent where the persons of concern reside: Americas, Africa, Asia, Americas, Africa, Asia, Europe or Oceania
country_res	Categorical	Country where the persons of concern reside, includes 199 countries
origin	Categorical	Country or origin of the persons of concern, includes 226 countries
pop_type	Categorical	Types of populations of concern: Refugees (incl.refugee-like situations), Returnees, Asylum-seekers, Stateless, Internally displaced persons, Returned IDPs, Resettled or Others of concern.
value	Continuous	Number of persons of concern that fulfill the conditions in a row (e.g. the number of refugees from Ghana living in Italy in 1984 was 102)

Variables in "demo_tidy"

Variable	Type	Description
year	Categorical/ continuous	Year when data was recorded, from 2001 to 2016
continent_res	Categorical	Continent where the persons of concern reside: Americas, Africa, Asia, Americas, Africa, Asia, Europe or Oceania
country_res	Categorical	Country where the persons of concern reside, includes 191 countries
location_res	Categorical	Location (city, region, state, province, etc.) where the persons of "Various", etc.)
gender	Categorical	Gender of the persons of concern (Female or Male)
age	Categorical	Age group of the persons of concern: "0-4", "5-17", "18-59", "over 60"
value	Continuous	Number of persons of concern that fulfill the conditions in a row (e.g. the number of female persons of concern, between 0-4 years old,living in Luanda in 2001 was 28)