HarvardX Research: education composition world map data specification

The interactive visualization appearing at harvardx.harvard.edu/harvardx-insights/worldmap-gender is based on data on HarvardX enrollment for courses offered through the edX platform. Enrollment is defined as individuals on HarvardX course lists. This document aims to describe (1) the way data was prepared and its possible sources of error and (2) possible misinterpretations of the data.

Preparation and possible sources of error

The data was prepared by using IP addess geolocation of HarvardX registrants, as well as parsing self-reported mailing address provided at registration if IP address is unavailable. Data on the highest completed level of education of registrants is obtained as described in the document located at http://ow.ly/tL7YW. Data is obtained as of the date specified on the visualization. The map shows proportion of registrants with Bachelor's degrees and above from each country. The lists of countries and country codes used in parsing, can be found at http://ow.ly/tL0EU. The Python code used to obtain the data can be found at http://ow.ly/tL8iC. The resulting datasets are at http://ow.ly/tL8qS and http://ow.ly/tL8uz.

Specification of possible sources of error:
1. The mailing address data is self-reported and may not be fully correct.
2. The data is time-invariant, so it actually records the country where a person resided at the time of their registration, not currently.
3. IP geolocation is performed using MAXIMIND IP location database and is subject to its accuracy.
4. In case IP geolocation information is unavailable, the data is inferred from a parser developed in Python using regular expressions to match country, state and province names from self-reported addresses. It does not work well with accents such as é. We estimate the error level due to parsing errors to be up to 10% of the inferred total student count for each country, but it could be much higher or lower on average and specific countries might be much higher.
5. Mailing address data were not collected for the first few months of edX, so we rely solely on IP geolocation capabilities for that set of registrants.
6. IP geolocation is prone to registrants' using proxy servers to connect to edX, thus distorting their location in case the proxy IP used is geolocated in a different country.
7. Normally, for around 15-20% of students we are not able to detect geographic location by either IP geolocation or self-reported address parsing. The numbers of students for each country and course were inferred using the Missing At Random (MAR) assumption for missing data and has error. It is possible that MAR assumption is not fully satisfied as, for example, people from certain countries might be less likely to choose to share their address, and therefore could be underrepresented in the set of people for whom address parsing was used.
8. Possible sources of error associated with highest completed level of education data are

described in the document located at http://ow.ly/tL7YW.

Possible misinterpretations

1. It is important to note that this figure represents all registered students, without any kind of classification filter based on activity in the course. It may be that the distribution of active students by country is different than the distribution of registered students by country.
2. As noted above, we cannot verify the accuracy of the location data, and we know there are several possible sources of error. For analyses where precise residence data is important, it might be better to investigate further the possible sources of error listed above, as well as other methods of location detection which account for possibly systemic biases in the information.