International stocks and flows of students and researchers reconstructed from ORCID biographies

Orazbayev, Sultan

6 April 2017

International stocks and flows of students and researchers reconstructed from ORCID biographies

Sultan Orazbayev[1]

April 6, 2017

1. Independent researcher: contact@econpoint.com; ORCID: 0000-0003-4097-4830.

Abstract

This document describes a dataset of estimated bilateral flows and stocks of students and researchers (including some other types of high-skilled workers) for more than 200 countries (and territories) since 1990. The data is derived by analysing education and employment histories of more than 650 thousand individuals registered with ORCID. Comparison with independent data sources supports technical validity and representativeness of this data. The dataset provides new measures of the geography of a subset of high-skilled labour and opens opportunities for exploring hypotheses related to migration and agglomeration, impact of immigration policy, scientific production and development, academic mobility, and brain drain.

Keywords: high-skilled migration; high-skilled diasporas; student mobility; scientific mobility; ORCID.

Background & Summary

High-skilled workers are an important driver of economic growth and technological development. Measurement of high-skilled workers' mobility and location decisions using traditional methods (including surveys, census data and CV analysis) is costly and time-consuming [1, 2, 3]. Recently, large-scale bibliometric databases allowed measuring high-skilled mobility by tracking changes in affiliations provided by researchers when identifying themselves as authors of a publication [4, 5, 6]. This document describes a new source of data, Open Researcher and Contributor ID (ORCID) registry, which allows tracking stocks and flows of students and researchers (including some additional types of high-skilled workers) across countries and across time.

The motivation for creation of ORCID was to provide a central, free and open registry for contributors, which would help with contributor disambiguation, reduce the reporting burden and provide other benefits to the registered users [7]. This data also is a valuable source of education and employment histories of highly-skilled individuals — PhD students, academics, engineers,

medical professionals, researchers and, more broadly, contributors. In comparison with the common definition of a high-skilled worker as an individual with a tertiary education, the majority of ORCID users have or are in the process of obtaining a PhD, which means that ORCID data captures a subset of 'super' high-skilled workers. Figure 1 shows an example of an ORCID profile, which includes information on location and years of education and employment.

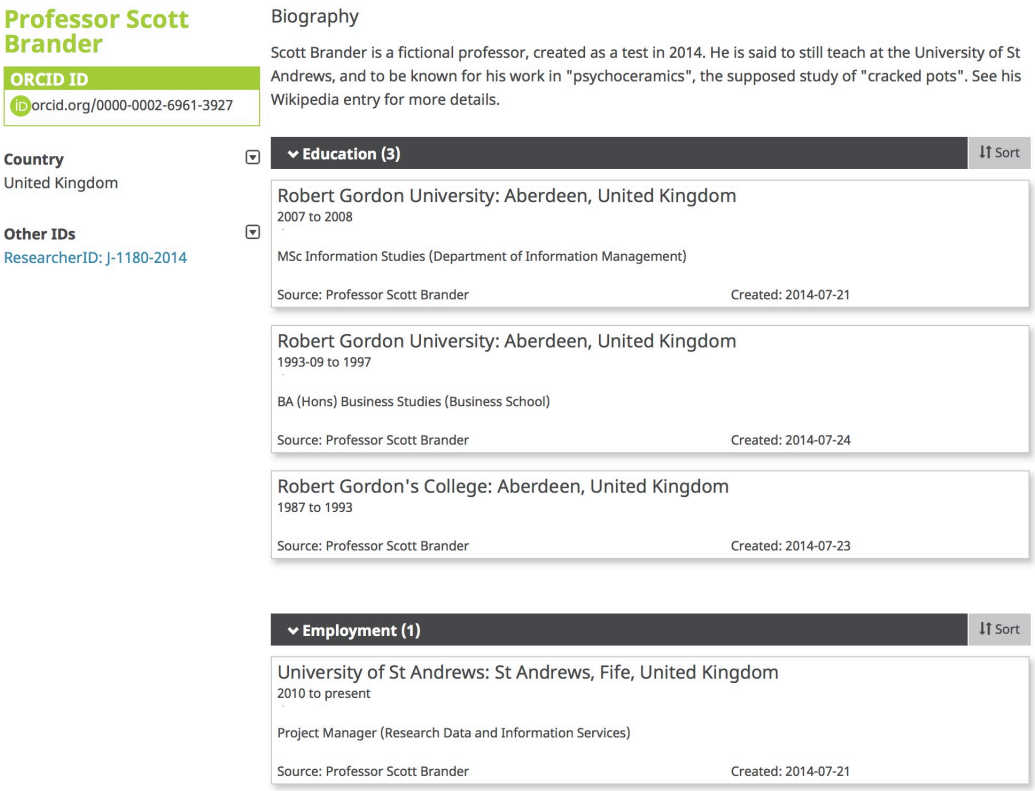


Figure 1: Sample ORCID profile.

ORCID profiles contain standardised information on education and employment, which allows tracking individuals across time. Specifically, it's possible to identify a user's geographic mobility and, with some additional assumptions, their most likely country of origin. The most comprehensive dataset to date which has this type of information is the GlobSci project based on a survey of approximately 17 thousand scientists in 16 countries [1]. ORCID dataset increases the sample size to approximately 650 thousand individuals located in more than 200 countries (and territories). Since ORCID was introduced only recently, the historical information is limited by the lifespan of the existing users. There are observations before 1990, however technical validation at the aggregate level shows that observations prior to 1990 are less consistent with

independent sources of data, so the core sample is restricted to 1990–2015. The next section describes the methodology used to create this longitudinal dataset of stocks and flows of students and researchers.

Methods

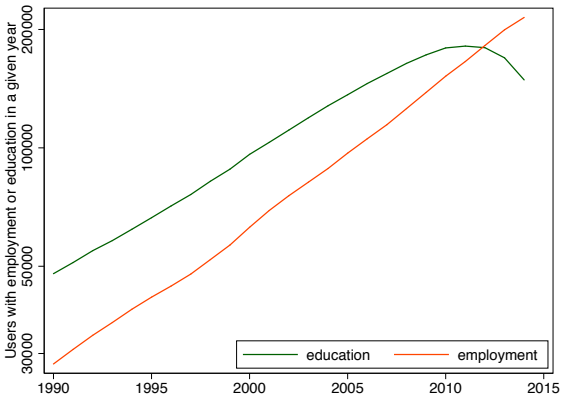
The dataset is constructed using ORCID 2016 Public Data File [8]. The raw data contains 2.5 million unique ORCIDs, of which 658 thousand are associated with a publicly-visible education or employment history. A few profiles are created for testing or amusement purposes, e.g. the fictional character whose profile was shown on Figure 1. Profiles that are known or suspected to be fake were removed from the sample, the full list of excluded ORCIDs is included among the dataset files.

The education and employment history includes a brief description of the position (e.g. PhD student, associate professor), affiliated organisation (e.g. university, medical hospital) and the starting and finishing dates. The level of detail and completion varies across users. Using information of users registered with ORCID as of 2016, it's possible to calculate their employment or education status in the past. Figure 2a shows the total number of users by their status in the past, sample size exceeds 30 thousand researchers and 50 thousand students per year after 1990. Combining the time and location data allows tracking a user's career over time and space. The career profile can be used to identify a user's most likely country of origin and international mobility events.

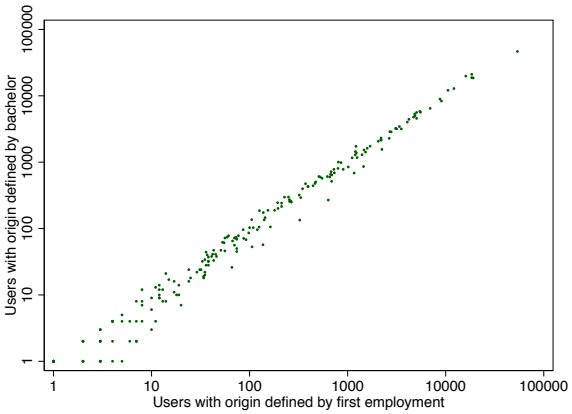
Key Assumptions

There are two key assumptions that have to be made during processing of the data: (1) how to treat incomplete information on the years of employment or education and (2) how to determine the most likely country of origin for a user.

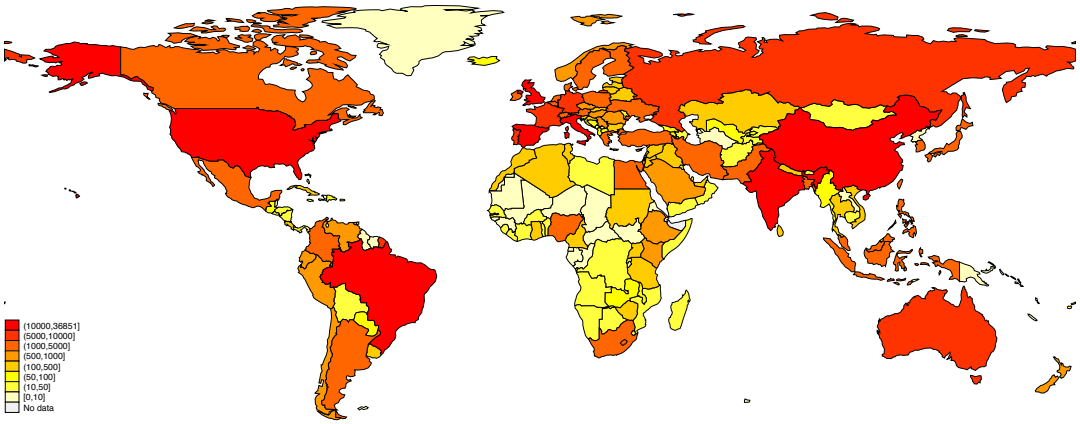
Some users provide only the starting or ending date for their education or employment event. For example, a user might specify obtaining a PhD in 2010, without specifying the starting year of the program. To impute the missing starting or ending date in such situations the following approach was used. If the position was identified as education, then it's title was processed and categorised as bachelor, master or PhD degree program. Categorisation was done using a manual list of patterns that capture standard formulations of the relevant program names, for example 'master in ...', 'bachelor of ...', 'PhD in ...' or 'DPhil in ...'. This matching procedure identifies education type (bachelor, masters or PhD) for 60% of the observations. Once the category is known, then the missing year information is added using assumption of 1-year for masters program (or for unidentified programs) and 4-years for bachelor and PhD programs. This length slightly underestimates the average length of the PhD program imputed from observations that contain both the starting and the ending dates (mean value is 4.5 years).



(a) ORCID users in October 2016 by their occupation in previous years.



(b) Comparison of origin assumptions.



(c) Number of users by imputed origin in 2015.

Figure 2: Selected aggregate patterns. Note: panel (a) shows the occupation type of ORCID users in previous years imputed from their public biographies in ORCID's October 2016 Public File [8]; panel (c) is drawn using *spmap*, see [9].

For employment records, if only the starting year was provided, then the employment was assumed to continue for at most 5 years, and if only the ending year was provided, then the employment was assumed to have started in the previous year. If both the starting and ending year were missing, then these observations would not be used in calculation of aggregate annual flows and stocks. Note that the missing end year for employment could also indicate ongoing employment, this is how the information is displayed in the ORCID web interface, so the 5-year limit is a fairly conservative assumption. However, variation of this assumption does not appear to have any meaningful impact on the aggregate patterns.

Each user's record was processed to identify their country of study or employment in different years. The user profile doesn't contain information on country of birth or nationality, so an assumption about the user's origin country is needed. The main approach used in constructing the dataset is to define a user's origin country as the country of the first education or employment, whichever is earliest. Other approaches are the country of undergraduate education and the country of first employment. These approaches give similar aggregate results, see Figure 2b, however technical validation of the data at individual level has higher match rate for origin imputed based on the country of the first education or employment, whichever is earliest. Hence, bilateral stocks were estimated using the country of origin imputed on the basis of the country of the first education or employment, whichever is earliest. The imputed origins cover almost every country (and territory) in the world, see Figure 2c.

The identification of the most likely origin could also be modelled using a more advanced process. For example, as suggested by an anonymous referee, the number of international students at an ORCID user's university could be used to assess the reliability of using the country of education as the most likely origin. If a person attended a very international university, then chances that the student's true origin is different from the university's country are higher than if the student attended a university with fewer international students. Such approaches however might also lead to increased error for domestic students, who might be mistakenly assigned a foreign origin based on attendance of an international university in their own country. However, with careful modelling it should be possible to increase the accuracy of identifying 'true' country of origin, which is an important area for future research since it will allow performing analysis at the individual level.

Calculating Bilateral Flows

Given information on a user's location in every year, it's possible to identify when the user relocates to a different country. First, each user's profile was converted into a long form: listing the user's location over all of the reported years. Then, a mobility event was identified as the combination of year, country of location during the last year (origin) and country of location during the current year (destination). These mobility events could correspond to short-term events, such as going on an exchange academic visit, or to longer-term

events, such as moving to a new country for work.

A small number of users report affiliations in multiple countries in a given year, for example in 2010 more than 96% of users had single-country affiliation(s). In cases of multiple-countries affiliation, it is assumed that the individual moves between all bilateral combinations of the reported countries. This assumption is not critical, excluding such users does not have a meaningful affect on the aggregate flows (correlation of 0.99).

The mobility events were aggregated to origin-destination country pair level for each year in 1990–2015 to obtain annual gross bilateral flows. Additional bilateral flows were calculated on the basis of the user's status in the destination (education and employment).

Calculating Bilateral Stocks

A user's (most likely) origin in combination with their current location allows calculating bilateral stock of migrants. The gross bilateral stocks were calculated by aggregating the number of registered users to destination-origin country pair for each year in the sample period, with additional variables capturing separately stocks of students and researchers (imputed based on the purpose of user's stay in the destination country).

Code availability

The computer code necessary for replication of the methods outlined above is archived and available at [10]. The computer code requires access to bash, jq (http://stedolan.github.io/jq/) and Stata (any version can be used to replicate the data, although the .dta file released with the data is in Stata 14 format).

Data Records

The dataset contains five files:

- the main table, which is provided in two formats: fully-labelled Stata 14 .dta file (data-v1.0.dta) and .csv file (data-v1.0.csv). Each of 90'839 observations in these files contains 9 variables described in Table 1;
- Stata .do file (rectangle.do) to 'rectangularise' the dataset (please see **Usage Notes**);
- the list of ORCIDs excluded because they are known or suspected to be fake (excluded.txt);
- the list of country codes and country names (country-names.txt).

Please see **Usage Notes** for additional comments regarding the proper interpretation of origin country for stocks/flows.

Variable name	Description
year	Calendar year in which the stocks of flows were observed
origin	Country from which the flow originated or the country of origin for the stock of individuals abroad
destination	Country to which the bilateral flow is flowing to or the current location for the stock of individuals
flow_all	Count variable that measures the number of students and workers that moved from origin to destination country
stock_all	Count variable that measures the stock of students and workers whose country of origin is origin and who are currently located in destination country
flow_students	Count variable that measures the number of students that moved from origin to destination country
stock_students	Count variable that measures the stock of students whose country of origin is origin and who are currently located in destination country
flow_workers	Count variable that measures the number of workers that moved from origin to destination country
stock_workers	Count variable that measures the stock of workers whose country of origin is origin and who are currently located in destination country

Table 1: Variable names and description.

Technical Validation

Micro-level Validation

ORCID profiles do not contain country of origin or birth, however many ORCID users have that information publicly available in an unstructured format, e.g. on their Internet homepage, in a publicly available CV or on their Wikipedia entry, if they have one. To confirm the validity of the assumption about the user's origin, information on all ORCID users with a Wikipedia page was processed to extract their place of birth, which was then matched to the relevant country. This step is needed to allow for changes in political borders over time, e.g. a person is born in a town in a country which splits at a later date, so the town becomes a part of a different country.

As of February 15, 2017, there were 1734 entries in Wikipedia (any language) that were associated with a unique ORCID. Information on the ORCID user's place of birth was recorded in 845 Wikipedia pages. After merging the Wikipedia information with ORCID data, there were 460 ORCID users that had both a Wikipedia-based place of birth and a public education/employment history. The Wikipedia-based and first country-based countries of origin matched for 348 users (out of 460), which is a 76% match rate.

The sample of ORCID users with a Wikipedia entry is not random and is likely to be positively selected on some measure of performance or publicity.

Although entries in any language were used, it's possible that ORCID users with a Wikipedia page come from or are currently based in a country with a large population (large origin or destination population makes it more likely, on the margin, that a Wikipedia entry will exist). Manual examination of the true (Wikipedia-based) origin countries for ORCID users that did not match with earliest-year based approach shows that these 112 users come from 43 unique countries, which range from large countries such as UK, Germany, France, Iran, Pakistan to smaller countries such as Romania, Kenya, Uruguay. The 348 users with matched information originated in 46 countries, with 28 countries present on both lists (of the matched and non-matched users). Overall, it's possible that some countries are under-represented when using the country of first education or employment year approach, but the relatively high match rate at the individual level suggests that the aggregated measures based on the proxied country of origin will be fairly reliable. In the absence of better data, further examination of the extent of any bias in determination of the country of origin at the individual level will have to be done through surveys or, in some cases, CV analysis.

Also, an indirect validation of the reliability of using the country of undergraduate education comes from the GlobSci survey which asked the respondents to report both their origin (defined as the country at age 18) and country of undergraduate degree. These countries were the same for 16'218 (out of 17'593) respondents, which is about 92% match rate. Note that defining origin as the country at age 18 as opposed to the country of birth could also explain the relatively lower match rate for Wikipedia data (which uses country of birth).

Macro-level Validation

An important consideration is that ORCID adoption (registration) can vary across countries, e.g. because researchers in some countries might not be aware of ORCID. In this case there will be an unaccounted-for measurement error affecting a country's bilateral stocks/flows. If the underlying 'true' population was known, then sampling weights could reduce the impact of this measurement error, however there is no data available that would capture the same geographical breadth and time sample. Instead, the computed aggregated flows and stocks are compared with similar measurements that are available from GlobSci survey and Scopus affiliation-based flows.

Bilateral stocks

The origin and destination data was compared with GlobSci results for year 2010 [1]. GlobSci survey examined approximately 17 thousand researchers in 16 'core' destination countries and four areas of science: biology, chemistry, materials and Earth and environmental sciences [1]. The survey asked the respondents for their country of residence at age 18, which roughly corresponds to the age at which individuals begin their undergraduate education or employment. This means that 'origin' definitions are quite similar between GlobSci and

the approach applied to ORCID data. The correlation of ORCID and GlobSci samples size is 0.87 by destination country and 0.78 by origin country (these and other correlations refer to the Pearson correlation coefficient).

GlobSci provides immigrant shares, calculated using information on the researcher's current location and their reported origin. Computing a similar measure for ORCID data results in a highly correlated measure, the correlation is 0.87 for immigrant share. Given that GlobSci was conducted in just 16 countries, their emigration share measure does not capture emigrants outside the sample countries, which could explain the relatively lower correlation of 0.58 between GlobSci and corresponding ORCID measure. The same sample restriction could explain the relatively low correlation of 0.58 for the emigrant's return rate, it's possible that GlobSci measure is affected by emigrants residing outside the sample countries.

The GlobSci's concentration rate measures the share of top 4 origin countries among all immigrants and it's highly correlated (0.82) with the corresponding statistic based on ORCID data. GlobSci defines international experience as working or studying outside the country of origin for at least one year, with the exception of work-related visiting scholar positions which need to last at least 6 months to qualify as international experience. Using ORCID data, it's possible to set international experience dummy equal to 1 if the individual was outside country of origin for at least one year. This depends critically on the information provided by ORCID users, who might under-report short-term visiting scholarships, especially if they are part of performing work for their main affiliation. Such underreporting or mis-measurement of international experience could explain the relatively lower correlation between GlobSci and ORCID-based measures of 0.46 for the share of individuals with international experience. Overall, the correlation of ORCID and GlobSci aggregate numbers is quite high, especially considering that ORCID data captures a broader population, so these results provide some support to the validity and representativeness of ORCID data at the aggregate level.

Bilateral flows

Flows of workers based on ORCID data were compared to the flows of authors of scientific publications over the sample period 1990–2014, as imputed from changes in reported affiliations in Scopus data [6]. The correlation between ORCID and Scopus measures of researcher flows using fill-forward approach (the correlation for fill-backward approach is given in brackets) varied from 0.69 (0.69) in 1990 to 0.80 (0.82) in 2014, with correlation for the whole sample at 0.68 (0.60). If the United States is excluded from the sample the correlations vary from 0.46 (0.44) in 1990 to 0.74 (0.75) in 2014, with correlation for the whole sample at 0.61 (0.53).

Affiliations-based flows of researchers rely on information derived from published research and will contain a measurement error in terms of occurrence of mobility events (e.g. due to researchers that visit a particular lab for a brief time, but claim affiliation of their home institution only) and timing of mobility

events (e.g. due to a publication lag). ORCID-based flows will capture mobility of researchers with a greater precision in terms of timing (reducing the impact of publication lag), but also will capture mobility of other high-skilled individuals (contributors), whose contributions are not necessarily visible in the form of authorship of academic publications. Given these considerations, the high correlation between ORCID and Scopus-based flows suggests that both approaches give mutually-consistent measures of international bilateral flows of researchers.

Usage Notes

Note that to reduce the file size the table includes only origins and destinations with at least one non-zero observations for any of the flow or stock variables. For empirical gravity-type of analysis, this table will need to be 'rectangularized' to make sure that all bilateral origin-destination combinations are included in every year of the observation. The relevant instructions are included in the accompanying computer code.

For stock data 'origin' country refers to the most likely country of individual's origin, but for flows 'origin' country refers to the country in which the individual was located last year (not necessarily their country of origin). This must be taken into account in calculation of brain drain/gain or other calculations that rely on the individual's (most likely) country of origin.

Furthermore, ORCID users can specify the location of education and employment events by choosing it from a specific list of countries (and territories) provided by the ORCID registry. This list is based on modern definitions of countries (and territories), which for some geographical locations might not correspond to a historical definition. For example, an ORCID user that studied in 1980 in city of Semey, formerly known as Semipalatinsk, would have studied in Kazakh SSR (part of USSR), but when entering this education record into ORCID registry the user will choose Kazakhstan as the country of education. This classification is advantageous for the purposes of calculating aggregate stocks and flows, because it allows tracking data associated with a set of fixed geographical locations (based on modern definition of countries and territories). If this data is merged with another dataset, however, it is important to cross-check the country and territory definitions for consistency.

The replication computer code allows reconstructing stocks and flows for any time period, however technical validation was performed only for the sample period (1990–2015). Researchers are advised to undertake additional validation when using data outside the sample period.

References

[1] Franzoni, C., Scellato, G. & Stephan, P. Foreign-born scientists: mobility patterns for 16 countries. *Nature Biotechnology* **30**, 1250–1253 (2012).

[2] Beine, M., Docquier, F. & Rapoport, H. Measuring international skilled migration: a new database controlling for age of entry. *The World Bank Economic Review* **21**, 249–254 (2007).

[3] Dietz, J. S., Chompalov, I., Bozeman, B., Lane, E. O. & Park, J. Using the curriculum vita to study the career paths of scientists and engineers: An exploratory assessment. *Scientometrics* **49**, 419–442 (2000).

[4] Moed, H. F., Aisati, M. & Plume, A. Studying scientific migration in scopus. *Scientometrics* **94**, 929–942 (2013).

[5] Moed, H. F. & Halevi, G. A bibliometric approach to tracking international scientific migration. *Scientometrics* **101**, 1987–2001 (2014).

[6] Czaika, M. & Orazbayev, S. Globalisation of scientific mobility. *Mimeo.* (2016).

[7] Haak, L., Fenner, M., Paglione, L., Pentz, E. & Ratner, H. Orcid: a system to uniquely identify researchers. *Learned Publishing* **25**, 259–264 (2012).

[8] Haak, L. *et al.* ORCID Public Data File 2016 (2016).

[9] Pisati, M. Spmap: Stata module to visualize spatial data (2008). Mimeo.

[10] Orazbayev, S. Replication code for: Reconstructing international stocks and flows of students and high-skilled workers from orcid biographies. figshare (2017). http://figshare.com/s/82f134e2568d9cfa5209.

Data Citation

1. Orazbayev, S. *Harvard Dataverse* http://dx.doi.org/10.7910/DVN/1RQAYC (2017).

Competing Financial Interests

The author declares no competing financial interests.