Data and replication files for 'City of dreams'

by Jorge De la Roca, Gianmarco I. P. Ottaviano, and Diego Puga

This page distributes and documents computer programs and data to replicate the results obtained by Jorge De la Roca, Gianmarco I. P. Ottaviano, and Diego Puga in their article **'City of dreams,'** to be published in the ***Journal of the European Economic Association***.

The primary data source is the National Longitudinal Survey of the Youth 1979 (NLSY79). This survey by the US Bureau of Labor Statistics (BLS) follows a nationally representative sample of men and women who were 14-22 years old in 1979. The computer programs on this page will serve NLSY79 data users to generate an individual-level panel for general purposes.

Bigger cities offer more valuable experience and opportunities in exchange for higher housing costs. While higher-ability workers benefit more from bigger cities, they are not more likely to move to one. The article proposes a model of urban sorting by workers with heterogeneous self-confidence and ability, which suggests that flawed self-assessment is partly to blame. Workers who misjudge their ability at an early career stage make location decisions they would not have made had they known their ability. By the time they learn enough about their actual ability, those early decisions have had a lasting impact, reducing their incentives to move and affecting their lifetime earnings.

Analysis of NLSY79 data shows that, in line with our model predictions, the location choices of young workers are guided by self-confidence rather than ability. Thus, some overconfident young workers start their career in a big city, while they would have chosen a small city with better self-assessment. That initial misjudged decision then becomes self-validating: having incurred a steep cost to gain more valuable experience, they find they might at least take advantage of this by remaining in the big city.

Conversely, some underconfident young workers spend their lives in a small city, even though a correct initial ability assessment would have made them self-select into a big city. Workers who severely underestimate their ability may nevertheless relocate from a small to a big city once labour market experience provides them with better information about their true capabilities. Young workers who are confident enough in their abilities locate in bigger cities to pursue their dreams, but those dreams do not come true for everyone.

The replication files

The full replication package is available for download from this site as a zip file: **`dreams_replication.zip`** (10.75 Mb) .

This replication package contains all the required data and code except for the restricted-access geocode files with the location (county) of respondents at birth, at age 14, and on every survey wave.

Obtaining access to the geocode NLSY79 data

Replicating the results of the published article requires, in addition to the code and data files provided here, access to the NLSY79 geocode data.

Only employees and students of US universities, employees of US federally-funded research centers, and employees of eligible US government institutions and non-profits can request access to the NLSY79 geocode data. The US Bureau of Labor Statistics (BLS) has no provisions for accessing the NLSY79 geocode data from outside the United States.

At the time of writing, one can find the application for obtaining access to the geocode NLSY79 data and information about the process at `https://www.bls.gov/nls/geocodeapp.htm`. In the application, the researcher must describe the project's research objectives in a few paragraphs. If the application is approved, the BLS will send the researcher a Letter of Agreement to be signed by an official institution signatory. The researcher must sign additional agreements, and in the case of students, their research advisor must be the signatory. Data access agreements are between the BLS and the recipient institution, not between BLS and individual researchers. All geocode data access must occur on the recipient institution's physical premises.

NLSY79 respondents. Shortly before the article's publication, the BLS transitioned its mode of provision of NLSY79 geocode data to a virtual data enclave (VDE). In this managed environment, researchers can analyze the geocode data. Statistical software available for use in the VDE includes Stata. Researchers can bring external files (such as the replication files for this project) and extract analysis results from the VDE, following a BLS approval process.

Instructions and overview of the replication files

These are the steps to construct the panel and replicate the results of the *Journal of the European Economic Association* article:

- Download and place the uncompressed replication package under some directory that will be the root directory of the replication files.
- Add to the `data/src/nlsy/geocode` directory four files that the BLS provides to authorised users of the geocode NLSY79 data: `location.dct`, `location-value-labels.do`, `survey_and_created_variables.dct`, and `survey_and_created_variables-value-labels.do`.
- Edit `code/_dreams_run.do` to specify in the line of `global PathProjectRoot` the path to the root directory of the replication files. The subdirectories `code`, `data`, and `results` will be under this directory on your computer.
- On the first run of the replication code, leave the flag `global InstallPackages = 1` in `code/_dreams_run.do` to install the required Stata packages (see Software and hardware notes below for details). Change the flag to `global InstallPackages = 0` for subsequent runs.
- Run `code/_dreams_run.do` in Stata.

The Stata script `code/_dreams_run.do` first runs `code/1_dreams_builddata.do` to perform the data construction. This code uses the data files described under Source data below, located in the directories `data/src/cbsa`, `data/src/nlsy`, `data/src/nlsy/cpi_ind_occ`, and `data/src/nlsy/geocode`. The script `code/1_dreams_builddata.do` first runs `code/builddata/nlsy_panel.do` to build a general-purpose panel that it saves as `data/processed/nlsy_panel.dta`. Next, it runs `code/builddata/nlsy_panel_sample.do` to prepare the panel for our estimations, adding core-based statistical area (CBSA) codes for locations, defining the junior and senior periods, constructing controls, and defining the final sample. This panel is saved as `data/processed/nlsy_panel_sample.dta`.

After the Stata script `code/1_dreams_builddata.do` creates the data file used for the analysis and places it in `data/processed/nlsy_panel_sample.dta`, the Stata script `code/_dreams_run.do` automatically runs `code/2_dreams_analysis.do` to perform the analysis of the processed data (described under Processed data below) and stores all the results (described under Results below) in the `results/` directory.

Using the code without the geocode NLSY79 data

While it is not possible to replicate the results of the *Journal of the European Economic Association* article without the restricted-access NLSY79 geocode data, researchers can run the code without these additional data under two scenarios.

Researchers who wish to use the public-use NLSY79 data for other projects that do not require the location of respondents can use our code as a starting point. To do this, edit `code/_dreams_run.do` and set the flag `global NLSYDisableGeocode = 1`. This adjustment will produce a general-purpose panel (described under Processed data below but without location data) saved as `data/processed/nlsy_panel.dta`. However, the panel will not produce any tables or figures with results.

Researchers without access to the geocode NLSY79 data that wish to check that the replication code runs smoothly can edit `code/_dreams_run.do` and set the flag `global NLSYGenerateFakeLocations = 1`. This adjustment will randomly generate a fake location history for each respondent, allowing the code to run but generating meaningless results with the same format but different values than the actual results in the article.

Software and hardware notes

The results and figures in the *Journal of the European Economic Association* article have been produced using the code and data provided in Stata version 17.

Stata. The following additional Stata packages are required:
  - estout: module to make regression tables from stored estimates, by Ben Jann.
  - grstyle: module to customize the overall look of graphs, by Ben Jann.

  These required Stata packages can be installed automatically by setting the flag `global InstallPackages = 1` in `code/_dreams_run.do`. This installation should be done on the first run of the code, then changing the flag to `global InstallPackages = 0` for subsequent runs.

  The code has been tested with Stata version 17. The code will impose `version 17` for more robust replicability when run on newer Stata versions.

- **Operating system**: None of the Stata code is operating-system-specific.

- **Hardware**: The run of the code producing the results reported in the published version of the article was performed on an Apple iMac Pro with a 3 GHz 10-Core Intel Xeon W processor and 128 Gb of DDR4-2666MHz RAM. However, running the code does not require a powerful computer. The run took place on 26 July 2022 and took 65 seconds. The data construction and analysis took 12 and 53 seconds, respectively. The log of this run is provided with the replication files in `code/logs/log_2022.07.27_12.34.44.txt`.

Source data

The primary source data combines public-use and restricted-access geocode data from the National Longitudinal Survey of the Youth 1979 (NLSY79). The replication code reads all source NLSY79 data files from the `data/src/nlsy` directory and its `data/src/nlsy/geocode` child directory. The required public-use NLSY79 data files are included with the replication file in the `data/src/nlsy` directory. The required public-use NLSY79 data files are the following:

- `raw_data.dct`: Fixed-format Stata data file with a dictionary containing a selection of NLSY79 variables. The data set has individuals' raw and percentile Z-scores on the Rosenberg test for 1980 (our self-confidence measure) and the AFQT for 1980 (Armed Forces Qualification Test, our ability measure). The data set also includes information on individuals' demographics (e.g., gender, age, race/ethnicity, educational attainment), household composition (e.g., marital status, number of children, age of children), and labor force status and job characteristics (e.g., experience, tenure, occupation, sector of economic activity, hours of work, and wages).
- `raw_data-value-labels.do`: Stata program provided by the BLS to assign value labels to variables contained in `raw_data.dct`.
- `lfstatus_data.dct`: Fixed-format Stata data file with a dictionary containing NLSY79 variables needed to complete information on respondents' labour force status. This status is available in `raw_data.dct` for all years except for 2000-2005 and from 2007 onwards. The file `lfstatus_data.dct` contains weekly arrays of labour force status for the missing years. We can recover the labour force status at the time of the interview for the missing years by combining the weekly arrays with the interview date for each year contained in `raw_data.dct`.
- `lfstatus_data-value-labels.do`: Stata program provided by the BLS to assign value labels to variables contained in `lfstatus_data.dct`.

If desired, the researcher can re-download from the BLS these files (for instance, to obtain additional variables). The files `raw_data.NLSY79` and `raw_data.NLSY79` are saved tagsets that make it easy to select the required variables for download in the NLSY Investigator platform (`https://www.nlsinfo.org/investigator`).

The restricted-access geocode NLSY79 data files are not included with the replication file. The researcher must request them from the US Bureau of Labor Statistics (BLS) and place them in the `data/src/nlsy/geocode` directory, as explained above. The required restricted-access geocode NLSY79 data files are the following:

- `location.dct`: Fixed-format Stata data file with a dictionary containing the locations (county) of respondents' residence at the time of the survey for different survey years.
- `location-value-labels.do`: Stata program provided by the BLS to assign value labels to variables contained in `location.dct`.
- `survey_and_created_variables.dct`: Fixed-format Stata data file with a dictionary containing the locations of residence of respondents at birth and age 14, among other variables.

In addition, complementary data files are needed to deflate nominal wages and create standardised time-consistent codes of occupation and sector. The researcher can find these files in the `data/src/nlsy/cpi_ind_occ` directory. Furthermore, researchers with access to the geocode NLSY79 can assign counties to metropolitan areas using the data files in the `data/src/cbsa` directory. We describe these auxiliary data files below:

- `occ1970_occ1990dd.dta`, `occ1980_occ1990dd.dta`, `occ2000_occ1990dd.dta`: Stata data files created by David Dorn (`https://www.ddorn.net/data.htm#Occupation%20Codes`). Each file provides a crosswalk of occupations in a decade (1970, 1980, and 2000) to their 1990 equivalent. Using these crosswalks, we create an occupational code variable that sequentially transforms occupation codes in 2002 and onwards into their 1990 equivalents, followed by occupation codes between 1982 and 2000 into 1990 codes. Lastly, we transform occupation codes before 1982 into 1990 occupation codes. The result is a standardised occupation variable with consistent codes during the analysis period.
- `ind1990_tt.xls`: Excel file that contains an IPUMS crosswalk between three-digit industry codes in 1990 and other code classifications from 1950 to 2005. The NLSY79 reported 1970 industry codes until 2000 and then switched to 2000 industry codes. We transform industry codes before 2000 into their 1990 equivalents and then replace industry codes since 2002 using their 1990 equivalents. The result is a standardised three-digit sector variable using the 1990 IPUMS industry classification.
- `cpi_sa_1947_2015.xlsx`: Excel file that reports the monthly Consumer Price Index from 1947 to 2015 by the BLS. The seasonally-adjusted index (ID=CUSR0000SA0) is an average across US cities and contains all expenditure items. We use December values to construct an annual index (Base 1982-1984 = 100) and deflate nominal wages.
- `CBSA_def2009.xls`: Excel file that contains a crosswalk between FIPS county codes and Core Based Statistical Areas (CBSAs) in 2009. The Office of Management and Budget (OMB) constructs CBSAs as metro areas of one or more counties (or equivalents) anchored by an urban center and any adjacent counties that are economically linked to the center by commuting patterns. This crosswalk allows the researcher to assign the counties at birth, age 14, and for every survey year to a CBSA or non-CBSA area.
- `metro_1980_2010.dta`: Fixed-format Stata file from US2010, a joint project between the Russell Sage Foundation and Brown University, that shows the population of metro CBSAs (i.e., those with an urban core above 50,000 people) in 1980, 1990, 2000, and 2010. We use the 2010 population to determine whether respondents live in a big metropolitan area (i.e., above 2,000,000 people).

We combine these source files to construct an annual panel from 1979 to 1994 and a biennial panel from 1994 to 2012.

Processed data

The replication code fully recreates the processed data from the original sources and performs the data analysis. The processed data consist of the following files and variables:

- `data/processed/nlsy_panel.dta`. This data file is a general-purpose panel with NLSY79 data and contains the following variables:
  - **person_id**. Individual identifier.
  - **year**. Year.
  - **non_int**. Non-interview indicator.
  - **birth_year**. Birth year.
  - **birth_month**. Birth month.
  - **age**. Age.
  - **sex**. Sex.
  - **race**. Race.
  - **sample_type**. Sample classification.
  - **educ_enrolled**. Educational enrollment status.
  - **educ_highest**. Highest education attained at survey date.
  - **educ_mother**. Mother's years of education.
  - **educ_father**. Father's years of education.
  - **marital**. Marital status.
  - **spouse_wkswk**. Number of weeks worked by spouse in past calendar year.

- **children**. \# of children (since 2000, only for female respondents and only bio children).
- **agechildren**. Age of youngest child (only individuals with children).
- **lfstatus**. Labor force status (generated for selected years).
- **wkswk_li**. Weeks worked since last interview.
- **worker_type**. Class of worker (job type), CPS definition.
- **wage**. Hourly real wage, main job, dollars, CPS definition (1982-84 = 100).
- **tenure**. Tenure with employer (years).
- **experience**. Cumulative experience (years).
- **unemployment**. Cumulative unemployment spells (weeks).
- **outlf**. Cumulative out-of-labor-force spells (weeks).
- **occupation3d**. Occupation, CPS definition, 1990 Census 3-digit (standardized).
- **occupation2d**. Occupation, CPS definition, 1990 Census 2-digit (standardized).
- **sector3d_90**. 3-digit sector (IPUMS consistent long-term classification, 1990 basis).
- **sector2d_90**. 2-digit sector (IPUMS consistent long-term classification, 1990 basis).
- **afqt_80**. AFQT percentile (full sample), 1980 (revised in 2006).
- **self_80_score**. Self-confidence, raw score, 1980.
- **self_80_zpctl**. Self-confidence, weighted z score percentile (full sample), 1980.
- **risk_scale**. Risk aversion (10 = fully prepared to take risks).
- **fips_birth**. Residence FIPS code of birth county.
- **fips_at14**. Residence FIPS code at age 14.
- **fips**. Residence FIPS code.

Note: the variables fips_birth, fips_at14, and fips require the restricted-access geocode NLSY79 data. As explained above, users without access to these data can edit `code/_dreams_run.do` and set the flag `global NLSYDisableGeocode = 1` to build the panel without these three variables. Alternatively, they can set the flag `global NLSYGenerateFakeLocations = 1` and get randomly-generated values for these three variables (obviously, researchers should only select this path to check that the replication code runs smoothly since the results generated will be meaningless).

- `data/processed/nlsy_panel_sample.dta`. This data file is the panel used for our analysis. Relative to `data/processed/nlsy_panel.dta`, it imposes our sample restrictions and contains the following additional variables:
  - **junior_period**. Junior period indicator: One year after completing highest level of education.
  - **senior_period**. Senior period indicator: 10 years after the junior period.
  - **hispanic**. Hispanic.
  - **black**. Black.
  - **educ_highest_life**. Highest level of education completed (lifetime).
  - **educ_high_school**. High-school completed (12 years of education).
  - **educ_some_college**. Some college completed (between 13 and 15 years of education).
  - **educ_college**. College completed (16 or more years of education).
  - **never_married**. Never married.
  - **married**. Married, spouse present.
  - **spouse_work_26w**. Spouse worked during previous calendar year (26 or + weeks).
  - **spouse_work_48w**. Spouse worked during previous calendar year (48 or + weeks).
  - **spouse_work_ft**. Spouse worked full-time during previous calendar year (40 or + weekly hours).
  - **child_under7**. At least one child under 7 years old in household.
  - **ln_wage**. Log hourly real wage in main job (1982-84 dollars).
  - **afqt_pctl**. Cognitive ability AFQT percentile, 1980.
  - **self_pctl**. Rosenberg self-confidence percentile, weighted z scores, 1980.
  - **cbsa_birth**. CBSA id of birth county.
  - **cbsa_birth_name**. CBSA name of birth county.
  - **cbsa_birth_pop2010**. CBSA 2010 population at birth.
  - **cbsa_birth_small**. Individual was born in small CBSA.
  - **cbsa_birth_type**. CBSA type of birth county (1 = Metro, 2 = Micro).
  - **cbsa_at14**. CBSA id of county at age 14.
  - **cbsa_at14_name**. CBSA name of county at age 14.
  - **cbsa_at14_pop2010**. CBSA 2010 population at age 14.

- **cbsa_junior_sameas_at14**. Same CBSA during the junior period as at age 14.
- **cbsa**. CBSA id (2009 definition).
- **cbsa_name**. CBSA name.
- **cbsa_type**. CBSA type (1 = Metro, 2 = Micro).
- **cbsa_pop2010**. CBSA 2010 population.

Note: the variables with names beginning with cbsa require the restricted-access geocode NLSY79 data. As explained above, users without access to these data can edit `code/_dreams_run.do` and set the flag `global NLSYGenerateFakeLocations = 1` to get randomly-generated values for these variables (obviously, researchers should only select this path to check that the replication code runs smoothly since the results generated will be meaningless).

Results

After running `code/1_dreams_builddata.do` to create the data file used for the analysis, the Stata script `code/_dreams_run.do` automatically runs `code/2_dreams_analysis.do` to perform the analysis of the processed data. Specifically, `code/2_dreams_analysis.do` runs in sequence the Stata scripts `code/analysis/dreams_table1.do`, `code/analysis/dreams_table2.do`, `code/analysis/dreams_table3.do`, `code/analysis/dreams_table4.do`, and `code/analysis/dreams_table5.do` to produce the LaTeX code for the tables in the article. Subsequently, it runs the Stata scripts `code/analysis/dreams_figure1.do` and `code/analysis/dreams_figure2.do` to produce the figures, and `code/analysis/dreams_text_results.do` to calculate various numbers mentioned in the text.

All the results are placed in the `results/` directory.

After running `code/_dreams_run.do`, the researcher must compile in LaTeX the file `results/dreams_tables.tex` to produce a PDF file with all the tables.

Figures are saved in Encapsulated PostScript format as `results/dreams_fig1.eps` (figure 1); `results/dreams_fig2a.eps`, `results/dreams_fig2b.eps`, and `results/dreams_fig2c.eps` (the three panels of figure 2); and `results/dreams_figb1a.eps` and `results/dreams_figb1b.eps` (the three panels of appendix figure B.1). They are also saved in PNG format with the same file names and extension `.png`.

The results mentioned in the text, besides those contained in tables, are calculated by `code/analysis/dreams_text_results.do`. This Stata script automatically writes the relevant paragraphs to the text file `results/dreams_text_results.txt`. This text file reads as follows:

```
'City of dreams', by Jorge De la Roca, Gianmarco I. P. Ottaviano, and Diego Puga

Results mentioned in the text not contained in tables.

Section 1

According to our data, 56% of all individuals (and 42% of the college-educated) in the
United States live in the same city at ages 14 and 40.

Our data show a low correlation of 0.21 between ability and self-confidence (our measure of
ability self-assessment). Among college graduates, this correlation falls to 0.02.

Our primary measure of ability is the individual's percentile score in the Armed Forces
Qualification Test (AFQT), a general ability test administered to respondents in 1980 when
they were between 15 and 23 (with a median age of 19).

Section 3

Our measure of ability is the individual's percentile score in the Armed Forces
Qualification Test (AFQT). This general ability test was administered in 1980 when NLSY
respondents were between 15 and 23 (with a median age of 19), regardless of their interest
in the military.
```

market experience provides workers with a better self-assessment of ability. Since the age of NLSY79 respondents ranged between 15 and 23 when tested in 1980, a way to see if self-assessment improves over time with job experience is to analyse whether self-confidence and ability are more correlated for older respondents at the time of the tests.

Regarding timing, we set the junior period for all respondents at the year after their highest level of education is completed, excluding educational periods that happen after more than two years away from education (median age of 20 for individuals without post-secondary education and 24 for the college-educated).

Based on these counties, we determine whether each respondent lives in a Core Based Statistical Areas (CBSA) with a 2010 population above two million. If so, we classify them as living in a big city, otherwise as living in a small city. This population threshold leads to 40% and 39% of individuals living in big cities during their junior and senior periods respectively.

The initial sample includes all 6,111 individuals in the cross-sectional sample of the NLSY79. We exclude individuals for whom the AFQT or the Rosenberg self-esteem scores are missing, which reduces the sample to 5,671 individuals. We can determine the junior period location of 5,462 of these individuals and, due to sampling attrition, the senior period locations of 5,180 of them. The availability of the demographic controls that we include further reduces our sample to 5,254 individuals in the junior period analysis and 4,985 individuals in the senior period analysis.

Section 6

One year after completing their education, 71% of individuals in our sample are in the same city as at age 14, and 61% remain there by age 40.

In table 2, 33.6% of individuals move between both periods while only 13.4% change city-size class (i.e, SB or BS).

Importantly, self-assessment of ability relative to people with the same education is so imperfect that there is virtually no correlation (0.02) between self-confidence and ability among college-educated workers.

References

De la Roca, Jorge, Gianmarco I. P. Ottaviano, and Diego Puga. Forthcoming. City of dreams. *Journal of the European Economic Association*.

Jann, Ben. 2004. estout: Stata module to export estimation results from estimates table.

Jann, Ben. 2017. grstyle: Stata module to customize the overall look of graphs.

27 July 2022, diegopuga.org
html5, css3

diego.puga@cemfi.es