This Readme documents the necessary steps to replicate all results in the paper and appendix of: "**The Margins of Global Sourcing: Theory and Evidence from U.S. firms**" by Pol Antràs, Teresa Fort, and Felix Tintelnot.

To reproduce the tables and figures in the paper:

1. All the results in the paper use confidential microdata from the U.S. Census Bureau. To gain access to the Census microdata, follow the directions here on how to write a proposal for access to the data via a Federal Statistical Research Data Center: https://www.census.gov/ces/rdcresearch/howtoapply.html.

2. You must request the following datasets in your proposal:
    a. Longitudinal Business Database (LBD), 1997, 2002 and 2007
    b. Foreign Trade Database – Import (IMP), 1997, 2002 and 2007
    c. Longitudinal Foreign Trade Transactions Database—LFTTD, 1997, 2002 and 2007
    d. Census of Construction (CCN), 1997, 2002 and 2007
    e. Census of Finance, Insurance, and Real Estate (CFI), 1997, 2002 and 2007
    f. Census of Manufactures (CMF), 1997, 2002 and 2007
    g. Census of Mining (CMI), 1997, 2002 and 2007
    h. Census of Retail Trade (CRT), 1997, 2002 and 2007
    i. Census of Services (CSR), 1997, 2002 and 2007
    j. Census of Transportation, Communications, and Utilities (CUT), 1997, 2002 and 2007
    k. Census of Wholesale (CWH), 1997, 2002 and 2007
    l. Standard Statistical Establishment List (SSEL), 1997, 2002 and 2007

3. Your proposal should also list the following external datasets that are available in "RDC_Replication_Files/Input_Data":
    a. china_input_shock_naicsb97.dta
    b. china_shock_naicsx97.dta
    c. country_data.dta
    d. final_dataset_v3.dta
    e. HIchina_input_shock_naicsb97.dta
    f. HIchina_ shock_naicsx97.dta
    g. HIindustry_dataset_naicsx97.dta
    h. hs6_naics_imp_year_97.dta
    i. hs6_naicsx_imp_year_97.dta
    j. hs10_naics07_imp_year_07.dta
    k. industry_dataset_naicsx97.dta
    l. io_table_1997.dta
    m. naics_97_02.dta
    n. naics_x97.dta
    o. naics97_naicsb97.dta
    p. naics97_naicsx97.dta
    q. naicsx97_naicsb97.dta

4. You should also reference "The Margins of Global Sourcing: Theory and Evidence from US Firms" by Pol, Antràs, Teresa Fort, and Felix Tintelnot, project number br1179 in the proposal. This may give you access to the programs and input datasets required to reproduce the results.

5. You should create a directory with the following subdirectories:
   - programs (place all Stata and SAS programs here)
   - output (figures and tables will be created here)
   - data
   - external_data (place all data in RDC_Replication_Files/Input_Data in here).
   - bs
   - matlab (place matlab programs in here)

6. You must cd to the main directory you create for replication in the beginning of each program, and/or define global macros for the appropriate sub directories at the top of each program. Every program has a "Set directory" command in the beginning where you can do this. Once set, each program is written to access automatically each of the subdirectories listed above.

7. You can now recreate all the descriptive tables and figures in the paper by running the script_moo.bash program. To do so, cd to the program directory and enter the command: "qsub script_moo.bash". All the output will be created in the output subdirectory. Please note that we often include variables in output datasets that are not used in our analyses. Use these variables at your own risk-- we have not verified that they were correctly constructed or that they are valid for research purposes.

8. You can now recreate the additional structural estimation output in the paper using output from step 7 above. To do so, follow the detailed steps in the Readme: "RDC Replication Files 5_3 Estimation and Bootstrap" in "RDC_Replication_Files\Programs\Section 5_3 estimation and bootstrap." While the bootstrapped standard errors can only be replicated at the RDC, a large set of results can be reproduced externally using a set of disclosed moments from the RDC (see 9 below).

9. Some of the structural results in the paper can be reproduced using a set of disclosed moments from the microdata. The requisite input files are provided in "Public_Replication_Files\Input_Data" and the programs to replicate the results are in "Public_Replication_Files\Programs".
   a. The set of disclosed moments is called: emp2_data_for_matlab_v3. Some of these are presented in Table 1. The full set of moments is used to estimate the model structurally and perform counterfactuals.
   b. Table 1 and a number of the structural estimation results can be replicated using these moments. With the disclosed moments in hand, the RDC files are required only to get standard errors for the estimates shown in Table 5. See the Readme "Readme public replication files" in: "Public_Replication_Files\Programs\Section_5_3_to_Section_6_3_and_selected_Appendix_tables".

10. Here we describe how the external datasets listed in 3 above were constructed. These data are all constructed using non-confidential data that are external to the US Census Bureau. The data were constructed as follows:
    a. (HI)china_shock_naicsx97.dta: Chinese import penetration in a collection of high income countries, by year, using public Comtrade data. HS to NAICS concordance done using the Pierce-Schott concordance. From NAICS, we aggregated up slightly to NAICS-X, which is the level at which all industries have positive imports. (Note NAICS-X is not a mutually exclusive classification but this is fine because the shocks are in shares). If there is prefix HI, means that we use the Autor-Dorn-Hanson set of high-income countries. Without the prefix HI, we use our preferred set of countries from Europe (see paper).

    b. (HI)china_input_shock_naicsb97.dta: For a given industry, the weighted average of the above China shock for all industries, where the weights are the input-use weights. (see paper). Requires slight aggregation of NAICS-X so that they can be concorded with BEA IO codes – we call this NAICS-B. This is the 'input-shock' analog of the above 'output shock'.

    c. (HI)industry_dataset_naicsx97.dta: Summary dataset with external trade volumes of (a) US and (b) collection of high income countries with (i) China and (ii) all countries including China. Trade data from Comtrade, using the same HS-NAICS concordance. Some other variables such as emp and vship come from the NBER-CES manufacturing database (available inside the RDCs). If prefix is HI, then the (b) set of high income countries is the Autor-Dorn-Hanson set.

    d. country_data.dta = final_dataset_v3.dta: A country-level dataset with information from the World Bank (WDI, WITS, Governance Indicators), Penn World Tables, Barro & Lee, and the ILO. Additional details are in the paper and online appendix, and in the data construction directory. These are the same dataset but saved with different names because different programs call this dataset differently.

    e. hs6_naics_imp_year_97.dta: a concordance between hs6 and naics codes from Pierce and Schott

    f. hs6_naicsx_imp_year_97.dta: a concordance between hs6 and our NAICS-X classification. NAICS-X is the least aggregated version of NAICS such that each industry has positive imports in the data, but it tries to preserve as much disaggregation as possible

    g. hs10_naics07_imp_year_07.dta: a concordance between hs10 codes and NAICS 2007 classification codes, form Pierce and Schott

    h. io_table_1997.dta: The BEA input-output use table 1997 in NAICS-B classification

    i. naics_97_02.dta: A mapping from NAICS 1997 to NAICS 2002 codes.

    j. naics_x97.dta: A list of codes in our NAICS-X classification

    k. naics97_naicsb97.dta: A mapping from NAICS to NAICS-B codes.

l. naics97_naicsx97.dta: A mapping from NAICS to NAICS-X codes

m. naicsx97_naicsb97.dta: A mapping from NAICS-X to NAICS-B codes