This Readme documents the necessary steps to replicate all results in the paper and appendix of: "**Technology and Production Fragmentation: Domestic versus Foreign Sourcing**" by Teresa Fort.

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), 2002 and 2007
    b. Foreign Trade Database – Import (IMP), 2002 and 2007
    c. Annual Survey of Manufactures (ASM), including the Computer Network Use Supplement (CNUS), 1999
    d. Census of Construction (CCN), 2002 and 2007
    e. Census of Finance, Insurance, and Real Estate (CFI), 2002 and 2007
    f. Census of Manufactures (CMF), 2002 and 2007
    g. Census of Mining (CMI), 2002 and 2007
    h. Census of Retail Trade (CRT), 2002 and 2007
    i. Census of Services (CSR), 2002 and 2007
    j. Census of Transportation, Communications, and Utilities (CUT), 2002 and 2007
    k. Census of Wholesale (CWH), 2002 and 2007
    l. Standard Statistical Establishment List (SSEL), 2002 and 2007

3. You should also reference "Technology and Production Fragmentation: Domestic versus Foreign Sourcing" by Teresa Fort, project number br1179 in the proposal. This will give you access to the programs and input datasets required to reproduce the results. Note that all the programs required to reproduce the results are also available on the REStud website.

4. You should create a directory with the following subdirectories:
    - programs (place all programs here)
    - output (figures and tables will be created here)
    - data
    - input_data (place all input data here)

5. You must cd to the main directory you create for replication in the beginning 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 each of the subdirectories listed above.

6. You can now recreate all the tables and figures in the paper by running the script_frag.bash program. To do so, cd to the program directory and enter the command: "qsub script_frag.bash". All the output will be created in the output subdirectory.

7. Note that figures 2 and A.4 are created using disclosed data. These have a separate Stata program (Make_figures_public.do) to create them. The requisite input data is also provided.

Public data:

1. Port and border crossings data: In the paper, I construct a new dataset with geocodes for deep water ports and border crossings. These data are posted online, along with the programs used to construct them.

2. CMS industry shares: In the paper, I calculate the share of establishments that purchase CMS by NAICS4 industry. These shares have undergone disclosure review and are posted online.

Dataset construction summary:

#1. DATASET: frag_dataset_orig.dta

1. Match 2007 and 2002 import data to firms
    a. I match the import files directly to the EC data as well as to the LBD. These data are aggregated to the firm level using the firmid variable. The EC concordances between EIN-firmid were based on all the EC data. I also made a bridge using the Business Register

2. Construct a panel establishment-level dataset with fragmentation questions
    a. These datasets are constructed using the CMF and the CWH data in 2002 and 2007. I pull in the other establishment activity questions, as well as all the other variables needed for the analysis. I merge these data to the LBD and use the lbdnum as the longitudinal identifier.

#2. DATASET: NL_dataset.dta

1. These data are constructed from the plant fragmentation data, but I create an alternative-specific measure of distance for the NL specifications.

#3. DATASET: firm_country_regs.dta

1. Construct a firm-level dataset with import information
    a. I aggregate the plant-level fragmentation data to the firm level and merge with the import data.
2. Create an observation for every firm-country combination and indicator for whether the firm imports from a particular country.
3. Note that these analyses are limited to firms with at least one offshoring establishment, to firms firms with some import activity, and to countries with at least 10 importing firms.

#4. DATASET: firm_imports.dta

1. Aggregate the firm-country-level data to the firm level to calculate summary statistics. This dataset is based on all firms in the sample

#5. DATASET: firm_imports_low_inc.dta

1. This dataset is made from the firm-country-level dataset, but is aggregated to the firm country-income level. This dataset is based on all firms in the sample.

#6. DATASET: NAICS4_variables.dta

1. This dataset contains NAICS4 variables constructed from the public NBER productivity database numbers.
2. These data are combined with Nunn's measure, and Lindsay Oldenski's Routineness measures. Please be sure to cite Costinot et al. (2011) if you use the routineness measures.

#7. DATASET: ind_vars.dta

1. These industry-level data are constructed from the 1999 CNUS data. I aggregate the plant data to the industry level to calculate the share of plants in an industry that used CAD/CAM in their production. I also calculate the share of plants that use networks to communicate with suppliers and other company units from the CNUS.
2. This information is merged with: a) capital and skill intensity constructed directly from the CMF data; b) the ratio of imports by wholesale firms relative to domestic production by manufacturers from the import data and the CWH data; and c) Nunn's measure of differentiated inputs concorded to NAICS2002 codes.

Notes:

- I use NAICS 2002 vintage codes throughout since these facilitate concording to older NAICS and SIC vintages, which is necessary for creating a number of industry level variables (e.g., CAD intensity, Nunn's measure, Routineness, etc.)