Please follow the instructions below to replicate the results from the paper

Trade, Gravity and Aggregation

16 June 2022

by Holger Breinlich, Dennis Novy and J.M.C. Santos Silva

0) Before running any of the do-files to replicate the empirical results, the user needs to create several folders for later use. The first is the main folder that will contain all data and results. This folder should contain the files *data_preparation_16062022.do* and *produce_results_16062022.do* and the path for this folder must be set in lines 7 and 3 of the files *data_preparation_16062022.do* and *produce_results_16062022.do*, respectively. The main folder must also contain a subfolder named *raw data* with the following files: *BB-countrylist.xlsx*, *data_fta_fixed.dta*, *dist_cepii.dta*, *geo_cepii.dta*, *limao_handbook_pta.dta*. Additionally, *raw data* must have a subfolder named *ComtradeBatch_1962-2010* containing the 15 Comtrade data files,[1] as explained in footnote 9 of the paper. As we discuss there, we use data from Comtrade for the period 1960-2000. More specifically, we use the value of bilateral imports in current US dollars on a c.i.f. basis. These original data need to be at least at the 4-digit level of the SITC classification (or more disaggregated than that). The do-files will collapse them to different levels of aggregation (0-, 2-, or 4-digit) as explained in the paper.

1) After creating the above folders, run the file *data_preparation_16062022.do*. This file uses the data in the *raw data* folder to create the Stata dta files that will be used to replicate the results in Step 2. So, the necessary data files need to be placed in the *raw data* folder as explained above before running this file. Note that lines 10-11 create the folder *dta-files* where the processed data will be saved and lines 13-14 create a folder named *results* where all results will be saved. This do-file creates two datasets in the *dta-files* folder. The first, BB_allyears_all_zeros_nomirror.dta, contains the fully rectangularized dataset, where all missing country-pair-year-product observations have been filled in and the corresponding value of the trade flow variable (imports) has been set to zero. The second is a version where all zeros are dropped, so that the data is broadly comparable to the data used in Baier and Bergstrand's (2007, JIE) paper. Both datasets will be used in the second step. Note that there is also the option of using mirror flows to impute missing trade flows (this is determined by how the global variable mirror is set in line 19 of the do-file). For the results in the paper, we didn't use mirror flows because the correlation with observable trade flows is rather low (i.e., when both sides of a trade flow are observed, the correlation between what's reported by the exporter and the importer is low).

2) Next run the file *produce_results_16062022.do*, which reproduces the results from the paper. Note that the used needs to preinstall several commands such as *reghdfe* and *ppmlhdfe*. Also, as noted above, the file requires setting a global variable in line 3 to indicate the location of the relevant files. The do-file will save a few auxiliary files in the *dta-files* folder created before and will used the folder *results* to save all the output from the do-file. This output consists mainly of various excel files generated by the *outreg* command at various points throughout the do-file, but also of dta files with coefficient estimates that are used at later points in the do-file.

[1] These files are named *ComtradeBatch_01-02.csv*, *ComtradeBatch_03-04.csv*, *ComtradeBatch_05.csv*, *ComtradeBatch_06.csv*, *ComtradeBatch_07.csv*, *ComtradeBatch_08.csv*, *ComtradeBatch_09-10.csv*, *ComtradeBatch_62-65.csv*, *ComtradeBatch_70-74.csv*, *ComtradeBatch_75-78.csv*, *ComtradeBatch_79-82.csv*, *ComtradeBatch_83-86.csv*, *ComtradeBatch_87-90.csv*, *ComtradeBatch_94-96.csv*, and *ComtradeBatch_99-00.csv*. Unfortunately, we cannot provide these files, but we are happy to share them with anyone who can provide proof that they have legitimate access to Comtrade.

After setting the global variables, the rest of the do-file can be run to reproduce all the results in the paper, or run in steps to reproduce the results in footnote 12, Table 1 and Figure 1 of the paper, as well as Table B1 of the online appendix.

The different outputs can be obtained by running the following parts of the code.

- Lines 19 to 64 re-estimate the key specification from Baier and Bergstrand (2007, JIE) that is referred to in Section 3 of the paper (as well as in footnote 12). Before running these lines, the global variable *zeros* must be set to *"no"* in line 12 to achieve comparability with B&B (recall that these authors use log trade flows, so have to drop observations with zero trade flows). Setting *$zeros ="no"* ensures that the do-file uses the data from step 1 that do not include zeros.

- Lines 71 to 273 reproduce the results from Table 1. Note that before running the actual regressions from line 120 onwards, the global variable *$zero* is to be set back to *"yes"* on line 71, so that the fully rectangularized dataset is used; one then needs to run lines 94-98 to drop sectors with fewer than 2,000 positive observations (see footnote 10 in the paper for a discussion of this point). Also note that some of the following regressions are quite computationally intensive (especially the PPML specifications with large sets of fixed effects and disaggregate data), so that it is also possible to only reproduce selected results by changing the inputs into the foreach loops on lines 120-123. The *foreach* loop setting the local variable *method* specifies whether OLS or PPML estimation is used, the *foreach* loop *level* indicates at which level of aggregation results will be produced: 0 corresponds to 0-digit or aggregate bilateral trade, 2 indicates the use of bilateral trade data aggregated to the 2-digit level, and 4 indicates the use of bilateral trade data aggregated to the 4-digit level (in Table 1 in the paper, we report results for the 0, 2 and 4-digit level). Finally, the meaning of the *fe* loop in line 122 is as follows:
    - *imp_exp_bilateral*: exporter-year, importer-year and bilateral fixed effects are used; note that this corresponds to the entries in Table 1 without sector-level fixed effects. For example, setting *method* to *ppml*, *fe* to *imp_exp_bilateral* and *level* to *4*, produces the results from the third line of the PPML panel of Table 1. Likewise, setting *method* to *ols*, *fe* to *imp_exp_bilateral* and *level* to *4* produces the results from the third line of the OLS panel of Table 1.
    - *all*: the full set of exporter-product-year, importer-product-year and exporter-importer-product fixed effects are used. Note that this corresponds to the entries in Table 1 with sector-level fixed effects. For example, setting *method* to *ppml*, *fe* to *all* and *level* to *4* produces the results from the last line of Table 1.

- Lines 280-385 generate the sector-level estimates summarised in Figure 1 in the paper. Note that the code will produce all the results needed for Figure 1 (OLS and PPML results at the 2- and 4-digit level). However, one can also produce only a subset of these results by changing the first line of the *foreach* loops on lines 290-291.

- Lines 389-463 produce the kernel density plots displayed in Figure 1 in the paper; the graph is saved in the file *Figure 1.gph* in the *results* folder. For this, the code uses the files with coefficient estimates produced by lines 280-385 as described above. Note that the code also requires some of the estimates from Table 1 for one of the elements of the figure (the solid line indicating the coefficient estimate when imposing coefficient homogeneity). So, these files need to be generated first as described above.

- Finally, lines 470-615 generate the forecasts shown in Table B1 of the online appendix. Note that this part of the code again requires both the homogeneous and heterogeneous coefficient estimates generated earlier on, so the respective parts of the do-file need to be run first. Once lines 470-615 have been run, lines 618-633 append the estimates just generated to generate a table with treatment effects on which Table B1 of the online appendix is based and saves it in the file *Treatment Effects.dta* in the *results* folder.

3) The file *Aggregation_Cases1234_19March_2022.do* replicates the simulation results found in the Appendix.