Data and code README for

Search Frictions in International Good Markets

by Clémence Lenoir, Julien Martin and Isabelle Mejean

To be published in the *Journal of the European Economic Association*

0. Description

This readme file lists the data and code associated with the paper "Search frictions in international good markets." The main dataset on firm-to-firm trade used in the paper is based on administrative data and, as a consequence, cannot be shared. In Bergounhon et al (2018), we explain how to get granted access to similar data by applying to the French *Comité du Secret.* We make available all the other datasets to replicate the results in the paper.

This readme file is organized as follow:

- Section 1 describes the computational environment
- Section 2 presents the replication instructions
- Section 3 describes the administrative data used in the paper and how to get granted access
- Section 4 describes the raw datafiles
- Section 5 lists the datafile that we can make publicly available
- Section 6 describes the Stata codes used to generate the estimation sample
- Section 7 describes the Matlab codes used to estimate search frictions, run the simulations and counterfactuals
- Section 8 describes the Stata codes used to compute summary statistics on search frictions and analyze their distortive impact

Should you have any questions regarding the data or the codes, please contact Isabelle Mejean at isabelle.mejean@sciencespo.fr or Julien Martin at martin.julien@uqam.ca.

1. Computational environment

The following computational environments were used to produce the results in the paper. All STATA and Matlab outputs were computed on a server running Windows Server 2016 Standard (64-bit).

The following software was used:

- Stata/MP 14.1 (64-bit) with additional packages

- Matlab R2018b

2. Replication instructions

- Define your own working paths in DefineGlobal.do. The baseline uses 2007 as reference and thus the "year" global is set to 2007.
- Organize your folders:
  - Save do files in the $dopath folder
  - Save external data in the $sourcepath folder
  - Save firm-level data in the corresponding $FirmDataPath folders
  - Create the $graphpath $outputpath and $datamatlab folders
  - The $datamatlab folder should have three sub-folders, "FIG", "Data" and "results_lambda"
- Obtain the intra-EU trade data and save them in your $customsdata folder
- Run the file `Master_LMM2022_PrepareData.do` in Stata
- Run `Master_LMM2022_Estimation.m` in Matlab
- Run `Master_LMM2022_PostEstimation.do` in Stata. Note that this program includes shells that runs `model_fit_byhs6.m, figure7.m and figure8.m` in Matlab

3. Required firm-level administrative data

The analysis mostly exploits the firm-to-firm intra-EU export data (DEB files) collected by the French customs (DGDDI). These data are used in motivating stylized facts and the structural estimation. These data are complemented with two datasets:

- The firm-level export data (DEB+DAU files) collected by the French customs (DGDDI)
- The balance-sheet data produced by INSEE based on data from the tax authority (DGFiP). We use the BRN files for firms in the normal and the simplified regimes.

Finally, Figure 4 correlates estimated frictions with a measure of the prevalence of intrafirm trade that is recovered from the EIIG (Échanges Internationaux Intra Groupe) survey run by INSEE. The EIIG survey provides a detailed geographical breakdown of the trade value of French firms at the product level (HS4) and their sourcing modes – arm's-length trade or intra-firm trade in 1999.
All firm-level datasets can be made available to researchers by the French "Comité du Secret". The procedure to access the data is described here. Once granted access, the research team can access the data through a secured data hub called CASD. Access to the data is restricted to researchers located within the European Union.

4. Data files (raw)

Trade at the product*country level

- `BACI_HS02_Y2007_V202102.csv` [publicly available]
  Contains the baci dataset for 2007, using the HS2007 nomenclature
  Source: CEPII

- `country_codes_V202102.csv` [publicly available]
  Contains correspondence between the country codes in baci and ISO3, ISO2 and the names of countries
  Source: CEPII

Sirene

- `StockUniteLegale_utf8.csv` [publicly available data]
  Contains some characteristics of firms recovered from INSEE
  Source: INSEE- Sirene

Country-specific variables

- `Dist_cepii.dta` [publicly available data]
  Distance data from CEPII
  Source: CEPII

- `GDP.xls` [publicly available data]
  GDP data from The World Bank
  Source: World Bank

- `GDPperCapita.xls` [publicly available data]
  GDP per capita data from The World Bank
  Source: World Bank

- `Ling_web_raw.dta` [publicly available data]
  Language proximity from Melitz & Toubal (JIE, 2014)
  Source: CEPII

- `UN_MigrantStockByOriginAndDestination_20002019.xlsx`
  Stock of migrants by origin and destination countries
  Source: United Nations

- `WDIPopulationData.xlsx`
  Population by countries, over time
  Source: World Bank

- `WGI_data.csv`
  Country-specific governance indicators
  Source: World Bank

- `sci_stroebeljie.dta`
  Index of social connectedness from Stroebel (JIE, 2021)
  Source: Johannes Stroebel's webpage

- hief_data.dta
  Historical Index of Ethnic Fractionalization Dataset (HIEF)from Drazanova (2019).
  Source: Harvard Dataverse

- gini_eurostat_data.dta
  Gini coefficient of equivalised disposable income - EU-SILC survey
  Source: Eurostat ILC DI12

Country-sector variables

- WIOT2007_October16_ROW.dta
  WIOD table used to construct absorption at the country sector level
  Source: WIOD

Correspondance files

- Corres_NAF2-NAF1.dta
  Correspondance between revision 2 and revision 1 of the French NAF nomenclature
  Source: INSEE

- HS2007Description.txt
  Product names for the hs2007 nomenclature
  Source: Eurostat

- Corres_hs2007_2002.dta
  Correspondance from hs2007 to hs2002
  Source: Eurostat

Product-level variables

- product_characteristics_hs6_2002.dta
  This Stata file contains various product-level characteristics recovered from several sources:
  - *Correspondence between hs6 (2002 version) and SITC sectors by UNSTAT*
  - *Rauch (1999) classification of products (conservative and liberal versions)*
  - *Nunn (2007) measures of relationship-specificity*
  - *Antras et al. (2012) measure of upstreamness*
  - *Imbs & Mejean (2015) estimates of substitution elasticities*

- ladder_hs10_rep.dta
  This Stata file contains estimates of quality ladder from Khandelwal (2010)

Source: Amit Khandelwal's website

5. Datafiles publicly available

In addition to the raw datafiles, we make available the dataset of estimated coefficients, EstimatedSearchParameters_LMM2022.dta. The dataset contains the following variables:

- hs6 product code

- iso2 country code

- lambda estimated meeting probability

- sd_lambda estimated standard deviation

- B_maxB number of buyers

- proba_nomatch probability of a buyer meeting with zero buyer in the destination

6. Stata code that generates the estimation sample

`Master_LMM2022_PrepareData.do`

- Master dofile that prepares the data used in the structural estimation
- 0. Defines the working paths and all global variables used in the code in DefineGlobal.do
- 1. Adds additional functions and packages, id_group.do
- 2. Prepares external data
- 3. Prepares firm-level datasets (some based on confidential data)
- 4. Merges the firm-level and the country-level data and construct the final dataset
- 5. Summary statistics in Figure 1 and Table 1
- 6. Constructs the datasets for the structural estimation

2. Prepares external data

`baci.do`

- Uses data from CEPII-BACI to compute the market share of France in each EU country, by hs6 product.

`Prep_controls.do`

- Country-specific variables used in a gravity equation
- Uses
  - distance data from CEPII
  - GDP and GDP per capita data from the World Bank
  - A measure of language proximity from Melitz & Toubal (JIE, 2014)
  - A measure of migrant networks using UN data on stock of migrants per country of origin and country of destination (calls ImportPanelUN.do)

  Note: These data are available every five year so we systematically choose the latest data
  - An indicator for the probability of sharing the same language

  - Data on number of French-speaking citizens in each destination
  - Data on the World Bank's World Governance Indicator

`RCA.do`

- Estimates revealed comparative advantage for France using the method in Costinot et al (2012) and BACI data from 2007

3. Prepares firm-level

`recup_APE_repertoireSiren.do`

- Uses data from INSEE-Sirene to recover information on firms' sector of activity

`customs_world.do`

- Takes data for 2007 and aggregate by firm, year and product

`brnrsi_panel.do`

- Constructs a panel of balance-sheet information
- Fills-in the information for missing values within a firm-level time-series for sector (ape), location (codep) and employment, value added and turnover. For the former three variables, we take the average between the values observed in t-1 and t+1 whenever the value for period t is missing.

`F2FData.do`

- Starts from the raw F2F data for 2007
- Cleans the data:
  - Remove firms which siren is not associated with a legal unit in France,
  - Remove importers which VAT number is considered valid by the French customs
  - Recover country code for the importer based on its VAT number, when the information is missing from the original variable
- Collapses data at the F2F-product-year level
- Merges with balance-sheet data
- Labels variables

`intrafirm.do`

- Uses 1999 data on intra-firm trade as an input
- Measures the share of intra-firm trade by product and country*product pair

4. Merge firm-level and country-level data and construct final dataset

`gendataset.do`

- Merges F2F data with Sirene dataset to fill in missing information on sectors and employment
- Saves summary statistics on the data coverage
- Removes non-core products of each firm
- Merges with all country-level data
- Aggregates across nc8 within a 6-digit HS product

5. Summary statistics

`Figure1.do`

- Produces figure 1 on the distribution of sellers' outdegrees (section 2.2)
- Uses degree.do and strength_distribution.do

`FigureA1.do`

- Produces figure A.1 on the distribution of buyers' indegrees

`Table1.do`

- Produces table 1, gravity equation at the product and firm level (section 2.2)
- Produces table A.1 (gravity equation with two alternative controls for search / information frictions)

6. Construct the datasets for the structural estimation

`count_producers.do`

- Creates a table for number of French producers at the product level
- The table combines information on trade at firm-level and number of firms at sector level (from balance-sheet). Potential suppliers of a product include all the firms from sectors that the firms exporting the product belong to.
- The do file also produces data on the firms' labor productivity and how they are positioned in the overall and sectoral distribution of labor productivities

`production_trade_wiod.do`

- Uploads the WIOD dataset for 2007 and

`import_absorption.do`

- Combines the WIOD and Baci data to compute the share of French products in absorption, by country and product
- The share of French products in absorption is calculated as the product of foreign goods in absorption which is defined at the sectoral level from WIOD and the share of French products in imports, taken from BACI
- The share of French products in absorption, together with the number of active buyers in the destination are used to compute a proxy for the total number of buyers in a destination, by product

`tailored_moments.do`

- Calculates the empirical moment to be used in the structural estimation (variance of M1, M2 and M3 ratios)

`Prep_Matlab.do`

- Exports the moments to csv files, country by country (Matlab file will then be run in parallel across countries)

- The program creates two types of csv files, one containing the moments, by product and country, one containing the firm-level data necessary to compute the variance-covariance matrix of the estimator

7. Matlab code that estimates search frictions

The Matlab code can either be launched in Stata using a shell (Master_LMM2022_Estimation.do) or directly in Matlab. In this paragraph, we describe the Matlab code. Before you start, all .m files need to be saved in the same folder, which also contains a "Data" folder in which the csv files from the `Master_LMM2022_PrepareData.do` program are saved and a "results_lambda" folder in which the estimated coefficients will be saved.

Master_LMM2022_Estimation.m

- Master m file that uploads the data and run the structural estimation (in parallel)
- 0. Defines the directory, the parameters of the algorithm and the country coverage
- 1. Uploads the data country by country
- 2. Organizes the data so that they can use in a parallelized loop that is run over country*product pairs
- 3. Launches the estimation
- 3.1 Uploads the raw data and compute the matrix of weights based on the empirical moments
- 3.2 Solves the algorithm over a grid to select the initial values (`Objectiv_function_grid_typeX.m`)
- 3.3 Solves the algorithm based on the previous step's initial values with the weight matrix calibrated from the firm-level data (`Objectiv_function1_typeX.m`)
- 3.4 Solves the algorithm based on the same initial values but with the optimal weight matrix (`Objectiv_function2_typeX.m`)
- 3.5 Computes the standard errors (`Var_est_typeX.m`)
- 4. Stores the results
- Note: All steps of the algorithm are defined for type1, type2 or type3 which correspond to the three possible definitions of the empirical moment
- Note2: The objective functions are written in a flexible way so that it is possible to estimate a single (constrained) coefficient on a pooled sample of products. In LMM, we estimate one coefficient per product, without any pooling.

3.2 Solve the algorithm over a grid to select the initial values

`Objectiv_function_grid_typeX.m`

Run twice, over a broad and narrower range of possible solutions.

Computes the objective function over each possible solution and pick the one that minimizes the output

The objective function is defined over the distance between the empirical and theoretical moments. The theoretical moment is defined in `var_tX.m.` It is the variance of the ratios of h(M) over h(1), over the three values of M that define typeX. These are defined in

`Vect_cum_tX.m` which output is the vector of the three ratios. `Vect_cum_tX.m` calls `functionP.m`, which simply uses the `betainc` function in Matlab for incomplete Beta functions.

3.3 Solve the algorithm based on the previous step's initial values with the weight matrix calibrated from the firm-level data

`Objectiv_function1_typeX.m`

Same steps as above but the minimization of the objective function uses the fmincon algorithm in Matlab using the output of the previous step as initial value.

The objective function is defined over the distance between the empirical and theoretical moments. The theoretical moment is defined in `var_tX.m.` It is the variance of the ratios of h(M) over h(1), over the three values of M that define typeX. These are defined in `Vect_cum_tX.m` which output is the vector of the three ratios. `Vect_cum_tX.m` calls `functionP.m`, which simply uses the `betainc` function in Matlab for incomplete Beta functions.

3.4 Solve the algorithm based on the same initial values but with the optimal weight matrix

`Objectiv_function2_typeX.m`

Same steps as above except that now the weight matrix entering the objective function is the optimal one.

The objective function is defined over the distance between the empirical and theoretical moments. The theoretical moment is defined in `var_tX.m.` It is the variance of the ratios of h(M) over h(1), over the three values of M that define typeX. These are defined in `Vect_cum_tX.m` which output is the vector of the three ratios. `Vect_cum_tX.m` calls `functionP.m`, which simply uses the `betainc` function in Matlab for incomplete Beta functions.

The matrix of optimal weights is defined in `dg_tX.m` which itself involves `functionH0.m` and `calculH.m` functions. The definition of the optimal weight matrix is available in section OA.2.3 of the online appendix.

3.5 Compute the standard errors

`Var_est_typeX.m`

This part of the code computes the variance of the estimator using the results in Gourieroux et al (1985) adapted to our context in section OA.2.3 of the online appendix.

The `Var_est_typeX.m` function involves two more functions, namely `dg_tX.m`

and d`Varlambda_tX.m`.

8. Stata and matlab codes that analyze the estimation results

`Master_LMM2022_PostEstimation.do`

- Master dofile for the post-estimation analysis
- 0. Defines the working paths and all global variables used in the code in DefineGlobal.do
- 1. Generates the dataset of estimated lambdas and come covariates
- 2. Tables and figures of main results
- 3. Analysis of the model fit
- 4. Analysis of the distortive impact of frictions

1. Generate the dataset of estimated lambdas

`gen_lambda.do`

Combines data on estimated lambda parameters with

- number of buyers to then compute no match probabilities
- trade shares
- various country-level controls
- number of French exporters

2. Tables and figures of main results

`Table2.do`

Computes the summary statistics in table 2 of the final manuscript

`Figure4.do`

Adds various country and product-level characteristics to the dataset of estimated lambdas and compute univariate and multivariate correlations between estimated no-match probabilities and these covariates

`Table3.do`

Merges data on export at country*product level with estimated lambdas to estimate the extended gravity equation in Table 3

3. Analyze the model fit

`model_fit.do`

- Combines data on estimated lambda together with i) information on the number of sellers and buyers per country*product, ii) information on the number of firms with 0 to ten buyers in the destination, iii) French market shares, iv) various measures of export premia recovered from the balance-sheet data.
- Note: export premia are recovered from the comparison of firms lying in the first half of their sectoral productivity distribution and firms in the top 20% of productivities. The export premium is either defined in terms of the export probability, the value of exports or the number of buyers per exporter
- Exports the data to csv

`model_fit_byhs6.m`

- Matlab file that computes the model fit
- Produces the empirical and theoretical CDFs in Figure 5

`Table4.do`

Imports empirical and theoretical moments from Matlab and computes the correlations in Table 4.

4. Analyze of the distortive impact of frictions

`Figure6.do`

Merges the estimated lambda coefficients and estimates of revealed comparative advantages and draws the correlation in Figure 6.

`Table5.do`

Merges the balance-sheet data with the firm-level export data and the estimated lambda coefficients to estimate the model in Table 5.

`figure7.m`

(Performed from Stata in a shell)

Compares the predicted export performances of firms along the distribution of productivities, in the data and in a counterfactual world with reduced search frictions.

`figure8.do`

(Performed from Stata in a shell)

Compares the interquartile range of export performances of firms, in the data and in a counterfactual world with reduced search frictions.