# Navigating International Taxation: The Effects of a Carbon Levy on Shipping[[1]](#footnote-1), Vianney Dequiedt, Audrey-Anne de Ubeda, Edouard Mien

Data replication read-me file (Repository: https://github.com/edmien/Navigating-International-Taxation-Replication)

For additional information and questions, feel free to contact edouard.mien@ferdi.fr

## Overview

The code in this replication package constructs the analysis file from several data sources (listed in the table below) using both Stata and R.

5 main files run all of the code to generate the data for 8 figures (labelled from 1 to 5 and from C1 to C2) and 12 tables (labelled from 1 to 7, A5, and from C1 to C4) used in the paper. Based on a hardware of 15.6Go of RAM, and using Stata 18 and R 4.3.2, the replicator should expect the code to run for between 2h30 half and 3 hours.

Before proceeding to the replications, the replicator is expected to chose a local working directory where all data and do-files will be stored. In this working directory, 3 folders must be created and named as follows: “Raw data”, “Refined data”, and “Results”. Within the folder “Results”, 2 sub-folders must be created and named “Figures” and “Tables”. The main dataset used here (“Database.dta”) and the different do-files must be located in the specified working directory but not within any of these folders.

## Data Files

In this replication file, data are provided in three different folders. The folder “Raw data” include all raw datasets exactly as they were downloaded from their sources. The folder “Refined data” include datasets processed by the authors based on the datasets included in “Raw data”. They are all constructed by running the do-file “01\_ Construction.do”. All datasets included in the folder “Refined data” are then merged into the core “Database.dta”. Finally, the folder “Result” includes both the datasets obtained after running the main estimations and the graphs and tables of results included in this article.

### Description and Provenance of Raw Data

This paper uses a combination of data sources. Access to the United States Census Bureau data is free but requires prior registration. All other data are publicly available and do not require any prior registration. The file “HS-Level Products.xlsx” is made by the authors and should not be downloaded contrary to all other datasets. This dataset only contains the name of each HS2-digit products as they will appear in the tables of results.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Dataset** | **Public?** | **Where and how to access** | **Last download** | **Reference** |
| “State Exports by HS Commodities.csv” | Yes, after registering | Register: <https://usatrade.census.gov/index.php>  Select: “State Export Data > HS”  Select the following criteria:   * Measures: “Total Value ($US)”, and “Vessel Value ($US)” * State: “All States” (only) * Commodity: All 6-digit * Country: “World Total”, “Canada”, and “Mexico” * Time: “2012”, “2013”, “2014”, “2015”, “2016”, “2017”, “2018”, and “2019”   Select: “Semicolon-delimited ASCII format .csv” | 10/18/24 | U.S. Census Bureau |
| “State Imports by HS Commodities.csv” | Yes, after registering | Register: <https://usatrade.census.gov/index.php>  Select: “State Import Data > HS”  Select the following criteria:   * Measures => “Total Value ($US)”, and “Vessel Value ($US)” * State => “All States” (only) * Commodity => Select all 6-digit * Country => “World Total”, “Canada”, and “Mexico” * Time => “2012”, “2013”, “2014”, “2015”, “2016”, “2017”, “2018”, and “2019”   Select “Semicolon-delimited ASCII format .csv” | 10/18/24 | U.S. Census Bureau |
| “Eurostat.xlsx” | Yes | Link:  <https://ec.europa.eu/eurostat/databrowser/view/ds-058213/legacyMultiFreq/table?lang=en&category=ext_go.ext_go_detail>  Select the following criteria:   * Flow => “IMPORT” and “EXPORT” * Frequency => “Annual” * Indicator => “VALUE\_IN\_EUROS” * Partner => “[EU\_EXTRA]”, “AD”, “AL”, “BA”, “BY”, “CH”, “LI”, “MD”, “ME”, “MK”, “NO”, “RU”, “TR”, “UA”, and “XS” * Product => All HS6-digit level products * Reporter => “[EU]” * Time Period => “2012”, “2013”, “2014”, “2015”, “2016”, “2017”, “2018”, and “2019” * Transport Mode => “Sea”, “Rail”, “Road”, and “Air”   Then select the following organization criteria:   * Row => PRODUCT * Column => PARTNER then TIME * Page => FLOW   Remark: due size limitations, the dataset might require to be downloaded in several sub-datasets and then manually merged into one. | 10/17/24 | Eurostat |
| “Exchange\_Rates\_incl\_Effective\_Ex\_Rat.xlsx” | Yes | Link:  <https://data.imf.org/regular.aspx?key=61545850>  Select “Annual” and then the following criteria:   * Country = > “Euro Area” * Date => “2012”, “2013”, “2014”, “2015”, “2016”, “2017”, “2018”, and “2019” * Variable => “Domestic Currency per U.S. Dollar, Period Average” | 10/29/24 | International Financial Statistics (IMF) |
| “BACI\_HS12\_Y2012\_V202401b.csv”  “BACI\_HS12\_Y2013\_V202401b.csv”  “BACI\_HS12\_Y2014\_V202401b.csv”  “BACI\_HS12\_Y2015\_V202401b.csv”  “BACI\_HS12\_Y2016\_V202401b.csv”  “BACI\_HS12\_Y2017\_V202401b.csv”  “BACI\_HS12\_Y2018\_V202401b.csv”  “BACI\_HS12\_Y2019\_V202401b.csv”  “country\_codes\_V202401b.csv”  “product\_codes\_HS12\_V202401b.csv” | Yes | Link: <http://www.cepii.fr/CEPII/en/bdd_modele/bdd_modele_item.asp?id=37>  Select “Download > HS12” | 10/21/24 | Gaulier and Zignano (2010) |
| “Gravity\_V202211.csv” | Yes | Link: <http://www.cepii.fr/CEPII/en/bdd_modele/bdd_modele_item.asp?id=8>  Select “Download > Stata” | 06/18/24 | Conte et al. (2022) |
| “CERDI-seadistance.dta” | Yes | Link: <https://zenodo.org/records/46822#.VvFcNWMvyjp> | 12/05/23 | Bertoli et al. (2016) |
| “TPc\_V202401.dta” | Yes | Link: <http://www.cepii.fr/CEPII/en/bdd_modele/bdd_modele_item.asp?id=5>  Select “Download > TPc\_V202401” | 07/04/24 | De Sousa et al. (2012);  Mayer et al. (2023) |
| “elasticity\_for\_publication\_2021\_09\_29.dta” | Yes | Link:  <https://sites.google.com/view/product-level-trade-elasticity>  Select “Dataset HS 6-digit classification UPDATED version Sept 2021” | 02/21/24 | Fontagné et al. (2022) |
| “Fuel Price.csv” | Yes | Links:  <https://www.insee.fr/fr/statistiques/serie/010002073#Telechargement> => provides data for the maritime fuel price at 1% sulphur content  In each page, select:   * Beginning date: January 2012 * End date: December 2019 * Transpose the table: Data in column   After downloading, delete lines 1-4 and column C | 11/20/24 | INSEE |
| “World\_Development\_Indicators\_GDPPC.csv” | Yes | Link:  <https://databank.worldbank.org/source/world-development-indicators>  Select:   * Database: World Development Indicators * Country: All countries * Series: “GDP per capita, PPP (current international $)” * Time: 2019   Select “Download options > Advanced options > Variable format > Codes only” | 10/22/24 | World Development Indicators (World Bank) |
| “2024-retrospective-review-official.xlsx” | Yes | Link:  <https://www.un.org/development/desa/dpad/least-developed-country-category/ldc-data-retrieval.html>  Select: “Time series estimates (LDC criteria) dataset (2002-2024)” | 11/28/24 | United Nations Department of Economic and Social Affairs |
| “world-administrative-boundaries.cpg”  “world-administrative-boundaries.dbf”  “world-administrative-boundaries.prj”  “world-administrative-boundaries.shp”  “world-administrative-boundaries.shx” | Yes | Link:  <https://public.opendatasoft.com/explore/dataset/world-administrative-boundaries/export/>  Select “Shapefile” | 06/20/24 |  |
| “UNSD Country Classification.csv” | Yes | Link:  <https://unstats.un.org/unsd/methodology/m49/overview> | 10/28/24 | United Nations Statistics Division |
| “HS2-Level Products.xlsx” |  | Authors | 11/22/24 | Authors |

Here is a short description of these datasets:

* “State Exports by HS Commodities.csv” contains data for all exports in value (current US$) from the United States to Canada, Mexico, and the rest of the world separately by product (at the HS6-digit level), year (all years between 2012 and 2019), and type of transportation (all flows and vessel flows only).
* “State Imports by HS Commodities.csv” contains data for all imports in value (current US$) of the United States from Canada, Mexico, and the rest of the world separately by product (at the HS6-digit level), year (all years between 2012 and 2019) and by type of transportation (all flows and vessel flows only).
* “Eurostat.xlsx” contains data for all exports and imports in value (current Euros) of the European Union to and from the rest of the World and with every neighbor of the European Union separated by product (at the HS6-digit level), year (all years between 2012 and 2019) and type of transportation (sea, rail, road, and air). Intra-European flows are not included.
* “Exchange\_Rates\_incl\_Effective\_Ex\_Rat.xlsx” contains annual data for the nominal bilateral exchange rate between the Euro and the U.S. dollar between 2012 and 2019. It is expressed in Euro per US$.
* “BACI\_HS12\_Y201X\_V202401b.csv” datasets contain data about all bilateral flows in quantity (metric tons) and value (thousands current US$) by pair of countries and products (at the HS6-digit level). Each dataset corresponds to a specific year (all years between 2012 and 2019).
* “country\_codes\_V202401b.csv” contains the table of equivalence between the country code number used in BACI datasets, the ISO3-digit country code used in most of our datasets to identify countries, and the full country name.
* “product\_codes\_HS12\_V202401b.csv” contains a table of equivalence between the product codes at the HS6-digit level and each product full name. This dataset is used for data processing.
* “UNSD Country Classification.csv” contains the regional area for each country (“Region”) and three binary variables indicating if the country is a Least Developed Country (“LDC”), a Landlocked Developing Country (“LLD”), and/or a Small Island Developing State (“SIDS”) according to the UN classification system.
* “HS-Level Products.xlsx” is a simple table of equivalence between each HS2-digit level product code and its full name as it will appear in the tables of results.
* “Gravity\_V202211.csv” contains variables that will enter the gravity model (common border, common language, former colonization relationships, free trade agreement, custom union, members of the European Union, members of the World Trade Organization) by pair of countries and year.
* “CERDI-seadistance.dta” contains (time-invariant) data about maritime distance, distance by air and the capital-to-port distance by road for each pair of countries.
* “TPc\_V202401.dta” contains data about all bilateral trade flows in value by pair of countries, years and types of industries (nine types of industries). Contrary to the BACI datasets, it includes intra-national trade flows (i.e. bilateral flows with the same exporter and importer).
* “elasticity\_for\_publication\_2021\_09\_29.dta” contains the tariff-based trade elasticities estimated by Fontagné et al. (2022) at the HS6-digit product level. It is used for the estimation of the structural gravity model of trade.
* “Fuel Price.csv” contains the annual average value of heavy fuel oil price in the port of Rotterdam (in US$ per metric ton).
* “World\_Development\_Indicators\_GDPPC.csv” contains national values of gross domestic product per capita (in current PPP US$) in 2019. It is used to estimate the relationship between the cost of maritime fuel taxation and domestic income per capita (Figure C4)
* “2024-retrospective-review-official.xlsx” contains various indexes from the United Nations, notably the Remotenes (REM) Index that we use in Table C3
* “world-administrative-boundaries” file contains countries coordinates and a map shapefile that will be used for the creation of Figure C3

### Description of Refined data

All refined datasets are created with the do-file “Step 1 - Creation of the Database” based on the raw datasets described above and saved in a .dta format. All variables of refined datasets are labelled to be easily understandable for the replicator. Most of them are then merged into the core “Database.dta”. Here is a short description of these datasets:

* “Country Codes.dta” contains the table of equivalence between the UN M49 country code number used in BACI datasets (“m49”), the ISO3-digit country code used in most of our datasets to identify countries (“Code”), and the full country name (“Country”), obtained by reformatting the “country\_codes\_V202301.csv” raw dataset. It also contains the regional area for each country (“Region”) and three binary variables for being a Least Developed Country (“LDC”), a Landlocked Developing Country (“LLD”), and/or a Small Island Developing State (“SIDS”) based on the “UNSD Country Classification.csv” raw dataset. The table of equivalence is primarily used to merge the different tables together (using the ISO-3 code as the reference) and to display the full country names in our tables of results (Tables 4, 5, 6, and C4), while the “Region” variable is used for the PPMl estimations and the “LDC”, “LLD”, and “SIDS” are used for estimating the welfare impact by categories of countries.
* “Elasticities.dta” contains tariff-based trade elasticities at the HS2-digit product level. It is constructed by aggregating HS6-digit level product elasticities from “elasticity\_for\_publication\_2021\_09\_29.dta” for each HS2-digit level product. For each HS2-digit Product, the main elasticities (denoted “sigma\_x”) is the average value of the elasticities of its HS6-digit level subproducts, while the two other elasticities (denoted “sigma\_x\_low” and “sigma\_x\_high”) correspond to the value of the 1st and 3rd quartile of its HS6-digit level subproducts.
* “Flows.dta” is an intermediary dataset used for the construction of the main “Database.dta” and contains all exporter-importer-product-year flows data (in value and in volume), as well as the explanatory variables that will enter the gravity model. In that base, all required “refined datasets” have already been merged and most required changes and reformatting have been made. However, in “Flows.dta”, all exporter-importer-product-year flows are distinguished between “total flows” (denoted Ytot) and “maritime flows” (denoted Y0, Y50, Y60, and Y70) for each existing product (numbered from 1 to 97 at the HS2-digit level). On the contrary, in “Database.dta”, the “total flows” variable has been dropped and all non-maritime flows have been grouped in a separate product (numbered 99). The dataset “Flows.dta” is used for Table 1 and Table 2.
* “Fuel Price.dta” contains the annual average value of heavy fuel oil price in the port of Rotterdam (in US$ per metric ton) in a Stata format. It is constructed by reformatting the “Fuel Price.csv” raw dataset.
* “GDPPC.dta” contains national values of gross domestic product per capita (PPP US$) in 2019. It is constructed by simply reformatting the dataset “World\_Development\_Indicators\_GDP.csv”.
* “REM.dta” contains the UNDESA Remoteness index. It is constructed by simply reformatting the dataset “2024-retrospective-review-official.xlsx”.
* “Gravity Variables.dta” contains variables that will enter the gravity model (common border, common language, former colonization relationships, free trade agreement, custom union, members of the European Union, members of the World Trade Organization) by pair of countries and year. It is constructed by simply reformatting the dataset “Gravity\_V202211.csv”.
* “HS6 Codes.dta” contains a table of equivalence between the product codes at the HS6-, HS4-, and HS2-digit levels and each (HS6-digit) product full name. It also stores the four binary variables indicating whether each product is mainly exported by sea or not (“D0”, “D50”, “D60”, and “D70”). This dataset is used for the construction of the main “Database.dta” and is obtained by reformatting the dataset “HS2-Level Products.xlsx” and merging it with “EU+US Trade.dta” dataset.
* “HS2 Codes.dta” contains a table of equivalence between each HS2-digit level product code and its full name as it will appear in the tables of results. It is constructed by simply reformatting the “HS-Level Products.xlsx” raw dataset.
* “NER.dta” contains annual data for the nominal bilateral exchange rate between the Euro and the U.S. dollar between 2012 and 2019. It is constructed by simply reformatting the dataset “Exchange\_Rates\_incl\_Effective\_Ex\_Rat.xlsx”
* “Seadistance.dta” contains (time-invariant) data about maritime distance, distance by air and the capital-to-port distance by road for each pair of countries. It is constructed by simply reformatting the dataset “CERDI-seadistance.dta”.
* “TradeProd Ratios.dta” contains the intranational-to-international trade ratios by country, year, and industry (nine categories of industries). These ratios are then multiplied with each HS2-digit level product to approximate the value of intranational trade by country, year, and product. This dataset is constructed using the dataset “TPc\_V202401.dta”.
* “US+EU Trade.dta” contains data about the USA and the European Union total and maritime trade flows (imports + exports) at the HS6-digit level with neighboring countries (“Neighbors”), the rest of the World (“ROW”), and the total of both (“World”). All flows are expressed in US dollars. EU Trade data do not include intra-EU trade flows. This dataset is used to determine the products that are the more likely to be transported by sea at the world level. Table 1 and Table 2 are drawn from it. It is constructed using “State Exports by HS Commodities.csv”, “State Imports by HS Commodities.csv” (US trade flows), “Eurostat.xlsx” (EU trade flows), and “NER.dta” (conversion of EU trade flows in US dollars)
* “Map Coordinates.dta” and “Map Countries.dta” are created using the “world-administrative-boundaries” file and contains country coordinates and a map shapefile used for the creation of Figure C3

## Computational requirements

Computations were made using both Stata and R Studio. More precisely, the do-files “01\_Construction.do”, “02\_Descriptive\_Statistics.do”, “03\_PPML\_Estimations.do”, and “05\_Final\_Results.do” were run with Stata 18, while the do-file “04\_Structural\_Gravity.Rmd” was run with R 4.3.2.

The following packages and libraries are required for the estimations:

* Stata (code was last run with version 18)
* For PPML estimations (“02\_PPML\_Estimations.do”):
  + reghdfe (version used: 6.12.3 08aug2023)
  + ftools (version used: 2.49.1 08aug2023)
  + ppmlhdfe (version used: 2.3.0 25feb2021)
* For the maps and presentation of graphs (05\_Final\_Results.do”)
  + spmap
  + shp2dta
  + mif2dta
  + palettes
  + colrspace
* R (code was last run with version 4.3.2) (“04\_Structural\_Gravity.Rmd”)
* For computations:
  + tidyr (version used: 1.3.0)
  + dplyr (version used: 1.1.4)
  + tensor (version used: 1.5)
* For the importation and exportation of data in .dta format:
  + haven (2.5.4)
  + foreign (0.8)

Code was last run on a 10-core Intel i5 based laptop with 16 Go of RAM and using Windows version 11. Using a similar computer machine, the replicator should expect the computations to take between 2h30 and 3 hours. More precisely, each of the following .do required:

* “01\_Construction.do” => approximately 25 minutes.
* “02\_Descriptive\_Statistics.do” => less than 2 minutes.
* “03\_PPML\_Estimations.do” => approximately 135 minutes.
* “04\_Structural\_Gravity.Rmd” => approximately 5 minutes.
* “05\_Final\_Results.do” => less than 2 minutes.

## Description of programs/code

All replication files are numbered from “01” to “05” and should be run in the given order. The replicator should run first “01\_Construction.do”, “02\_Descriptive\_Statistics.do” and “03\_PPML\_Estimations.do” with Stata, then “04\_Structural\_Gravity.Rmd” with R, and finally “05\_Final\_Results.do” with Stata again. The file “00\_Packages.do” only aims at downloading all required Stata packages. It must be run before the first replication (to ensure all packages are correctly installed) but can be skipped after that.

* “01\_Construction.do” will extract and reformat all raw datasets and create the refined datasets described in the above section. It will then merge the main refined datasets into the core “Database.dta” that will be used for the econometric estimations.
* “02\_Descriptive\_Statistics.do” will generate the main figures and tables of descriptive statistics based on either the core “Database.dta” or on any datasets stored in the “Refined data” file. This corresponds to Table 1, Table 2, Table A5, Table C4, and Figure 1.
* “03\_PPML\_Estimations.do” will estimate the Poisson Pseudo-Maximum Likelihood estimators on the variables included in the main “Database.dta”. PPML estimations are made separately for each product at the HS2-digit level (97 products in the base). Results are then stored in the “Estimation\_Results.dta” and “Betas.dta” datasets. Finally, it will export the main results from PPML estimations that are Table 3, Table C1, Table C2, Figure 2, and Figure C1). This file requires the package *ppmlhdfe* from Correia et al. (2020), which itself requires the packages *ftools* and *reghdfe*.
* “04\_Structural\_Gravity.Rmd” will estimate the structural gravity model of trade using R and based on the partial elasticities estimated in the previous step. The new flows will then be exported in the dataset “Final Results.dta”. This file requires the packages *tidyr*, *dplyr*, and *tensor* for the computations, and the packages *haven* and *tensor* for the importation and exportation of data from and to Stata.
* “05\_Final\_Results.do” will finally compute and estimate all the main results of this article based on the new flows estimated in the previous step. It will then create Table 4, Table 5, Table 6, Table 7, Table C3, Figure 3, Figure 4, Figure 5, and Figure C2. It will also create another excel file titled “Main Results.xlsx” storing some key results used in this article and not directly included in the different Tables (average welfare loss and price increase, marginal cost of funds, linear regressions for welfare / price and GDP per capita…)

## List of tables and programs

The provided code reproduces the following tables and figures:

|  |  |  |
| --- | --- | --- |
| Figure/Table | Program | Lines |
| Table 1 | 02\_Descriptive\_Statistics.do | 53 - 125 |
| Table 2 | 02\_Descriptive\_Statistics.do | 131 - 159 |
| Table 3 | 03\_PPML\_Estimations.do | 152 - 187 |
| Table 4 | 05\_Final\_Results.do | 48 - 85 |
| Table 5 | 05\_Final\_Results.do | 129 - 166 |
| Table 6 | 05\_Final\_Results.do | 194 - 216 |
| Table 7 | 05\_Final\_Results.do | 313 - 438 |
| Table A5 | 02\_Descriptive\_Statistics.do | 165 - 169 |
| Table C1 | 03\_PPML\_Estimations.do | 99 - 124 |
| Table C2 | 03\_PPML\_Estimations.do | 152 - 187 |
| Table C3 | 05\_Final\_Results.do | 235 - 296 |
| Table C4 | 02\_Descriptive\_Statistics.do | 175 - 234 |
| Figure 1 | 02\_Descriptive\_Statistics.do | 17 - 47 |
| Figure 2 | 03\_PPML\_Estimations.do | 131 - 145 |
| Figure 3 | 05\_Final\_Results.do | 91 - 108 |
| Figure 4 | 05\_Final\_Results.do | 172 - 188 |
| Figure 5 | 05\_Final\_Results.do | 235 - 296 |
| Figure C1 | 03\_PPML\_Estimations.do | 131 - 145 |
| Figure C2 | 05\_Final\_Results.do | 114 - 123 |

Tables A1, A2, A3, and A4 have been written manually by the authors and do not appear in these replication files.

Tables C5 and C6 present the main results that can be found in Table 4, Table 5, Table 7, Equation (4), and Equation (5) under the different scenarios presented in the last paragraph of section 7, i.e., with alternative elasticities or alternative criteria to isolate maritime trade flows. To obtain the results for each alternative scenario, it is necessary to run each time the code for Table 4, Table 5, Table 7, Equation (4), and Equation (5) with the corresponding alternative inputs, alternative elasticities, or alternative trade flows.

## Acknowledgements

This README file has been written based on the README template of Vilhuber et al. (2022) available at <https://zenodo.org/records/7293838> and under a CC-BY-NC license.

## References

Bertoli M. Goujon M. and Santoni O. (2016) The CERDI-seadistance database, *CERDI Working Paper* 2016-07, CERDI.

Conte M., Cotterlaz P. and Mayer T. (2022) The CEPII Gravity Database. *CEPII Working Paper* 2022-05, CEPII.

Correia, S., Guimarães, P., & Zylkin, T. (2020). Fast Poisson estimation with high-dimensional fixed effects. The Stata Journal, 20(1), 95-115.

de Sousa J., Mayer, T. and Zignago, S. (2012) Market Access in Global and Regional Trade. *Regional Science and Urban Economics*, vol 42:6, pp 1037-1052.

Fontagné L., Guimbard H. and Orefice G. (2022) Product-Level Trade Elasticities*. Journal of International Economics*, vol 137

Gaulier G. and Zignago S. (2010) BACI: International Trade Database at the Product-Level. The 1994-2007 Version. *CEPII Working Paper* 2010-23, CEPII.

Mayer T., Santoni G. and Vicard, V. (2023) The CEPII Trade and Production Database, *CEPII Working Paper* 2023- 01, CEPII.

Vilhuber L., Connolly M., Koren M., Llull J. and Morrow P. (2022) A template README for social science replication packages (v1.1). *Social Science Data Editors*. https://doi.org/10.5281/zenodo.7293838

1. The authors are grateful to Issakha Thiam for his excellent research assistance. [↑](#footnote-ref-1)