CGPA WORKING PAPER
2025-27

The Granular Trade and Production Activities (GRANTPA) Database

Sebastien Bradley
Javier Flórez
Mario Larch
Yoto V. Yotov

CENTER FOR GLOBAL POLICY ANALYSIS
LeBow College of Business
Drexel University

The Granular Trade and Production Activities (GRANTPA) Database*

Sebastien Bradley Javier Flórez Mario Larch Yoto V. Yotov†

May 23, 2025

Abstract

This paper introduces the Granular Trade and Production Activities (GRANTPA) database, which covers international trade flows for 3,124 products and 247 countries over the period 1995-2019 as well as domestic trade flows and production data for the same number of products and years for a subset of 35 European economies. The original data sources that we employ are Eurostat's *Comext* and *Prodcom* databases. A gravity application delivers a large set of product-level 'home bias' estimates, which cannot be obtained without domestic trade flows. The average estimates on the standard gravity variables in our model (e.g., distance) are comparable to those from the related literature. However, our disaggregated estimates are very heterogeneous across products, thus highlighting the importance of our new database.

JEL codes: C81, F13, F14.
Keywords: Gravity Data, Structural Gravity, Domestic Trade Flows, Disaggregated Gravity Estimates, Home Bias Estimates.

*We are very grateful to the Eurostat team for research support and clarifications concerning *Prodcom* codes, combined nomenclature classifications, and complementary concordance files provided by Erlend Firth and Laia Guinovart. This paper expresses the authors' views solely, and Eurostat or any institution is not responsible for any errors. We are grateful to Andrew Bernard, Ilke Van Beveren, and Hylke Vandenbussche for providing an earlier version of this concordance and for sharing the corresponding codes. We also thank David Zenz for his guidance in retrieving the production of manufactured goods database. To access the GRANTPA database, please e-mail us at grantpadatabase@gmail.com.

†Contact Information: Bradley: School of Economics, Drexel University, e-mail: sbradley@drexel.edu; Flórez Mendoza: The Vienna Institute for International Economic Studies (wiiw), Vienna University of Economics and Business (WU), e-mail: mendoza@wiiw.ac.at; Larch: Department of Law, Business & Economics, University of Bayreuth, CEPII, CESifo, GEP, ifo Institute, e-mail: mario.larch@uni-bayreuth.de; Yotov: School of Economics, Drexel University, CESifo, ifo Institute, e-mail: yotov@drexel.edu

1 Introduction and Motivation

Many important trade policies (e.g., tariffs), as well as other policies that may impact international trade flows (e.g., technical barriers to trade (TBT), sanitary and phytosanitary standards (SPS), and maximum residue levels (MRL)), are designed and implemented at a very disaggregated level (e.g., 6-digit HS level, which covers more than 5,000 categories). Moreover, much of the current policy debate is on the effects of non-discriminatory policies that are country-specific by definition (e.g., TBT, SPS, MRL). However, due to their country-specific nature, the effects of such policies cannot be identified in a properly specified econometric model that only includes international trade flows. A simple theory-consistent solution to identify the effect of non-discriminatory trade policies and of country-specific policies on international relative to domestic trade is to rely on domestic (along with international) trade flows (see Heid, Larch and Yotov (2021), Beverelli et al. (2023), and for a survey Yotov (2022)). However, available datasets that include international and domestic trade flows, such as USITC's ITPD-E or CEPII's TradeProd database, are relatively aggregated and do not match the product-level dimension of many trade policies.

To fill this gap, we capitalize on disaggregated data from Eurostat and we implement consistent concordance procedures to construct the GRANnular Trade and Production Activities (GRANTPA) database. Specifically, the sources that we relied on for the construction of the GRANTPA database are Eurostat's *Comext* database, which we used for the bilateral trade data, and Eurostat's *Prodcom* database, which we relied on for the production data. Both of these datasets are collected and maintained by Eurostat. We are indebted to experts at Eurostat, whose guidance was instrumental in using their bulk download facilities to obtain the raw trade and production datasets.

Despite the genuine intent for the European international trade and production classifications (from *Comext* and *Prodcom*, respectively) to be internally consistent over time and also consistent with each other, each of the two databases and corresponding classifications have gone through several changes over time, and many of these changes were specific to each database and independent from each other (both over time and between *Comext* and *Prodcom*). Thus, our first and most demanding task was to construct separate internally consistent concordances for the international trade data and production data over time, and combine these into a single concordance. To this end, we capitalized on previous work by Van Beveren, Bernard and Vandenbussche (2012) and Pierce and Schott (2012*a*,*b*) and extended this to a broad set of countries (without the use of firm-level data).

The construction of the GRANTPA database necessitated four additional steps. First, we cleaned and prepared the raw trade and production datasets by eliminating duplicate observations and taking full advantage of the raw data (e.g., by using reported export values to replace corresponding missing import values). Second, we applied our new, consistent concordance to the international trade and production data. Third, we used the bilateral trade data to construct total exports at the country level for each product, country, and year in the data, and we combined these total exports with the corresponding production data to construct domestic trade as the difference between production and export values at the product-country-year level. Finally, we combined the bilateral trade flow data with the domestic trade data to construct the GRANTPA database.

The GRANTPA database covers international trade data for 3,124 products and 247 countries over the period 1995-2019, along with production and domestic trade data for the same number of products and years for 35 European economies, including the 28 EU member states plus Norway, Iceland, Turkey, Montenegro, Bosnia and Herzegovina, Serbia, and Macedonia.[1] To help users who may want to limit their sample and corresponding analysis to only countries for which there is consistent international and domestic trade data, the GRANTPA database includes a 'flag' variable to denote the country-year combinations

[1] Given the period of investigation, the United Kingdom is included as an EU member. The GRANTPA database features data on international trade flows for a much broader set of countries but does not include data on their domestic trade.

for which domestic trade data are available.[2]

We demonstrate the usefulness of the GRANTPA database with an application to the workhorse model of trade—the gravity model. Specifically, we obtain estimates of several standard gravity variables including distance, contiguity, common language, and international borders, and we benchmark our results against the large set of existing gravity estimates from the related literature. We draw three main conclusions about the usefulness of the GRANTPA database for gravity analysis.

First, the average estimates of the gravity variables that we obtain are comparable to the gravity estimates from the existing literature. Second, while it is possible to obtain gravity estimates of the effects of distance, contiguity, and common language with datasets that only include international trade, our 'home bias' effects can only be obtained with the use of domestic trade flows data, highlighting this important dimension of our data. The home bias estimates that we obtain are large, positive, and statistically significant, which is consistent with the existing literature. However, we are not aware of 'home bias' estimates at such a disaggregated level. Finally, the disaggregated estimates on all gravity variables in our model vary significantly across the products in the GRANTPA database. The implication is that more aggregated gravity analysis may mask significant heterogeneity, which may be important from a policy perspective. Accordingly, we see value in using the GRANTPA database to analyze the effects of various bilateral and country-specific policies.

The rest of the paper is structured as follows. Section 2 summarizes our concordance procedures, describes the raw data, and highlights key features of the resulting GRANTPA database. Section 3 provides a proof of concept by employing the GRANTPA database to obtain benchmark gravity estimates. Section 4 offers concluding remarks and points to directions for possible uses and improvements of the database. A Technical Appendix

[2]We end up with 3,124 products, i.e., less than the over 5000 6-digit HS categories, because the trade and production data are recorded using different product classification codes that do not fully correspond, i.e., not all product codes in the international trade data have a correspondence to product codes in the production data and vice versa. However, we also provide a version of the GRANTPA database that corresponds to the 6-digit HS level.

includes all the detailed steps that we took to construct the GRANTPA database.

2 Sources and Methods

This section briefly describes the methods used to construct and prepare the trade and production data used for the GRANTPA database (Subsection 2.1), the sources for the raw data that we use (Subsection 2.2), and the steps taken to combine the raw databases for constructing domestic trade and combining it with the international trade data (Subsection 2.3). Subsection 2.3 concludes with a description of the main features, dimensions, and limitations of the GRANTPA database. For a detailed description of the data sources and the steps taken to construct the GRANTPA database, we refer the reader to the Technical Appendix.

2.1 Concordance of Trade and Production Data

2.1.1 General Procedure

Despite the genuine intent of the European international trade and production classifications (from *Comext* and *Prodcom*, respectively) to be internally consistent over time and also consistent with each other, each of the two databases and corresponding classifications have gone through several changes over time, and many of these changes were specific to each database and independent from each other (both over time and between *Comext* and *Prodcom*). As a result, the most demanding task in the creation of the GRANTPA database was to construct a consistent concordance between the international trade and production datasets.

We note at the outset that we benefited from and expanded upon the methods from several related efforts to create such data concordances. Van Beveren, Bernard and Vandenbussche (2012) (henceforth, VBBV) focus on the implications of changing product classifications using Belgian firms. Even though our focus is broader (i.e., we aim to construct a consistent trade and production database for Europe), we benefited tremendously from the guidance and concordances created by VBBV.[3] In addition, we implemented some of the algorithms provided by Pierce and Schott (2012*a*,*b*) for concording the US Harmonized System codes over time and with the SIC/NAICS product classes and industries. Since the GRANTPA database includes many new/additional years (the concordances from VBBV end in 2010, whereas ours run through 2022) during which the underlying international and production databases and classifications have changed significantly, we had to address some new challenges (taking care of duplicates, including accession countries),[4] and we utilized new concordance files extracted from Eurostat.

Following the aforementioned studies, we constructed our new concordance in three broad steps. Each of these steps is described in detail in the Technical Appendix.

1. First, we need internal consistency of the international trade data (from *Comext*) over time. The European international trade data at the product level from *Comext* is recorded according to the *8-digit Combined Nomenclature* (CN8) classification. Due to various changes in the product classification and the composition of products in the trade data (e.g., some products disappear while new products emerge),[5] to have a consistent international trade data, we need to construct a common and consistent concordance for the products over time, which is labeled CN8+. Therefore, the first step to ensure internal consistency over time is to create a variable that identifies each 'family' of codes, i.e., codes that are connected over consecutive years. Table 2 reports the number of obsolete and new codes in each year, the number of families, and the number of simple changes for the years in the trade data. For the many-to-many and one-to-many mappings between two years, we rely on a 'feedback' loop from Pierce and Schott (2012*b*). An additional challenge arises because some product codes change in

[3] The concordances from VBBV are available at:
https://sites.google.com/site/ilkevanbeveren/concordances.

[4] Tables 7 and 8 illustrate how we handle complex (many-to-many) mappings. Furthermore, Tables 5 and 6 in the Technical Appendix show examples of how we detect and remove duplicates in the production data.

[5] Table 1 illustrates the evolution of the 8-digit combined nomenclature classifications through 2022.

more than one year. Hence, we implement a procedure that ensures consistency over the whole coverage period of the database. The last step in the construction of the international trade data is to merge the concordance between CN8 and CN8+ with the trade data, taking care of the time variation and 'many-to-one', 'one-to-many', and 'many-to-many' cases, the latter making it necessary to aggregate/collapse the data.

2. Second, we need consistent production data over time. We use *Prodcom* ("*Production Communautaire*"), which is a system used in the EU to compile statistics on the production of manufactured goods in member states. Production activities are reported at the 8-digit *Prodcom* level (PC8) on a monthly basis. The *Prodcom* declaration includes data on the physical volume and value of production sold for each product during the survey period. As highlighted by VBBV, the *Prodcom* list changes over time and hence we exclude codes that cannot be consistently tracked over time in the concordance. Table 3 reports the number of obsolete and new codes, the number of families, and the number of simple changes in each year of the production data. Furthermore, we focus on mandatory 8-digit *Prodcom* codes. To avoid double counting, we use a concordance procedure that flags aggregate codes and drops the corresponding disaggregated codes to achieve consistency. We align the *Prodcom* codes with CPA6 and NACE 4 classifications. As with the trade data, we need to ensure consistency over time. Furthermore, we have to align the production data with the refined PC8+ classification.

3. Finally, we concord the international trade and production data. We use the PC8+ classification to bridge CN8 product codes (for international trade) and PC8 codes (for domestic production). As not all CN8 products are covered by the *Prodcom* list, we exclude them from the international trade data. On the other hand, some PC8 products are not covered by the CN8 classification. We drop some of these (industrial services) and recode others into their mandatory and aggregate counterparts to enable a PC8+

classification. Furthermore, we take care of the changing coverage of the *Prodcom* list over time. After this, we follow the concordance procedure outlined by VBBV which consists of several steps. The first step deals with concording product classifications within a single year. The remaining steps focus on the actual implementation of these concordances in the international trade and production data, ensuring consistency and accuracy in the alignment between *Comext* and *Prodcom*.

2.1.2 An Alternative HS6+ Concordance

The preceding procedure is deliberately intended to exploit the high degree of granularity and fine disaggregation afforded by the PC8+ classification. Nevertheless, many recent trade-related policies (e.g., SPS or TBT measures) are often implemented at the 6-digit Harmonized System (HS6) level. As such, we also develop an HS6+ concordance, which aggregates both trade (CN8) and production (PC8) data to a consistent 6-digit HS classification over time.[6] In order to ensure internal consistency, we limit this alternative HS6+ dataset to relatively straightforward code mappings (i.e., simple or many-to-one). This inevitably means dropping some codes that have one-to-many or many-to-many mappings, but the resulting HS6+ database remains sufficiently comprehensive for policy analysis while maintaining temporal consistency and data integrity.

The construction of an HS6+ concordance follows closely the methodology laid out in VBBV, mirroring the PC8+ procedure described above. All CN8 codes in the international trade data are truncated to HS6, and any codes that do not map into a valid HS6 category are dropped. For the production data, we take the PC8-based Prodcom data, drop optional or aggregated codes that complicate mapping, and merge the remaining mandatory codes into HS6+ identifiers. Code pairs that fail to match (i.e., those involving one-to-many or many-to-many mappings) are excluded to ensure a unique product classification. We then merge trade and production at the HS6+ level at the country-year level, ensuring that each

[6]The first 6 digits of the Combined Nomenclature correspond to the HS6 level.

HS6+ code is uniquely represented. Table 4 summarizes the primary differences between the PC8+ and HS6+ procedures including coverage, code structure, and how mappings are tracked over time.

The resulting HS6+ dataset involves two main compromises. First is the loss of granularity resulting from merging multiple PC8 codes into a single 6-digit code which may mask the finer product level heterogeneity that may be critical for specific research questions. Second is the aforementioned need to drop codes that do not fit neatly into a single HS6 category due to one-to-many or many-to-many mappings. Fortunately, the net effect on coverage therefore remains modest and the HS6+ dataset retains most of the underlying information on export, import, and production values. This alternative dataset thus allows direct integration with widely available 6-digit policy data without sacrificing the time consistency of trade and production flows.

2.2 Raw Data and Sources

In this section, we briefly describe the raw trade and production data and the sources that we used to retrieve it. We provide further details in the Technical Appendix. The original database for our international trade data is *Comext*, which we extracted using Eurostat's bulk download facility. *Comext* has several dimensions which we capitalize on in the construction of the GRANTPA database. First, it records international trade values for countries within the EU as well as between EU and non-EU member countries. The group of destination and origin countries in the intra-EU and extra-EU declaration has changed over time due to changes in EU membership.[7] In addition, *Comext* records two flows for each pair of countries, e.g., exports from Austria to Bulgaria and imports in Bulgaria from Austria. We utilize this feature by replacing some missing values in one direction of the trade flows with

[7] In 1995, Austria, Finland, and Sweden joined the EU. Then, in May 2004, ten new countries joined: Cyprus, Czech Republic, Estonia, Hungary, Latvia, Lithuania, Malta, Poland, Slovakia, and Slovenia. In 2007, Bulgaria and Romania joined. Lastly, Croatia joined the EU in 2013 and more recently the United Kingdom exited the EU in 2019. When a country joins the EU, its exports and imports vis-à-vis EU countries are no longer reported in the extra-EU trade statistics. Instead, they have to be declared in the intra-EU trade statistics.

the corresponding values in the other direction, which are not missing. Finally, *Comext* reports data on trade values and quantities. For the GRANTPA database, we use trade values. However, as described in the next section, we also use trade quantities to distinguish between missing values in the data versus true zero trade values.

Production data are only available for European countries. These data are downloadable from Eurostat's bulk download facility as '*Europroms*'. Due to changes in classifications over time, the production data must be downloaded separately for different periods. Similar to *Comext*, we capitalize on several dimensions of the *Prodcom* database. Specifically, in addition to the values of production at the product level, *Prodcom* reports quantities, which we use to distinguish between true zeroes versus missing production values when the latter values are not reported. In addition, *Prodcom* classifies some values as confidential (C:). Those have to be treated as missing values. Finally, *Prodcom* includes data on the value of total exports, which we use where available to construct domestic trade flows—i.e., the difference between total production and exports—for each country, year, and product in our sample.[8]

2.3 The GRANTPA Database: Construction and Coverage

In this subsection, we deploy the raw trade and production data that we described in Subsection 2.2, and we use the concordances that we created in Subsection 2.1 to construct the GRANTPA database. We proceed in 5 steps. We start with a description of the additional steps that we took to prepare the raw trade database (Step 1) and raw production database (Step 2) for merging them with each other. Then, in Step 3, we apply the concordances to the trade and production databases. In Step 4, we combine total product-level exports and production values for each country and year in our sample to construct domestic trade. Finally, in Step 5, we combine the international and domestic trade data to construct the GRANTPA database. We conclude with a description of the main

[8]To improve data coverage—as discussed below—we use total export data calculated from Comext.

features, dimensions, and limitations of the GRANTPA database, and we flag alternative subsets of the data based on country coverage. The Technical Appendix provides a more detailed description of these steps.

- ***Step 1: Prepare the international trade data.*** The international trade data from *Comext* has several dimensions (e.g., intra- versus extra-EU trade and imports versus exports), which we strive to exploit to the fullest in our processing of the raw data.

  First, we eliminate any duplicate observations, which could appear due to double reporting or because of the letter codes (T-,Q-,V-,E-). We focus on import values from *Comext* as our main trade variable. Thus, if *Comext* provides data on both exports from Germany to France and imports to France from Germany for a particular product, we use the reported imports as our measure of trade. The reason for selecting imports is that, by definition, the reported import values include trade costs (e.g., cost, insurance, and freight (CIF)). This is consistent with trade/gravity theory, which is derived at delivered prices. Furthermore, for non-intra-EU-trade, importer-reported data is typically considered to be more reliable, as they are based on custom statistics.

  Even though we rely on imports as our main variable, we also take advantage of some of the information contained in the reported export values. For example, if data on product-level imports for a given pair are not reported or zero, we replace these import values with their trading partner's corresponding (non-zero or non-missing) exports. Processing the *Comext* data in this manner yields an unbalanced product-level bilateral trade variable for trade between each of the European countries and all other countries in the world, including the European countries themselves.

- ***Step 2: Prepare the production data.*** We need the production data (in combination with total exports) to calculate domestic trade flows. To construct product-level production for the years and countries in our sample, we take advantage of several dimensions of the *Prodcom* database. Similar to the trade data, we first eliminate duplicate production-value observations, which could appear due to double reporting or because of the letter codes (T-,Q-,V-,E-). Then, we select the reported product-level values of production for each country and year as our main production value variable. We take advantage of the fact that *Prodcom* reports quantities as well as values by making sure that we treat missing production values as true missing values (as opposed to zeros) when the corresponding reported quantities are not missing. Finally, we also make sure that we treat missing production values as true missing values when *Prodcom* includes a corresponding code for confidential data. The outcome of this step is an unbalanced production value variable for each product, country, and year in our sample.

- ***Step 3: Add concordances to the trade and production data.*** We use the concordances described in Subsection 2.1 to make sure that the trade and production data are consistently classified. We apply the concordances to the bilateral trade data, the total exports data, and the production values data. As expected, due to the presence of one-to-many, many-to-one, and many-to-many combinations, we have to aggregate some of the product categories in each of the datasets, so that the product classification is unique and common across the trade and production datasets.

- ***Step 4: Construct domestic trade.*** In this step we combine the data on total product-level exports and product-level production values for each country and year in our sample, and we construct domestic sales as the difference between total production and total exports. For consistency, we first use the total exports data that is reported in *Prodcom*. In addition, we also use the *Comext* trade data to create total exports when those are missing in in *Prodcom*. To construct total exports based on *Comext*, we use positive bilateral import data to replace corresponding missing or zero bilateral export values.[9] Next, we sum all bilateral exports for each year and country. Importantly, we utilize the fact that *Comext* includes trade between each of the European economies and all other countries in the world.[10]

- ***Step 5: Construct the GRANTPA database.*** In this step, we construct the full and final version of the GRANTPA database. To this end, we combine the bilateral trade data that we constructed in Step 3 with the domestic trade data that we constructed in Step 4.

After all these steps, we arrive at our final product, the GRANTPA database. The GRANTPA database covers 247 countries and 3,124 products over the period 1995-2019. Table 4 of the Technical Appendix includes the list of all countries in the GRANTPA database, while a list of the products that are covered in the GRANTPA database, along with the corresponding codes from *Comext* and *Prodcom* can be found here. The GRANTPA database includes the following variables: (i) *'exporter'* is a 3-letter ISO code for each exporter; (ii) *'importer'* is a 3-letter ISO code for each importer; (iii) *'year'* captures the year when trade took place; and (iv) *'product'* is a numeric product code which is unique to the GRANTPA database. A complete concordance file, which includes product IDs, product names, and the corresponding codes from *Comext* and *Prodcom* can be found here. (v) *'trade'* denotes nominal (domestic and international) trade values in thousands of Euros.[11] Finally, (vi) *'flag'* is an indicator variable that takes a value of one for the countries (and years) for which the database includes domestic trade data and zero otherwise.

[9]Depending on the year, mirrored import values account for approximately 28-32% of the Comext data that we use to aggregate to the country-product-year level. In turn, the Comext observations account for only a fraction of the export data (i.e., the extra-EU portion) that are used to construct domestic trade. This will tend to modestly understate CIF-based trade values, but this is nevertheless superior to erroneously treating missing export values as zeroes.

[10]Similar to Bernard et al. (2018), we observe cases where exports exceed production. The percentage of negative domestic trade values thus ranges between 9% in 1995 to 22% in 2019. 'Carry Along Trade' may provide an explanation (Bernard et al., 2018), as could changes in inventories, or the presence of multinational enterprises and cross-border production networks. We replace all negative domestic trade with missing values.

[11]We note that, by construction, if we sum the values of trade (including domestic trade) for each exporter and year for which the GRANTPA database includes domestic trade, we will obtain the corresponding production values for each product, country, and year.

As discussed earlier, we also construct a version of the GRANTPA database at the HS6+ level, 1995-2019, which covers trade values at a time-consistent HS6 level for 2,406 "product" categories, some of which have been aggregated to `hs6_plus` to handle classification splits. Furthermore, the GRANTPA HS6+ alternative includes 244 reporting and partner countries. Within the data, 35 European economies report both international trade and domestic trade, identified by a dedicated variable "`flag`". Each observation is sorted by `year`, `exporter_iso3`, `importer_iso3` and a `product_id` that maps to a synthetic HS6+ code `product_over_time`, ensuring that products remain comparable over time despite nomenclature changes.[12] In addition, it includes the HS2 and HS4 equivalent codes to provide intermediate groupings of the HS classification, while an "`hs6_synthetic`" indicator distinguishes between "pure" HS6 categories and the "synthetic" codes that link split or merged lines generated from VBBV.

We conclude this section by discussing select limitations of the GRANTPA database and the alternative HS6+ database. First, both the main (PC8+) and HS6+ versions of the GRANTPA database cover bilateral trade between 35 European economies and all other countries in the world. However, it should be noted that (i) neither include trade between non-European countries, and (ii) the domestic trade data are available exclusively for European countries for which we have total trade and production data, i.e., no domestic trade data are available for non-European economies, and domestic trade data may even be missing for some European countries (e.g., late EU joiners). To help users who may want to limit their sample and corresponding analysis to only countries for which there is consistent international and domestic trade data, the GRANTPA database includes a 'flag' variable to denote the country-year combinations for which domestic trade data are available. Meanwhile, a limitation that is specific to the HS6+ version of GRANTPA is that it cannot/should not be aggregated by summing up individual HS6 categories as it does not provide comprehensive data coverage on account of excluding complex mappings. For more

[12]HS6+ correspondence and description tables are available here. Note, that this correspondence table includes also one-to-many and many-to-many codes.

details, please see the Technical Appendix.

3 Gravity with GRANTPA: A Proof of Concept

The objective of this section is to deploy our new GRANTPA database in an application as a proof of concept. To this end, we selected a 'gravity' application for two main reasons. First, the main motivation for constructing the GRANTPA database was that it could be employed for disaggregated gravity analysis at the product level. Second, the gravity model is the workhorse model of trade and, as such, it has been employed in thousands of papers that study various determinants of trade flows. Thus, we can rely on a large set of existing gravity estimates against which we can benchmark our new results to establish the representativeness and credibility of the GRANTPA database. We proceed in three steps. First, we combine the GRANTPA database with some existing gravity datasets. Then, we specify our estimating gravity model. Finally, we obtain and interpret our results.

The first gravity database that we combine with our GRANTPA database is the US International Trade Commission's *Dynamic Gravity Database* (DGD), which is created and maintained by Gurevich and Herman (2018). We use the DGD to obtain the covariates for bilateral distance and contiguity. In addition, we rely on the *The Domestic and International Common Language* (DICL) database of Gurevich et al. (2023) to obtain a variable for common language. We rely on the DICL dataset because it includes a continuous variable for common international language, which, as demonstrated by Gurevich et al. (2021), dominates the use of a dummy variable for common language, which is the standard approach in the literature. Finally, we construct a dummy variable that takes a value of one for domestic trade and is equal to zero otherwise. The estimates of this variable will reflect the effects of forces that drive a wedge between domestic and international trade, which are referred to in the literature as 'home bias' effects. Identifying such effects is not possible without the availability of domestic trade data—one of the core attributes of the GRANTPA database and one of the main motivations for its construction.

Capitalizing on some of the current gravity estimation techniques, as summarized by Larch, Shikher and Yotov (2025), we specify the following simple gravity model:

$$\begin{aligned} X_{ij,t}^k &= \exp[\gamma_1^k DIST_{ij} + \gamma_2^k CNTG_{ij} + \gamma_3^k LANG_{ij}] \times \\ &\quad \exp[\gamma_4^k SMCTRY_{ij} + \psi_{i,t}^k + \phi_{j,t}^k] \times \varepsilon_{ij,t}^k, \qquad \forall i, j. \end{aligned} \tag{1}$$

Here, $X_{ij,t}^k$ denotes the nominal exports (at delivered prices) of product k from exporter i to destination j at time t.[13] Following Santos Silva and Tenreyro (2006), we estimate equation (1) using the PPML estimator,[14] which accounts for potential heteroskedasticity issues inherent to trade data and enables us to take advantage of the information that is contained in the zero trade flows in the GRANTPA database. The gravity covariates in equation (1) include the logarithm of the bilateral distance between the trading partners ($DIST_{ij}$) and indicator variables for the presence of contiguous borders ($CNTG_{ij}$), common official language ($LANG_{ij}$), and domestic vs. international trade ($SMCTRY_{ij}$). Finally, following the literature, we cluster the standard errors by country-pair.

We rely on specification (1) to obtain a set of gravity estimates for each of the 3,124 products in the GRANTPA database.[15] Due to the large number of estimates, we report them, along with their corresponding confidence intervals, in Figure 1. For clarity of exposition (due to the presence of outliers), we do not include the largest and smallest five percent of point estimates for each of the gravity variables in our model. In addition, we

[13] The disaggregated estimating gravity equation (1) has solid theoretical foundations on the demand side, e.g., Anderson and van Wincoop (2004), and on the supply side, e.g., Shikher (2011) and Costinot, Donaldson and Komunjer (2012). Following Arkolakis, Costinot and Rodríguez-Clare (2012), Yotov et al. (2016) employ the notation of Anderson and van Wincoop (2003) to demonstrate the equivalence between the industry-level gravity equations on the demand side and the supply side and discuss the implications for gravity estimations.

[14] In practice, we use the fast and robust estimation command 'ppmlhdfe' of Correia, Guimarães and Zylkin (2020).

[15] In the Technical Appendix, we also obtain and report gravity estimates with the HS6+ version of GRANTPA – at the 2-digit, 4-digit, and 6-digits HS levels. The pooled gravity estimates are plausible, and we do not detect any significant biases across the average estimates that are obtained at each of the three alternative levels of aggregation. For further details, see Figures 1, 2, and 3 in the Technical Appendix.

drop the top and bottom five product-level estimates with the widest confidence intervals. The four panels of Figure 1 report the estimates for each of the four gravity variables in our model, and in each case, we have ordered them from smallest to largest.

Panel A of Figure 1 reports the results for distance—the most widely used and robust gravity covariate. The main conclusions that we draw from this figure are threefold. First, most of the estimates (about 94%) of the effects of distance on product-level trade are negative and statistically significant, which is consistent with the voluminous gravity literature. Second, in terms of magnitude, the average of the distance estimates is -0.769 (std.dev. 0.618), which is also readily comparable with the vast majority of the distance estimates from the existing literature. Third, the estimates of the effects of distance are quite heterogeneous across the products covered by the GRANTPA database. This is important for the current purposes because the wide variation in the estimates of the distance effects that we obtain suggests that more aggregate gravity estimates mask significant heterogeneity, which may be very important from a policy perspective.

Without going into too much detail, we note that the estimates on contiguity and common international language are both mostly positive and statistically significant. Specifically, 78% of the estimates of the effects of contiguity that we obtain are positive and most of them are statistically significant. Similarly, 80% of the estimates of the effects of common language are positive and, once again, most of them are statistically significant. These results are also consistent with findings from the existing literature and imply that sharing a common border and speaking the same language promote international trade. In terms of magnitude, the average estimates on common borders (0.316, std.dev. 1.174) and common language (0.876, std.dev. 2.572) are also very similar to corresponding estimates from the existing literature. In addition, we observe very heterogeneous estimates for these two variables, thus reinforcing the argument for using disaggregated data for gravity estimations.

Finally, we turn to the estimates on the $SMCTRY$ variable, which are reported in Panel D of Figure 1. Importantly, these estimates can only be identified due to the domestic trade dimension of the GRANTPA database. As expected, most of the $SMCTRY$ estimates (more than 90%) that we obtained are positive, and most of them are statistically significant. This result, sometimes dubbed as the 'home bias' effect, is well-established in the gravity literature and reflects the fact that *ceteris paribus*, most sales are domestic. What is novel, however, is that for the first time in the literature, we confirm this result with very disaggregated data. In terms of magnitude, the average estimate on $SMCTRY$ that we obtain is 1.741 (std.dev. 1.513), and it implies that *ceteris paribus* domestic trade is about 4-5 times larger than international trade. We find this implication plausible, and it is comparable to recent estimates from the gravity literature.

Finally, and similar to the estimates on the other gravity variables, we observe very wide heterogeneity in the 'home bias' effects at the product level. We believe that exploring this heterogeneity further, e.g., investigating its drivers or variation across countries, etc., could be very interesting and important from a policy perspective. Similarly, we know that our gravity specification can be improved and extended to include several other important determinants of trade flows, e.g., various bilateral as well as country-specific trade policies. However, since our current purposes are simply to demonstrate the usefulness and applicability of the GRANTPA database for gravity estimations, we leave this type of more detailed analysis for future work.

4 Conclusion

This paper introduced *The Granular Trade and Production Activities* (GRANTPA) database, which covers international trade data for 3,124 products and 247 countries over the period 1995-2019 and production and domestic trade data for the same number of products and years for 35 European economies. After describing the methods that we employed to construct the GRANTPA database, we demonstrated its usefulness with a gravity application that delivers estimates of several standard gravity variables. We draw two main conclusions about the usefulness of GRANTPA based on this gravity analysis. First, the average estimates that we obtain on each of the standard gravity variables in our econometric model are comparable to the gravity estimates from the existing literature. This reveals that the GRANTPA database is representative in the sense that it captures and reflects the gravity forces that have already been established to shape international (and domestic) trade flows. An alternative interpretation is that gravity works at the very disaggregated level. Second, the disaggregated estimates of all gravity variables in our model vary widely across the products in the GRANTPA database. Consistent with the main motivation for constructing the GRANTPA database, the implication for our database is that more aggregated gravity analysis masks significant heterogeneity, which may be very important from a policy perspective. Accordingly, we expect that the GRANTPA database will be useful for analyzing the effects of various bilateral and country-specific policies.

The norm is that trade theory and trade policy are done in a general equilibrium (GE), e.g., a bilateral free trade agreement or a tariff war between two countries, which may have significant implications for other countries that are not part of the agreement or the tariff war. Proper GE analysis requires consistent trade and production data, and we are aware of some excellent databases that can be used for GE analysis, e.g., Timmer et al. (2015) (WIOD), OECD (2023) (ICIO), and Aguiar et al. (2019) (GTAP) database. However, all existing GE datasets are relatively aggregated (e.g., covering around 50 sectors). As demonstrated, the GRANTPA database can be used to obtain product-level estimates. In terms of GE analysis, we are aware that the GRANTPA database only covers a limited number of countries and that the data is heavily unbalanced. Hence, for future research, we may harmonize and expand the non-EU countries' trade and production data and expand the scope of the database.

References

Aguiar, Angel, Maksym Chepeliev, Erwin L. Corong, Robert McDougall, and Dominique Van Der Mensbrugghe. 2019. "The GTAP Data Base: Version 10." *Journal of Global Economic Analysis*, 4(1): 1–27.

Anderson, James E., and Eric van Wincoop. 2003. "Gravity with Gravitas: A Solution to the Border Puzzle." *American Economic Review*, 93(1): 170–192.

Anderson, James E., and Eric van Wincoop. 2004. "Trade Costs." *Journal of Economic Literature*, 42(3): 691–751.

Arkolakis, Costas, Arnaud Costinot, and Andrés Rodríguez-Clare. 2012. "New Trade Models, Same Old Gains?" *American Economic Review*, 102(1): 94–130.

Bernard, Andrew B, Emily J. Blanchard, Ilke Van Beveren, and Hylke Vandenbussche. 2018. "Carry-Along Trade." *The Review of Economic Studies*, 86(2): 526–563.

Beverelli, Cosimo, Alexander Keck, Mario Larch, and Yoto V. Yotov. 2023. "Institutions, Trade, and Development: Identifying the Impact of Country-specific Characteristics on International Trade." *Oxford Economic Papers*.

Correia, Sergio, Paulo Guimarães, and Tom Zylkin. 2020. "Fast Poisson estimation with high-dimensional fixed effects." *Stata Journal*, 20(1): 95–115.

Costinot, Arnaud, Dave Donaldson, and Ivana Komunjer. 2012. "What Goods Do Countries Trade? A Quantitative Exploration of Ricardo's Ideas." *Review of Economic Studies*, 79(2): 581–608.

Gurevich, Tamara, and Peter Herman. 2018. "The Dynamic Gravity Dataset: 1948-2016." USITC Working Paper 2018-02-A.

Gurevich, Tamara, Peter Herman, Farid Toubal, and Yoto Yotov. 2021. "One Nation, One Language? Domestic Language Diversity, Trade and Welfare." *CESifo Working Paper.*

Gurevich, Tamara, Peter Herman, Farid Toubal, and Yoto Yotov. 2023. "The Domestic and International Common Language Database (DICL)." Unpublished Manuscript.

Heid, Benedikt, Mario Larch, and Yoto V. Yotov. 2021. "Estimating the Effects of Non-discriminatory Trade Policies within Structural Gravity Models." *Canadian Journal of Economics*, 54(1): 376–409.

Larch, Mario, Serge Shikher, and Yoto Yotov. 2025. "Estimating Gravity Equations: Theory Implications, Econometric Developments, and Practical Recommendations." Center for Global Policy Analysis, LeBow College of Business, Drexel University Working Papers 2025001.

OECD. 2023. "OECD Inter-Country Input-Output Database." Accessed: December 13 2023.

Pierce, Justin R., and Peter K. Schott. 2012*a*. "A Concordance Between Ten-digit US Harmonized System Codes and SIC/NAICS Product Classes and Industries." *Journal of Economic and Social Measurement*, 37(1-2): 61–96.

Pierce, Justin R., and Peter K. Schott. 2012*b*. "Concording US Harmonized System Categories over time." *Journal of Official Statistics*, 28(1): 53–68.

Santos Silva, João M.C., and Silvana Tenreyro. 2006. "The Log of Gravity." *Review of Economics and Statistics*, 88(4): 641–658.

Shikher, Serge. 2011. "Capital, Technology, and Specialization in the Neoclassical Model." *Journal of International Economics*, 83(2): 229–242.

Timmer, Marcel P., Erik Dietzenbacher, Bart Los, Robert Stehrer, and Gaaitzen J. De Vries. 2015. "An Illustrated User Guide to the World Input–Output Database: The Case of Global Automotive Production." *Review of International Economics*, 23(3): 575–605.

Van Beveren, Ilke, Andrew B. Bernard, and Hylke Vandenbussche. 2012. "Concording EU Trade and Production Data over Time." *National Bureau of Economic Research*, Working Paper (18604).

Yotov, Yoto V. 2022. "On the Role of Domestic Trade Flows for Estimating the Gravity Model of Trade." *Contemporary Economic Policy*.

Yotov, Yoto V., Roberta Piermartini, José-Antonio Monteiro, and Mario Larch. 2016. *An Advanced Guide to Trade Policy Analysis: The Structural Gravity Model.* Geneva, Switzerland:United Nations and World Trade Organization, available for download at https://www.wto.org/english/res_e/publications_e/advancedguide2016_e.htm.

Table 1: Structure of the Combined Nomenclature (CN8) Classification (Extended)

Combined Nomenclature 8-digit(CN8)		*Harmonized System 6-digit (HS6)*
Year	*# of CN8 products*	
1988	9506	HS6 1988 (# HS6 = 5019)
1989	9579	
1990	9695	
1991	9743	
1992	9837	HS6 1992 (# HS6 = 5018)
1993	9906	
1994	10108	
1995	10448	
1996	10495	HS6 1996 (# HS6 = 5113)
1997	10606	
1998	10587	
1999	10428	
2000	10314	
2001	10274	
2002	10400	HS6 2002 (# HS6 = 5224)
2003	10404	
2004	10174	
2005	10096	
2006	9841	
2007	9720	HS6 2007 (# HS6 = 5051)
2008	9699	
2009	9569	
2010	9443	
2011	9294	
2012	9383	HS6 2012 (# HS6 = 5205)
2013	9376	
2014	9379	
2015	9386	
2016	9414	
2017	9528	HS6 2017 (# HS6 = 5387)
2018	9533	
2019	9533	
2020	9483	
2021	9494	
2022	9736	HS6 2022 (# HS6 = 5612)

Note: All classification files are obtained from Eurostat Ramon server.

Table 2: Changes in the Combined Nomenclature Classification Over Time: Extension

Effective year	Number of obsolete codes	Number of new codes	Number of families (including simple changes)	Number of simple (one-to-one) changes
1989	76	149	58	1
1990	122	238	111	11
1991	85	133	64	8
1992	128	222	85	2
1993	276	345	171	14
1994	233	435	197	11
1995	531	871	383	31
1996	1257	1304	792	435
1997	170	281	130	0
1998	334	315	175	0
1999	303	144	132	3
2000	223	109	96	0
2001	90	50	42	1
2002	847	973	504	311
2003	16	20	12	0
2004	503	273	211	7
2005	186	108	95	5
2006	743	489	281	11
2007	1202	108	630	387
2008	96	75	54	2
2009	257	127	111	0
2010	381	255	151	1
2011	282	133	124	0
2012	959	1048	637	357
2013	43	36	24	1
2014	40	43	22	2
2015	18	25	11	0
2016	27	55	18	0
2017	766	876	414	133
2018	13	18	9	0
2019	9	9	4	0
2020	104	54	42	1
2021	9	20	9	0
2022	535	769	332	135

Note: This table shows the number of obsolete and new codes for each year, as well as the number of families (shrinking, growing, or simple) and the number of simple changes (one-to-one). The effective year refers to the year in which the change becomes effective. HS6 codes have been revised in 1992, 1996, 2002, 2007, 2012, 2017 and 2022. The main changes in the combined nomenclature (CN8) classification over time are obtained from Eurostat Ramon server as shown in Van Beveren, Bernard and Vandenbussche (2012).

Table 3: Changes in the Prodcom Classification Over Time: Extension

Effective year	Number of obsolete codes	Number of new codes	Number of families (including simple changes)	Number of simple (one-to-one) changes	Number of codes that are dropped (exit)	Number of codes that are new on the list (entry)
1994	32	46	29	17	4	3
1995	33	52	15	12	19	29
1996	118	80	54	12	14	15
1997	0	0	0	0	0	0
1998	2	0	1	0	2	0
1999	68	92	31	2	3	62
2000	16	12	9	1	0	0
2001	113	76	57	0	0	0
2002	82	54	29	3	1	3
2003	363	296	215	190	1	13
2004	35	24	17	1	1	2
2005	305	105	91	0	67	1
2006	4	2	2	0	0	0
2007	184	131	76	13	3	9
2008	4396	3864	3651	3258	52	19
2009	28	15	15	1	1	1
2010	45	26	23	4	0	0
2011	61	28	28	0	0	0
2012	68	53	40	11	0	5
2013	11	8	1	0	11	8
2014	4	2	1	0	4	2
2015	9	6	1	0	9	6
2016	141	135	95	72	23	28
2017	105	43	1	0	105	43
2018	0	0	0	0	0	0
2019	79	217	1	0	79	217
2020	0	0	0	0	0	0
2021	5	12	1	0	5	12

Note: This table shows the number of obsolete and new codes in each year, as well as the number of families (shrinking, growing, simple, entry or exit) and the number of simple changes (one-to-one). The effective year refers to the year in which the change became effective. Some PC8 codes are not covered throughout the whole sample period, resulting in new codes (*entry*) appearing on the list and old codes (*exit*) disappearing from the list. All changes in the PC8 classification over time are obtained from Eurostat Ramon server. Following closely Van Beveren, Bernard and Vandenbussche (2012), optional codes have been removed (or replaced by their mandatory aggregates) to ensure comparability over time and across countries.

Table 4: Comparison of PC8+ and HS6+ procedure

Procedure Step	**PC8+ Approach**	**HS6+ Approach**	**Key Difference**
A. Reading and cleaning	Reads CN8 for trade and PC8 for production; keeps complete 8-digit codes unless special breakdowns (Z, T, Q, V, E) must be recoded.	Truncates CN8 to 6 digits (HS6), then merges with PC8 if possible; letter-coded breakdowns also get recoded.	PC8+ procedure remains fully at 8-digit detail; HS6+ starts at 6 digits, inherently more aggregated.
B. Identify code types	Uses CN8–PC8 cross-sectional concordance, flags the codes as one-to-many, many-to-one, many-to-many, or simple at the 8-digit level.	Applies the same logic, but to 6-digit HS codes matched to PC8.	Both rely on iterative "feedback" loops developed by Pierce and Schott (2012*a*); However, PC8+ is finer in granularity (8 digits), HS6+ (6 digits).
C. Matching correspondances	Merges CN8 with PC8 at the 8-digit level to obtain `pc8plus`. Codes not covered by PC8 are dropped.	CN8 is truncated to HS6; merges with PC8 if possible. Multiple HS6 codes linking to PC8 become `hs6plus`.	PC8+ mostly remains at 8 digits. HS6+ collapses synthetic codes and retains many-to-one and simple codes for consistency.
D. Handling optional and breakdown codes	Recodes and drops optional B-/N-lists and breakdown PC8 codes to avoid double-counting; only "mandatory" PC8 codes remain.	Similar approach, but after truncation of HS6. Creating a unique group of codes, where simple and many-to-one codes remain at 6-digit.	Both recode PC8 special codes; HS6+ entails further aggregation at the 6-digit scope.
E. Aggregation	If multiple CN8 or PC8 map to one product, they merge into a single `pc8plus`; simple codes remain unchanged.	HS6 lines are collapsed into `hs6plus` families as needed. Multiple CN8 codes or partial coverage lead to a single HS6+ code.	PC8+ can preserve more granular codes if no merges are required; HS6+ systematically aggregates at 6 digits.
F. Linking codes over time	Uses iterative "feedback" loop from Pierce and Schott (2012*b*) to unify code changes year by year PC8→PC8+, giving one final identifier `pc8plus` which is consistent over time.	The same principle applied on truncated 6-digit HS, HS6→HS6+ producing `hs6plus` over time.	Both rely on year-by-year code families; PC8+ is finer, HS6+ remains more aggregated.

Note: Both procedures follow closely Van Beveren, Bernard and Vandenbussche (2012) framework for identifying code mappings (one-to-many, many-to-one, many-to-many, simple). For further details on HS6+ procedure see VBBV's supplementary: "Concording trade and production data in a single year" and "Concording HS6 products over time": Readme files.

Table 5: Mapping: Combined Nomenclature (CN8) and Prodcom

Year	Many to One		One to Many		Many to Many		Simple		Total
	N	%	N	%	N	%	N	%	N
1995	5995	63.87%	4	0.04%	790	8.42%	2597	27.67%	9386
1996	6091	64.68%	4	0.04%	805	8.55%	2517	26.73%	9417
1997	6231	65.36%	4	0.04%	805	8.44%	2494	26.16%	9534
1998	6292	65.76%	4	0.04%	787	8.23%	2485	25.97%	9568
1999	5716	56.17%	4	0.04%	1968	19.34%	2489	24.46%	10177
2000	5652	55.91%	4	0.04%	1968	19.47%	2485	24.58%	10109
2001	5726	55.58%	4	0.04%	2091	20.30%	2481	24.08%	10302
2002	5857	56.63%	4	0.04%	2042	19.74%	2440	23.59%	10343
2003	6403	63.88%	4	0.04%	1072	10.69%	2545	25.39%	10024
2004	6238	63.87%	6	0.06%	950	9.73%	2572	26.34%	9766
2005	6383	68.79%	6	0.06%	302	3.25%	2588	27.89%	9279
2006	6094	66.98%	6	0.07%	302	3.32%	2696	29.63%	9098
2007	6101	67.97%	6	0.07%	224	2.50%	2645	29.47%	8976
2008	6658	75.06%	6	0.07%	221	2.49%	1985	22.38%	8870
2009	6529	74.55%	6	0.07%	228	2.60%	1995	22.78%	8758
2010	6489	76.14%	0	0.00%	0	0.00%	2033	23.86%	8522
2011	6345	75.76%	0	0.00%	0	0.00%	2030	24.24%	8375
2012	5781	60.42%	23	0.24%	1771	18.51%	1993	20.83%	9568
2013	5778	60.44%	26	0.27%	1767	18.48%	1989	20.81%	9560
2014	5780	60.43%	26	0.27%	1770	18.50%	1989	20.79%	9565
2015	5786	60.38%	26	0.27%	1785	18.63%	1985	20.72%	9582
2016	5812	60.20%	29	0.30%	1832	18.98%	1981	20.52%	9654
2017	5915	59.41%	41	0.41%	2111	21.20%	1890	18.98%	9957
2018	5920	59.43%	41	0.41%	2111	21.19%	1890	18.97%	9962
2019	5764	57.80%	41	0.41%	2111	21.17%	2057	20.63%	9973

Note: This table shows the evolution of code mappings between the Combined Nomenclature (CN8) and Prodcom (PC8) classifications from 1995 to 2019. For each year, columns 2–9 report the absolute number (N) and percentage (%) of CN8–PC8 code pairs in four mapping categories: Many-to-One, One-to-Many, Many-to-Many, and Simple. The final column indicates the total number of CN8–Prodcom pairs in the respective year. A Many-to-One relationship occurs when multiple CN8 codes map into a single Prodcom code. A One-to-Many relationship occurs when a single CN8 code corresponds to several Prodcom codes. A Many-to-Many mapping indicates that multiple CN8 codes are linked to multiple PC8 codes. These complex mappings often require additional steps in the concordance procedure (Pierce and Schott, 2012*b*). A Simple (One-to-One) mapping denotes that each CN8 code uniquely matches exactly one PC8 code.

Table 6: Mapping: Prodcom (PC8)

Year	Many to One		One to Many		Many to Many		Simple		Total
	N	%	N	%	N	%	N	%	N
1995	1404	32.90%	4	0.09%	263	6.16%	2597	60.85%	4268
1996	1432	33.93%	4	0.09%	267	6.33%	2517	59.64%	4220
1997	1455	34.48%	4	0.09%	267	6.33%	2494	59.10%	4220
1998	1461	34.65%	4	0.09%	267	6.33%	2485	58.93%	4217
1999	1420	33.41%	4	0.09%	337	7.93%	2489	58.56%	4250
2000	1417	33.40%	4	0.09%	337	7.94%	2485	58.57%	4243
2001	1434	33.87%	4	0.09%	315	7.44%	2481	58.60%	4234
2002	1451	34.45%	4	0.09%	317	7.53%	2440	57.93%	4212
2003	1549	35.37%	4	0.09%	281	6.42%	2545	58.12%	4379
2004	1518	34.77%	6	0.14%	270	6.18%	2572	58.91%	4366
2005	1582	37.49%	6	0.14%	44	1.04%	2588	61.33%	4220
2006	1472	34.90%	6	0.14%	44	1.04%	2696	63.92%	4218
2007	1456	35.13%	6	0.14%	38	0.92%	2645	63.81%	4145
2008	1520	42.71%	6	0.17%	48	1.35%	1985	55.77%	3559
2009	1495	42.16%	6	0.17%	50	1.41%	1995	56.26%	3546
2010	1480	42.13%	0	0.00%	0	0.00%	2033	57.87%	3513
2011	1449	41.65%	0	0.00%	0	0.00%	2030	58.35%	3479
2012	1377	38.41%	23	0.64%	192	5.36%	1993	55.59%	3585
2013	1376	38.40%	26	0.73%	192	5.36%	1989	55.51%	3583
2014	1374	38.37%	26	0.73%	192	5.36%	1989	55.54%	3581
2015	1375	38.43%	26	0.73%	192	5.37%	1985	55.48%	3578
2016	1366	38.11%	29	0.81%	208	5.80%	1981	55.27%	3584
2017	1379	39.08%	41	1.16%	219	6.21%	1890	53.56%	3529
2018	1379	39.08%	41	1.16%	219	6.21%	1890	53.56%	3529
2019	1348	36.78%	41	1.12%	219	5.98%	2057	56.13%	3665

Note: This table shows the distribution of code mappings within the Prodcom classification (PC8) from 1995 to 2019. For each year, columns 2–9 show the absolute number (N) and percentage (%) of PC8 codes in four mapping categories: Many-to-One, One-to-Many, Many-to-Many and Simple. The last column gives the total number of PC8 code pairs in that year.

Table 7: Example Many to Many Mappings: PC8+

Year	CN8	PC8	PC8+	Code	Declarant	Partner	Flow	Value II	Value IE	Value XI	Value XE	`drop_synth`
1995	62043290	18223331	1442	56.2005	FR	IT	1	4702012	.	.	.	1
1995	62043290	18223331	1442	56.2005	FR	IT	2	.	.	2217971	.	1
1995	62043100	18223334	1442	56.2005	FR	IT	1	10522374	.	.	.	1
1995	62043100	18223334	1442	56.2005	FR	IT	2	.	.	2903226	.	1
1995	62043390	18223339	1442	56.2005	FR	IT	1	3443395	.	.	.	1
1995	62043390	18223339	1442	56.2005	FR	IT	2	.	.	2490222	.	1
1995	62043919	18223339	1442	56.2005	FR	IT	1	4396515	.	.	.	1
1995	62043919	18223339	1442	56.2005	FR	IT	2	.	.	1013895	.	1
1995	62043990	18223339	1442	56.2005	FR	IT	1	3750361	.	.	.	1
1995	62043990	18223339	1442	56.2005	FR	IT	2	.	.	1974194	.	1
1995	62043100	182233Q0	1442	56.2005	FR	IT	1	10522374	.	.	.	0
1995	62043100	182233Q0	1442	56.2005	FR	IT	2	.	.	2903226	.	0
1995	62043290	182233Q0	1442	56.2005	FR	IT	1	4702012	.	.	.	0
1995	62043290	182233Q0	1442	56.2005	FR	IT	2	.	.	2217971	.	0
1995	62043390	182233Q0	1442	56.2005	FR	IT	1	3443395	.	.	.	0
1995	62043390	182233Q0	1442	56.2005	FR	IT	2	.	.	2490222	.	0
1995	62043919	182233Q0	1442	56.2005	FR	IT	1	4396515	.	.	.	0
1995	62043919	182233Q0	1442	56.2005	FR	IT	2	.	.	1013895	.	0
1995	62043990	182233Q0	1442	56.2005	FR	IT	1	3750361	.	.	.	0
1995	62043990	182233Q0	1442	56.2005	FR	IT	2	.	.	1974194	.	0

Note: This table shows how duplicate observations are flagged using the `drop_synth` column, allowing us to remove duplicates in further steps and providing a unique PC8+ over time code.

Table 8: Example Many to Many Complex Mappings: Multiple PC8+ over time

Year	CN8	PC8	PC8+	Code	Declarant	Partner	Flow	Value II	Value IE	Value XI	Value XE	`drop_synth`
2019	72085120	24103150	1385	17.2003	IT	AT	1	10951684	.	.	.	0
2019	72085191	24103150	1385	17.2003	IT	AT	1	317515	.	.	.	0
2019	72085198	24103150	1385	17.2003	IT	AT	1	125962	.	.	.	0
2019	72085291	24103150	1385	17.2003	IT	AT	1	163221	.	.	.	0
2019	72085299	24103150	1385	17.2003	IT	AT	1	153151	.	.	.	0
2019	72085310	24103150	1385	17.2003	IT	AT	1	1954	.	.	.	0
2019	72089020	24103150	1385	17.2003	IT	AT	1	1496	.	.	.	0
2019	72089080	24103150	1385	17.2003	IT	AT	1	1875694	.	.	.	0
2019	72111300	24103210	1385	17.2003	IT	AT	1	16875	.	.	.	0
2019	72192110	24103340	1385	17.2003	IT	AT	1	181648	.	.	.	0
2019	72192190	24103340	1385	17.2003	IT	AT	1	133770	.	.	.	0
2019	72259900	24103550	1385	17.2003	IT	AT	1	6002248	.	.	.	0
2019	72109030	24105150	1385	17.2003	IT	AT	1	831136	.	.	.	0
2019	72085120	2410T222	1385	2410T222	IT	AT	1	10951684	.	.	.	1
2019	72085191	2410T222	1385	2410T222	IT	AT	1	317515	.	.	.	1
2019	72085198	2410T222	1385	2410T222	IT	AT	1	125962	.	.	.	1
2019	72085291	2410T222	1385	2410T222	IT	AT	1	163221	.	.	.	1
2019	72085299	2410T222	1385	2410T222	IT	AT	1	153151	.	.	.	1
2019	72085310	2410T222	1385	2410T222	IT	AT	1	1954	.	.	.	1
2019	72085390	2410T222	1385	2410T222	IT	AT	1	176273	.	.	.	1
2019	72089020	2410T222	1385	2410T222	IT	AT	1	1496	.	.	.	1
2019	72089080	2410T222	1385	2410T222	IT	AT	1	1875694	.	.	.	1
2019	72109030	2410T222	1385	2410T222	IT	AT	1	831136	.	.	.	1
2019	72111300	2410T222	1385	2410T222	IT	AT	1	16875	.	.	.	1
2019	72254012	2410T222	1385	2410T222	IT	AT	1	2005078	.	.	.	1
2019	72254040	2410T222	1385	2410T222	IT	AT	1	12432648	.	.	.	1
2019	72084000	24103130	1385	6.2005	IT	AT	1	7237	.	.	.	1
2019	72085390	24103130	1385	6.2005	IT	AT	1	176273	.	.	.	1
2019	72085400	24103130	1385	6.2005	IT	AT	1	296742	.	.	.	1
2019	72254012	24103530	1385	6.2005	IT	AT	1	2005078	.	.	.	1

Note: This table highlights that the same product can appear under multiple classifications or under aggregated categories (e.g. PC8 2410T222). Observations with flagged `drop_synth` represent duplicates, for treatment of double-counting. In this case, the code over time (e.g. 17.2003) is uniquely carried over time.

Table 9: Example Many to One Mappings: PC8+

Year	CN8	PC8	PC8+	Code	Declarant	Partner	Flow	Value II	Value IE	Value XI	Value XE	`drop_synth`
1995	2021000	15111200	15111200	423.2008	FR	IT	1	76235.0	.	.	.	.
1995	2021000	15111200	15111200	423.2008	FR	IT	2	.	.	6363.0	.	.
1995	2022030	15111200	15111200	423.2008	FR	IT	1	8926.0	.	.	.	.
1995	2022050	15111200	15111200	423.2008	FR	IT	1	4788.0	.	.	.	.
1995	2022050	15111200	15111200	423.2008	FR	IT	2	.	.	186493.0	.	.
1995	2022090	15111200	15111200	423.2008	FR	IT	1	63565.0	.	.	.	.
1995	2022090	15111200	15111200	423.2008	FR	IT	2	.	.	39119.0	.	.
1995	2023010	15111200	15111200	423.2008	FR	IT	1	2569560.0	.	.	.	.
1995	2023090	15111200	15111200	423.2008	FR	IT	1	1970130.0	.	.	.	.
1995	2023090	15111200	15111200	423.2008	FR	IT	2	.	.	2308153.0	.	.

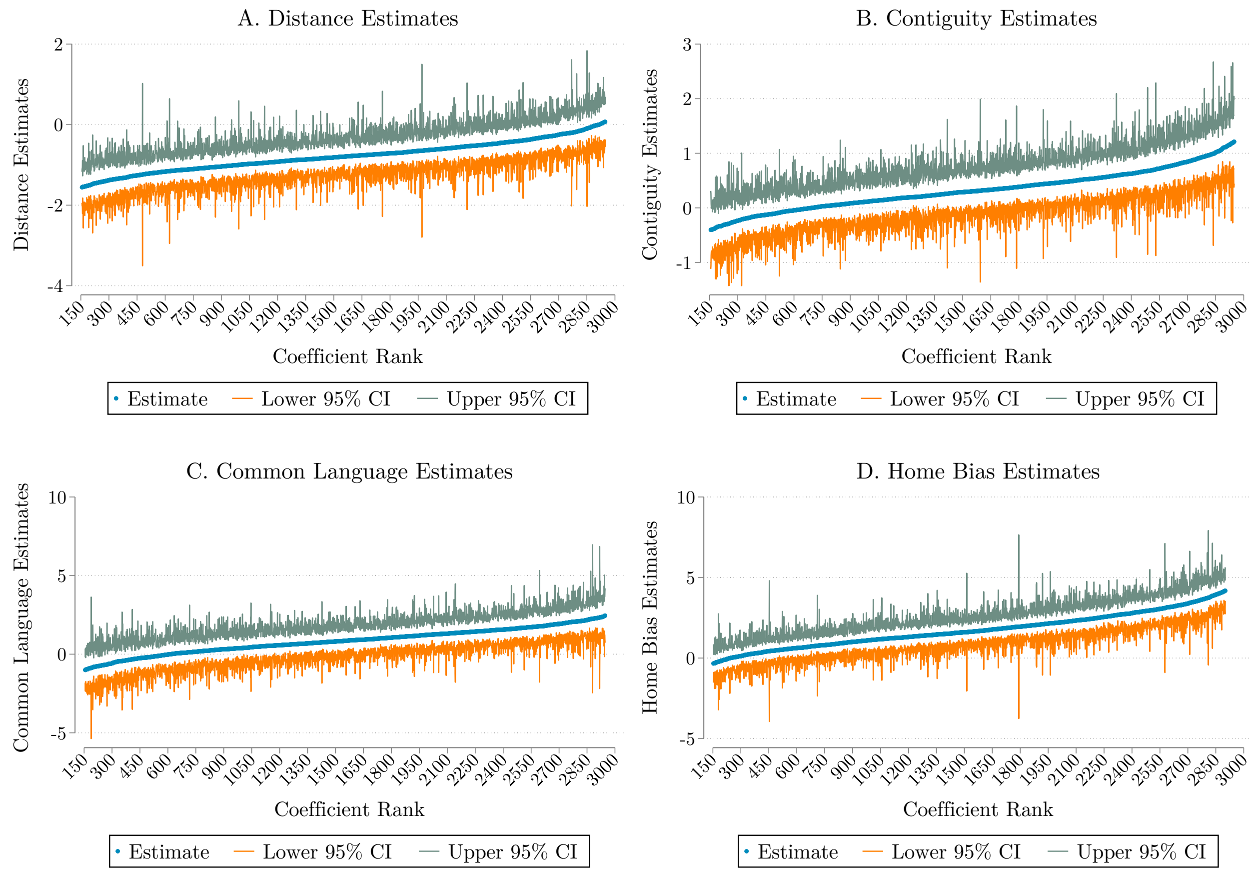
Note: This table shows the Many to One mapping, where several CN8 codes match into one PC8 which contains a unique PC8+ over time code. In this case, there is no need to drop data given the unique structure.

Table 10: Example One to Many Mappings: PC8+

Year	CN8	PC8	PC8+	Code	Declarant	Partner	Flow	Value II	Value IE	Value XI	Value XE	`drop_synth`
1995	72131000	27105010	1	43.2003	FR	IT	1	406457.0	.	.	.	1
1995	72131000	27105010	1	43.2003	FR	IT	2	.	.	45348.0	.	1
1995	72131000	271050V0	1	43.2003	FR	IT	1	406457.0	.	.	.	0
1995	72131000	271050V0	1	43.2003	FR	IT	2	.	.	45348.0	.	0

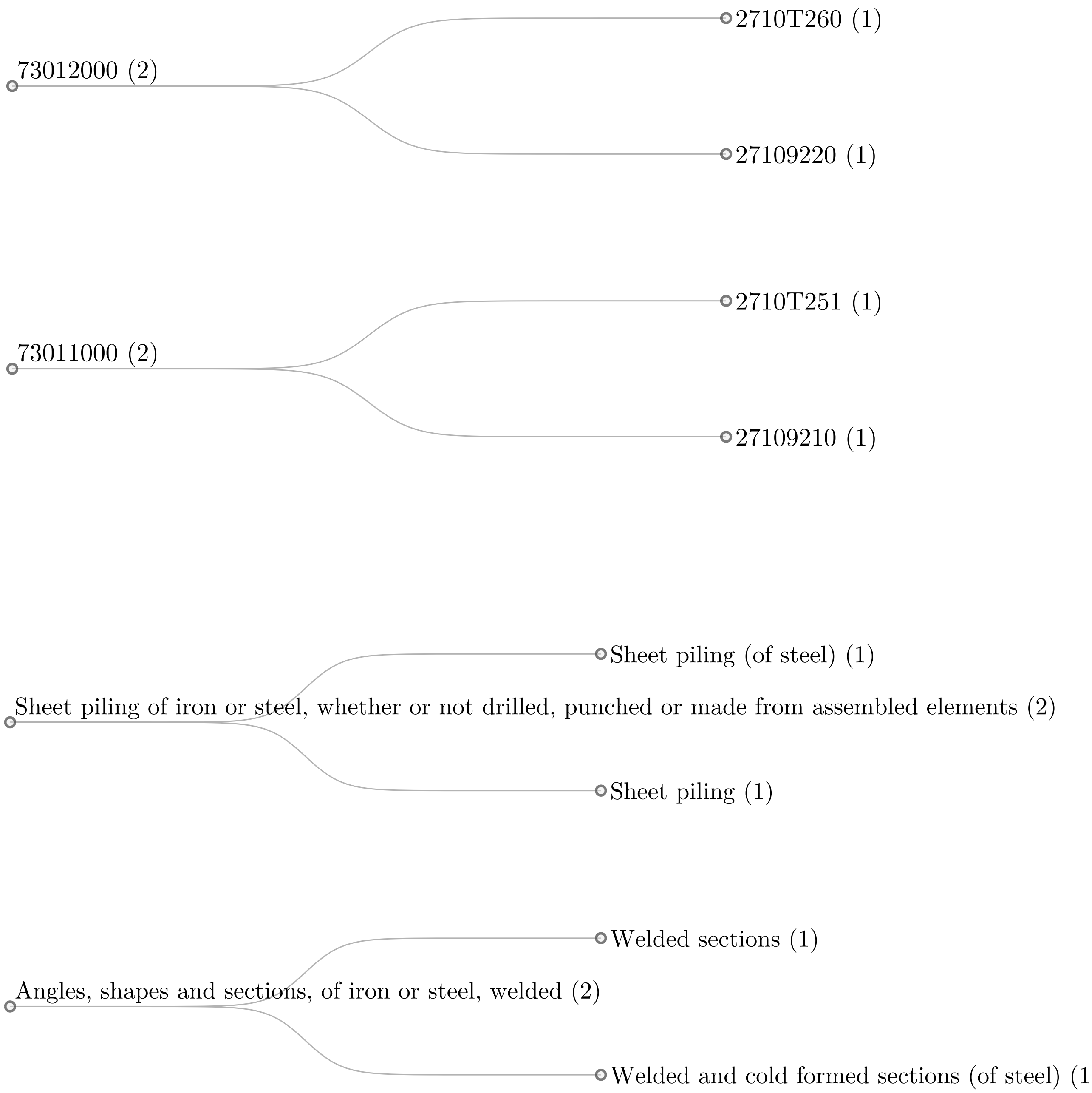
Note: This table shows the One to Many mapping, where multiple PC8 codes match into one CN8 code and the issue of potential duplicates. The variable `drop_synth` solves the problem of double-counting while addressing unique product codes that are traceable over time. Notice that we keep the aggregated version (e.g. 271050V0) but we use the code over time of the mandatory code (e.g. 43.2003).

Figure 1: Gravity Estimates with the GRANTPA Database, 1995-2019



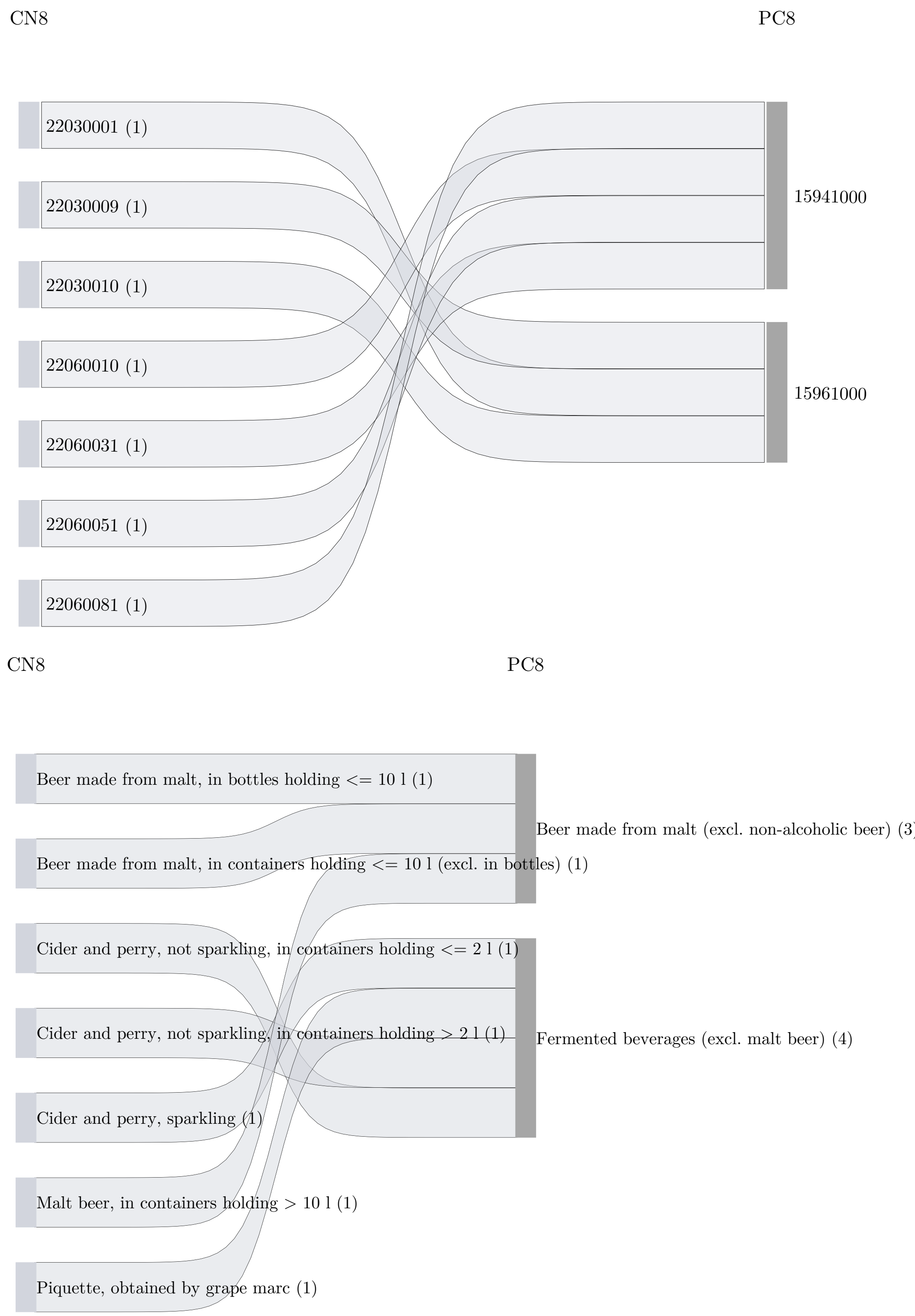
Notes: This figure reports the estimates, along with the corresponding confidence intervals, of the effects on four standard gravity variables. All estimates are obtained with the PPML estimator and exporter-time and importer-time fixed effects according to specification (1), where the dependent variable is always product-level nominal trade in levels from the GRANTPA database. Panel A graphs the estimates of the effects of the log of bilateral distance. Panel B shows the estimates of the effects of contiguous borders. Panel C plots the estimates of the effects of common language. Finally, Panel D visualizes the estimates of the 'home bias'. See text for further details.

Figure 2: Example Complex Codes in 2003: One-to-Many



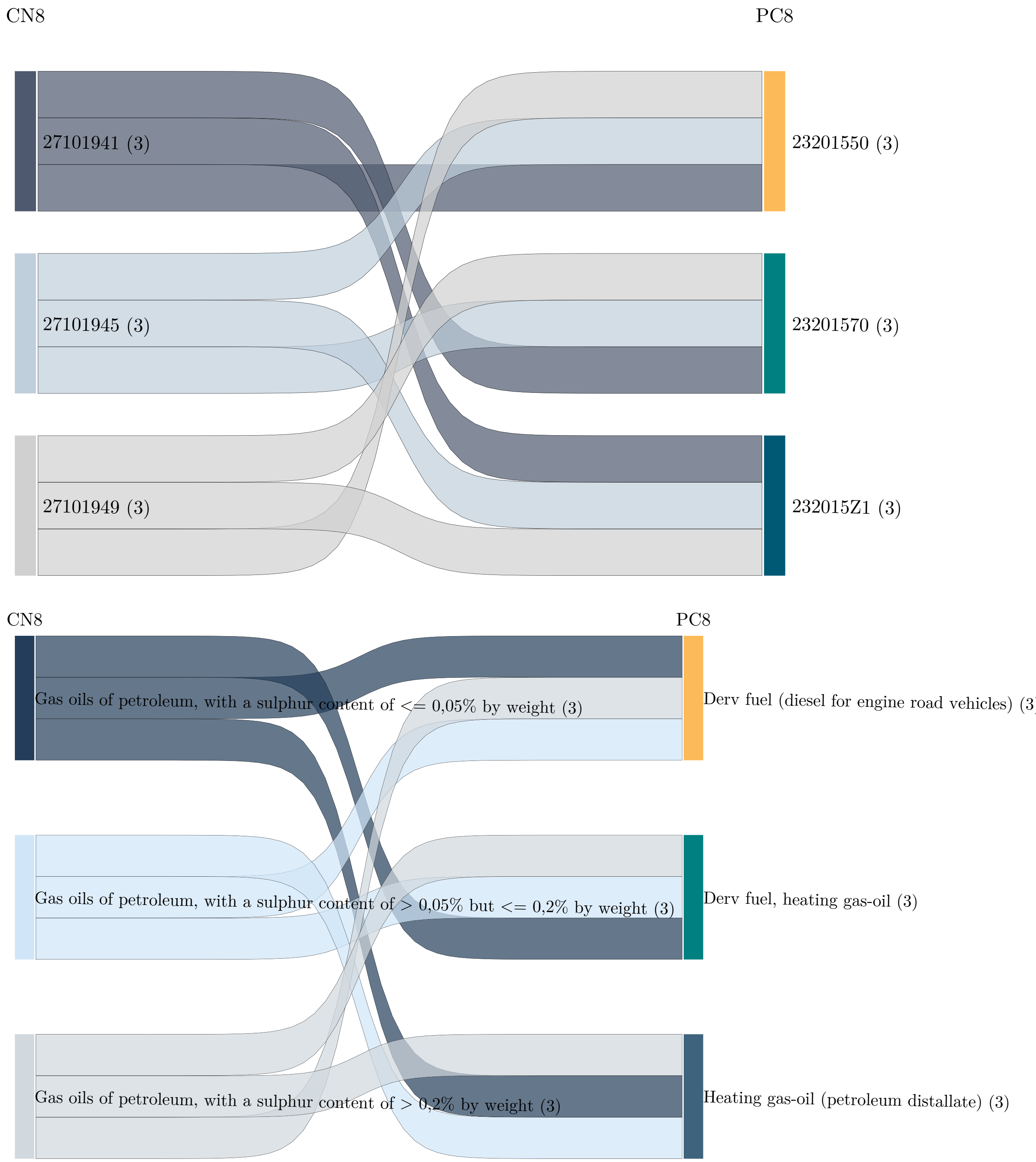
Notes: Mapping of Combined Nomenclature (CN8) to *Prodcom* Codes (PC8). The 8-digit CN code 73012000 (left) is shown mapping to multiple 8-digit *Prodcom codes* 2710T260 and 27109220 (right), demonstrating a one-to-many relationship. Each line represents the transition from a single CN8 code to its corresponding PC8 variants. The numbers in parentheses indicate the frequency relevant to the mapping.

Figure 3: Example Complex Codes in 2003: Many-to-One



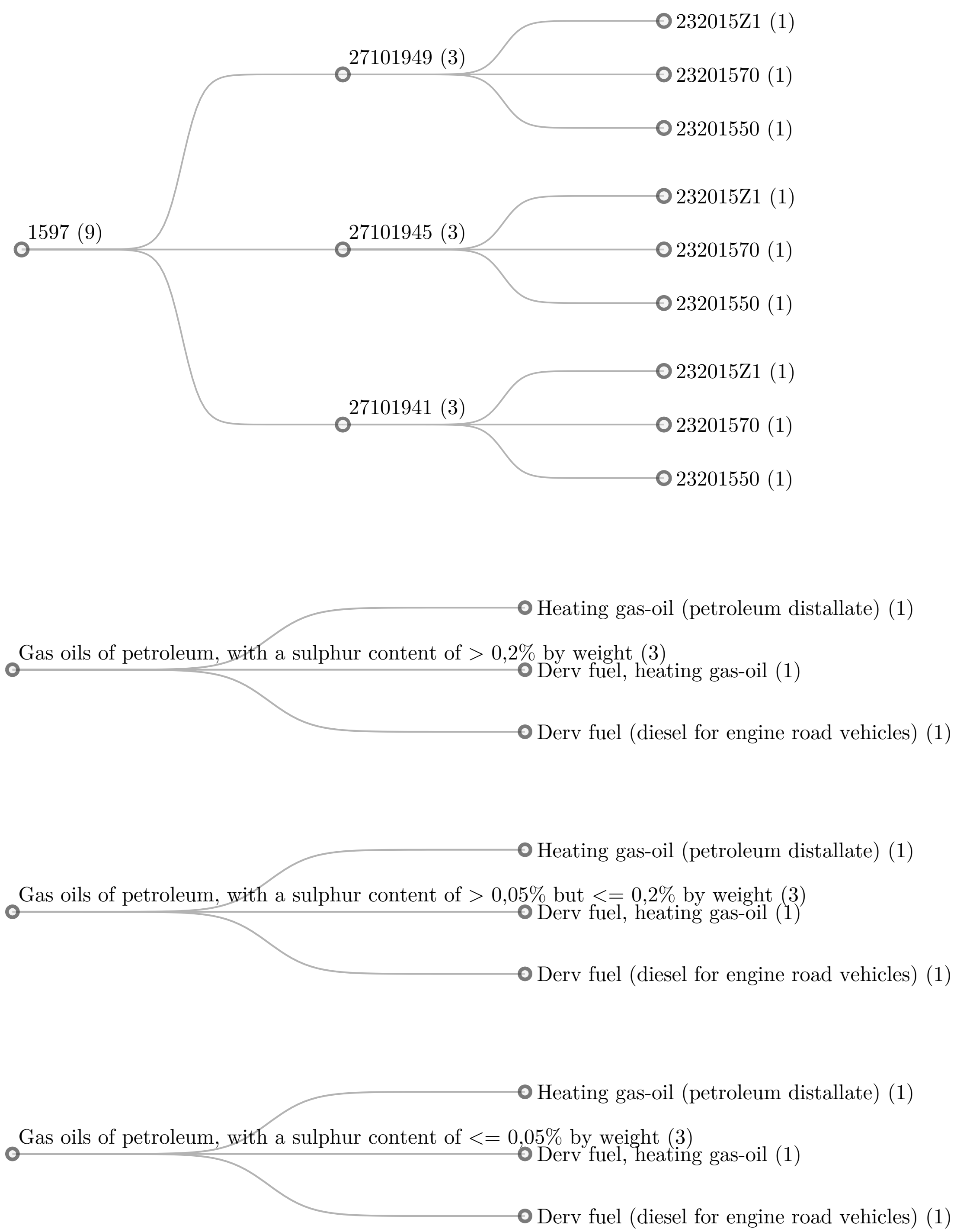
Notes: Mapping of Combined Nomenclature (CN8) (left) to Prodcom codes (PC8) (right). The figure shows a many-to-one relationship, where multiple CN8 codes (22030001, 22030009, 22030010) are linked to one PC8 code 15961000. The lines indicate the directional flow from CN8 to PC8 codes, with the numbers in parentheses denoting the frequency of the mapping process.

Figure 4: Example Complex Codes in 2003: Many-to-Many



Notes: Mapping of Combined Nomenclature (CN8) (left) to Prodcom codes (PC8) (right). This figure presents multiple CN8 codes (left) and their complex mappings to several PC8 codes (right). Each coloured pathway represents the interconnections between a CN8 code and its multiple PC8 counterparts. The numbers in parentheses denote the frequency of the mapping process.

Figure 5: Example Synthetic Codes in 2003: Many-to-Many



Notes: Complex Mapping from '*synthetic*' code (1597) through Combined Nomenclature (CN8) to Prodcom Codes (PC8). This figure illustrates the mapping process starting with PC8+ codes (left), linking through CN8 codes (center), and PC8 codes (right). The numbers in parentheses indicate the frequency.

Figure 6: Growing Family Tree

Period t	Period t+1	Period t+2

$$10101010\ a \left\{ \begin{array}{l} 10101010\ b \left\{ \begin{array}{l} 10101010\ d \\ 10101010\ e \end{array}\right. \\ 10101010\ c \left\{ \begin{array}{l} 10101010\ f \\ 10101010\ g \end{array}\right. \end{array}\right.$$

Figure 7: Shrinking Family Tree

Period t	Period t+1	Period t+2

$$\left. \begin{array}{r} \left.\begin{array}{r} 10101010\ a \\ 10101010\ b \end{array}\right\} 10101010\ e \\ \left.\begin{array}{r} 10101010\ c \\ 10101010\ d \end{array}\right\} 10101010\ f \end{array}\right\} 10101010\ g$$