A guideline to French firm-level trade data

Flora Bergounhon* Clémence Lenoir† Isabelle Mejean‡

September 27, 2018

Abstract

French customs data have been widely used in the recent empirical literature to study various trade-related topics. While access to these data is relatively easy, there is little documentation that can help researchers identify questions that such data can tackle and construct a dataset which best fits the question under study. This note provides such guidance, with a special focus on longitudinal studies. We describe the characteristics of data collection which has consequences for the scope and information available to researchers. We provide a list of minimum data trimming steps that allow recovering consistent time-series as well as more advanced algorithms that we argue are useful when using the data for more particular purposes. Finally, we provide broad statistics about the structure of the data.

1 Introduction

French customs data have been widely used in the recent empirical literature to study various trade-related issues. The list of such topics includes the participation of individual firms to international markets (Eaton et al., 2011; Blaum et al., 2018), the dynamics of exports (Bricongne et al., 2012), the product scope of exporters (Mayer et al., 2014), the vertical differentiation of traded goods (Crozet et al., 2012; Martin and Mejean, 2014), the sensitivity of trade to various shocks (Berman et al., 2012; Fontagné et al., 2015, 2017), the extent of price discrepancies in international markets (Mejean and Schwellnus, 2009), the international transmission of shocks (di Giovanni et al., 2018) or the structure of firm-to-firm trade relationships (Kramarz et al., 2016; Lenoir et al., 2018). While access to these data is relatively easy, there is little documentation that can help researchers identify questions that such data can tackle and construct a dataset which best fits the question under study. This note provides such guidance, with a special focus on longitudinal studies. We describe the characteristics of data collection which has consequences for the scope and information available to researchers. We provide a list of minimum data trimming steps that allow recovering consistent time-series as well as more advanced algorithms that we argue are useful when using the data for more particular purposes. Finally, we provide broad statistics about the structure of the data.

In France, the Customs administration is in charge of collecting information about trade in goods.[1] All data collected are obtained from French firms involved in an international transaction filling a compulsory form. As such, the data are of high quality. Researchers interested in using these data at a fine disaggregation level need to follow a strict but well-established procedure to be granted access. The procedure is open to any researcher, whatever his/her nationality or status. The first step consists in submitting a research proposal to the public institution in charge, called "*Comité du Secret*".[2] Crucially, the research proposal must describe the exact data for which access is to be granted, including the administration in charge of the collection, the name of the dataset, the level of disaggregation, the variables of interest, etc. One of this note's purposes is to help the researcher in this initial step by describing with as much details as possible the source of raw data.

Once access is granted, the French customs provide researchers with a single dataset which is customized to the particular research project under consideration. In many instance, this dataset combines information collected under two distinct legal frameworks. The first legal framework, called "DEB" for "*Déclaration d'Echange de Biens*" concerns intra-EU trade flows. Since 1993, the European

*CREST-Ecole Polytechnique, Email address: flora.bergounhon@sciencespo.fr.

†CREST, ENSAE, Email address: clemence.lenoir@ensae-paristech.fr.

‡CREST-Ecole Polytechnique and CEPR, Email address: isabelle.mejean@polytechnique.edu.

[1] Data about trade in services are collected by Banque de France and are out of the scope of this note.

[2] The procedure is described into details at https://www.comite-du-secret.fr, unfortunately in French.

Union is a single market and thus goods and services can freely move across European countries. These intra-EU trade flows are still recorded, though, for VAT purposes or to compute statistics on intra-EU trade imbalances. Information on trade between France and non-EU countries is collected through a second legal framework called "DAU" for "*Document Administratif Unique*".[3] Although the merge of information in DEB and DAU is possible, the content and scope of the data collected through these forms do vary and should be cautiously taken into account when setting up a research project.

The dataset made of all transactions collected from the DEB and DAU gives a quasi-exhaustive picture of all trade relationships between French firms and the rest of the world, whether through importing or exporting. The French Customs use these data to compute official trade statistics.[4] Data collected can also be used to study more specific questions since information on many potentially interesting variables is recorded, beyond the identity of the firm involved in the trade flow and the value of the transaction. The full list of these variables is discussed in Section 2. Among the most interesting variables, one can mention the quantity of goods traded, the product category, the destination/origin country, the transportation mode, etc. In Sections 2 and 4 we discuss in details the data coverage for these different variables. We also provide information on some regulatory changes that have affected the quality of the information collected, over time. Section 3 discusses some data treatments that can help harmonize this information. All codes necessary to perform these trimming procedures are available on this note's companion website.

The rest of the note is organized as follows. Section 2 describes the content and scope of French customs data for each collecting process. Section 3 presents a recommended cleaning procedure to work with French customs data. Section 4 presents general descriptive statistics on French trade and illustrates the different caveats evoked in Section 2.

2 Content and scope of collected information

In this section, we present the various variables that are collected by the customs through the DEB and DAU forms. As is standard in trade, two types of trade flows are to be distinguished, export transactions involving goods that are sold by a French firm abroad and import transactions for goods purchased by a French entity, be it a firm or an individual. The French customs distinguish export and import flows through a variable called ***flux*** (flow in English). By convention, import flows are coded 1 and 3, for extra- and intra-EU transactions respectively, while export flows are coded 2 and 4.

2.1 Data on intra-EU trade (DEB)

The legal framework concerning intra-EU transactions can further be divided into two parts. The first part is fiscal and is restricted to export flows (*flux* 4). The data are filled into the VAT Information Exchange System which is used by EU countries' tax administrations to control intra-EU VAT payments. Data collected within this context are highly reliable. The second part is statistical: information about transactions are collected to establish intra-EU trade balance statistics. For this purpose, the Customs collect intra-EU import flows as well as additional product-level variables regarding intra-EU exports (*flux* 3 and 4).

Information collected within the fiscal framework is limited to a small set of variables but is exhaustive. The value of each single transaction involving a French firm selling goods to a European importer is recorded into the VAT Information Exchange System, together with the European VAT numbers of the French exporter and its European partner.[5] VAT numbers are time-invariant identifiers of firms. For French firms, this number is constituted of the "FR" country code, a 2-digit key and the 9-digit "Siren" number of the firm.[6] For European importers, the Customs provide researchers with an anonymized version of their VAT numbers.

[3] Incidently, trade between metropolitan France and France overseas is also collected through DAU. Monaco is integrated to French territory to establish trade balance statistics.

[4] Note however that the data provided to researchers do not necessarily add up to official trade statistics. The reason is that the building of trade statistics follows a procedure which implies some treatment and selection over the raw data. Moreover, transactions on some goods such as arms are not provided to researchers.

[5] The value of the shipment is in euros *excluding VAT*. It corresponds to the value reported on the invoice.

[6] "Siren" is the French identifier of firms, which is available in most French firm-level datasets, e.g. tax forms, the employer-employee linked data, etc. This identifier can thus be used to merge customs data with other French firm-level administrative data.

Table 1: Description of the Customs procedures

Procedure	**Description**	**ComExt**
11	Imports of taxable goods	✓
19	Imports of non-taxable goods	✓
21	Exports of taxable goods	✓
25	Cash rebate	
26	Cash surcharge	
29	Exports of non-taxable goods	✓
31	Cash transfer under third-party trade	

Notes: The third column indicates the families of transactions that are included in official trade statistics. Regimes 25 and 26 record monetary transactions that sometimes follow a regular transaction under Regime 21, when the parties agree on a rebate or a surcharge over the value of the traded goods. Regime 31 involves specific monetary transactions occurring under third-party trade, when there is no physical trade flow mirroring the monetary transfer. Non-taxable goods include gifts or transactions associated with tolling agreements.

Information collected within the statistical framework is more detailed. It concerns both imports and exports but is not always exhaustive. Firms involved in intra-EU trade, whether as an exporter or as an importer, fill a DEB form concerning all transactions that have occurred over the last month towards a given EU partner and for a given product. The list of variables somewhat varies depending on i) the annual value of trade undertaken by this firm and ii) the "procedure" under which the transaction is made. Note that the value which is reported under the statistical framework is somewhat different to the "fiscal" value reported by French exporters because it not only includes the value of goods but also transportation and distribution services up to the French border. Duties, taxes and excise duties are excluded from the statistical value. Following the usual trade convention, export flows are valued FOB while import values are inclusive of insurance and freight costs.

Custom procedures: In the DEB form, the customs *procedure* qualifies the nature of the transaction. There are seven different procedures which are listed in Table 1. Transactions recorded under three of these procedures concern trade flows which are not included in official trade statistics because they do not actually involve the physical displacement of a good (see details in Table 1). Except if otherwise specified, the French Customs provide researchers with the sub-sample of transactions included in official trade statistics. This represents more than 99% of the overall number of trade flows recorded in the DEB database.[7]

This leaves the user of French data with two alternative custom procedures for each type of intra-EU trade flows (*flux* 3 and 4). The most common procedures, *Regimes* 11 and 21 for imports and exports respectively, concern transactions over standard taxable goods. *Regimes* 19 and 29 concern transactions over non-taxable goods, such as gifts, or transactions linked to a tolling agreement. They typically concern less than 5% of recorded transactions. Since these transactions do not induce the payment of a tax, the VAT number of the European partner is not necessarily provided.

Declaration thresholds: The number of variables collected under the statistical framework depends on the firm's size, as measured by the overall value of intra-EU trade the firm is involved in during the current civil year. [8] Depending on its size, the firm falls into a given stringency level which varies between 1 and 4, 1 being the most demanding regime. The size thresholds corresponding to each level are reported in Table 2. Notice that there are important changes over time in the definition

[7] Additional transactions excluded from the scope of the DEB data include imports delivered in France to a foreign resident, trade with overseas department, goods traded for repair operation purposes, temporary exports, bunkering, placement in bonded warehouse of goods other than oil, monetary gold transactions, and satellites.

[8] Since the declaration thresholds are based on the value of trade over the current year, it can happen that a firm passes a threshold during the year. In such case, it immediately starts providing information about the additional variables which it was exempted to disclose while below the threshold. This is one reason why, for instance, a number of firms declare a value of imports which does not cumulate to the threshold below which firms are not asked to declare their intra-EU imports. Another reason for why this can happen is that the regime under which a firm starts declaring is based on the cumulated value of its trade the year before. Thus a firm that has imported more than 460 thousands euros in $t-1$ starts declaring its imports in t, even if it may not actually pass the bar of 460 thousands euros during this particular year.

Table 2: Size thresholds, over time

	1993-1997	1998-2000	2001	2002-2006	2007-2010	2011-2018
Exports						
1	>10,000	>15,000	>15,000	>2,300	>2,300	>460
2	[1,400;10,000]	[3,000;15,000]	[3,000;15,000]	[460;2,300]	[460;2,300]	
3	[250;1,400]	[250;3,000]	[650;3,000]	[100;460]	[150;460]	
4	<250	<250	<650	<100	<150	<460
Imports						
1	>10,000	>15,000	>15,000	>2,300	>2,300	>460
2	[700;10,000]	[1,500;15,000]	[1,500;15,000]	[230;2,300]	[230;2,300]	
3	[250;700]	[250;1,500]	[650;1,500]	[100;230]	[150;230]	
4	<250	<250	<650	<100	<150	<460
Units	1,000 FF	1,000 FF	1,000 FF	1,000 €	1,000 €	1,000 €

Notes: This table reports the time-varying size intervals corresponding to decreasing stringency levels for intra-EU customs declarations.

of these thresholds, which can induce substantial selection biases. The most important revision took place in 2011 with a substantial increase in the size threshold corresponding to the lowest stringency level and the simplification into two regimes. Finally, the French customs do not always provide researchers with transactions flows declared under the less stringent level (*Obligation* 4). Interested researchers must specify in their data application that the data coverage must be exhaustive.

Table 3 displays the list of the most useful variables requested in the DEB form as a function of the stringency level.[9] In the less stringent procedure, importers are not compelled to a DEB declaration while exporters record the value of the transaction and the identity of the European partner. Export data are thus almost exhaustive but the simplified procedure does not provide key variables used in many empirical works, most notably the product category.

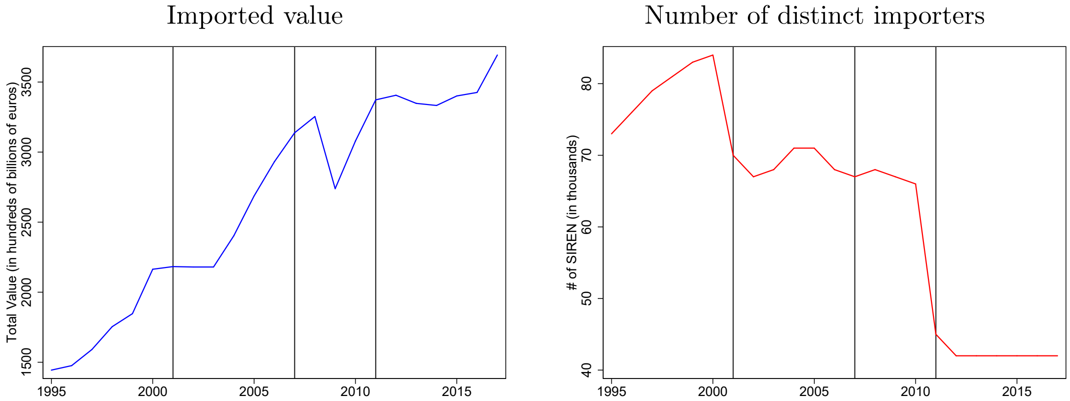
Table 3: Variables available in the intra-EU statements, as a function of the stringency level

Variable		**Stringency level**				**Note**
		1	2	3	4	
flux	Type of trade flow	✓	✓	✓	Export	3 = Import, 4 = Export
siren	French-firm identifier	✓	✓	✓	Export	9-digit Siren identifier
numvat	Id. foreign partner	Export	Export	Export	Export	Anonymized EU-VAT number
regdem	Customs Procedure	✓	✓	✓	Export	See Table 1
oblig	Stringency level	✓	✓	✓	Export	See Table 2
an/mois	Date (month-year)	✓	✓	✓	Export	Date of the declaration
valfac	Fiscal value	✓	✓	✓	Export	In euros
temo	Transportation mode	✓	✓	Since 2011	Since 2011, Export	See Table B1
nc8	cn8 product	✓	✓	✓		Current CN8 nomenclature
cpa6	cpa6 product	✓	✓	✓		Time-invarying product code
pyod	Destination/origin country	✓	✓	✓		iso2. See Tables B4 and B5
pypd	Last/next country of transit	✓	✓	✓		iso2. See Tables B4 and B5
natr	Nature of transaction	✓	✓			See Table B3
kgs	Quantity (kg)	✓	✓			
usup	Quantity (Physical units)[10]	✓	✓			List of products in cn8_usup.txt
incoterm	Incoterm	Until 2006	Until 2006			
dept	Departement	✓	✓			See Table B2
valstat	Statistical value	Until 2006				In euros

[9] The full list of variables collected can be found by looking at the DEB form at the end of the note.

The absence of collected data for intra-EU imports below a certain threshold implies a selection bias, which is all the more problematic since its size is likely to change over time, when the declaration threshold is updated in 2001, 2007 and 2011 (See Table 2). Quantifying the size of the bias precisely is almost impossible, although indicative evidence are provided in Figure 1. Intuitively, the size of the bias can be assessed by comparing statistics regarding intra-EU imports, before and after a change in the declaration threshold.

Figure 1: Evolution of intra-EU imports over time


Imported value

Number of distinct importers

Notes: Value and number of French firms recorded in the French intra-EU import files (DEB, *flux* 3), over time. The vertical lines correspond to the years of a declaration threshold adjustment.

Figure 1 shows the evolution over time in the value of imports (left panel) and the number of distinct importers (right panel) recorded in the DEB data. Vertical lines materialize changes in declaration thresholds. Starting with the left panel, there is no obvious discontinuity in the value of overall imports, before and after changes in the declaration threshold. This outcome is not surprising since the declaration threshold concerns very small importers, which are well-known to contribute little to overall imports (Bernard et al., 2009; di Giovanni et al., 2018, on US and French firms, respectively). The right panel in Figure 1 however shows strong discontinuities, notably in 2001 and 2011, when the number of distinct importers drops significantly. Our interpretation of these discontinuities is that a significant number of French importers purchase goods for a value in between the old and the new thresholds and are thus no longer required to fill the DEB form, after the reform. The subsequent decrease in the extensive margin of trade reflects a statistical bias, which is all the stronger since the threshold adjusts more.

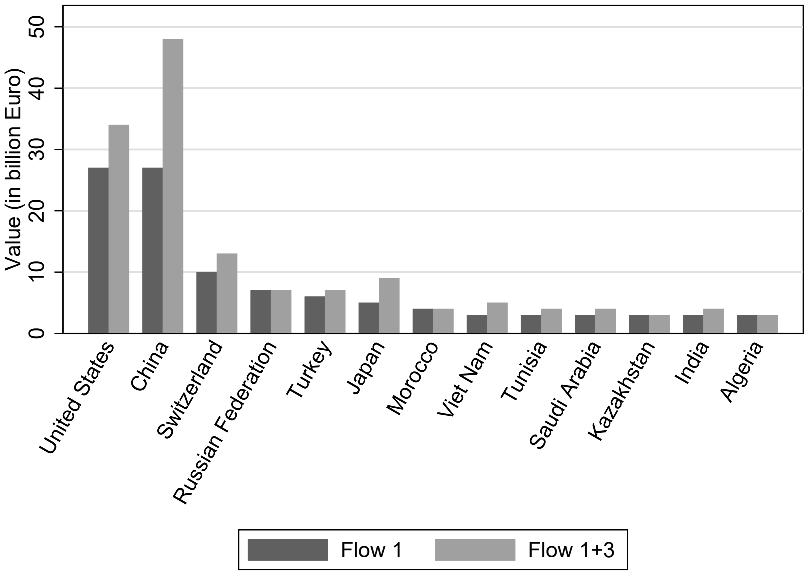
Third party trade: One caveat of intra-EU customs data is that goods in transit are included in the scope of collect. For instance, goods shipped to a European destination, say Belgium, to be further exported outside of the EU are recorded into a DEB export form with the EU country as the destination and the identity of the European intermediary for the importer. Likewise, some extra-EU officially French exports might actually correspond to export flows where France solely plays the role of an intermediate. This data restriction may contribute to explaining why Belgium is the first destination of French exports in terms of the number of transactions, although the country is substantially smaller than other neighboring countries such as Germany, the first destination in terms of the value of trade. The port of Anvers is indeed the second largest cargo port in Europe and is thus likely to intermediate a substantial share of French exports to remote countries. Unfortunately, quantifying the share of exports to Belgium (or the Netherlands) which transits to these countries before being exported further away is impossible and there is nothing that the researcher can do to correct for the bias.[11]

The situation is slightly better for extra-EU import flows. Here, it is fairly common that the country declared to be the *origin country* (recorded under the "pyod" variable) is not the same as the

[10] Since 2006, it is not required to provide the weight of the product whenever the physical quantity is required.

[11] Another form of third-party trade can be identified in the DEB export data, but is much less important quantitatively. Namely, it can happen that the recorded importer is located in a country which is not the destination country, because the importing firm has requested the French exporter to send the goods to a third party located in another country than its own, for instance an affiliate located in another country. Such transactions can be identified using the first two characters of the importer identifier to determine the location of the firm purchasing the good and the destination country as the country where goods are shipped.

Figure 2: Distribution of imports, by country of origin, in 2017



The graph compares the value of imports, by origin country, recorded in the DAU data (dark grey bars) and complemented with transactions recovered from the DEB forms using the information recorded in the "pyod" variable (light grey bars). Countries are ranked according to their decreasing importance in the DAU data.

country from which the good entered into France (recorded under the "pypd" variable). In particular, many goods enter France through Belgium but originate from China or the US. Using the "pyod" information to identify the origin of the import transaction allows to recover the "true" import flow, between the producing and the consuming country. Figure 2 shows that this is not a quantitatively small issue. Namely, the graph compares, in 2017, the value of imports by extra-EU origin country, when only the information in DAU is used (dark grey bars) versus when the information in DAU is complemented with trade flows recovered from the DEB forms (light grey bars). About one third of imports from China enter France through a third EU country. The proportion is smaller, although significant as well, for imports from the US. Note that, by default, when researchers do not ask access to the DEB and DAU forms separately, the French customs provide them with a single sample of firm-level imports that keeps the information in "pyod".

2.2 Data on extra-EU trade (DAU)

Data collected under the DAU form concern all extra-EU trade flows. The structure of the dataset is relatively standard which explains that this section is significantly shorter than the previous one.

Custom procedures: A large variety of customs procedures (*Régimes*) exist for extra-EU flows. The precise list of these procedures is available upon request but we do not detail it in this note since all these procedures contain transactions which are included in official trade statistics. For this reason, and except otherwise specified, the French customs do not provide researchers with information on these procedures.

Declaration thresholds: The extra-EU DAU data are almost exhaustive, since any transaction between France and a foreign country is recorded. Before 2010, there was a declaration threshold of 1,000 euros (or 1,000 kilos) below which firms were exempted from a declaration. This exemption no longer exists.[12] Figure 3 shows the evolution in the number of transactions below the threshold observed in the DAU database. As expected, this number starts being significant after 2010 to reach

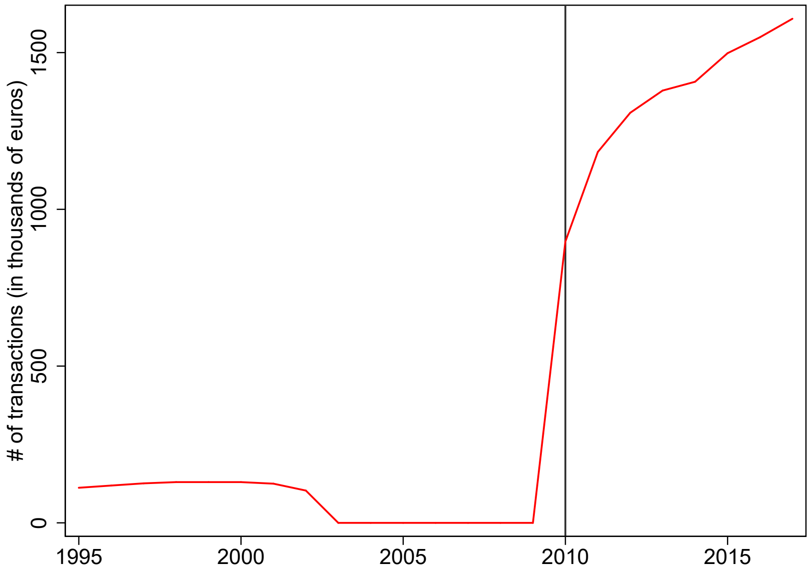
[12] There are a few exceptions, though, such as individuals' postal shipments under 200 euros which are still exempted from a declaration.

Table 4: Variables available in the extra-EU statements

	Variables	**Note**
statflux	Type of trade flow	1 = Imports, 2 = Exports
siren	French-firm identifier	
an/mois	Date (year/month)	Date of the declaration
prifac	Fiscal value	In euros
cn8	cn8 product	Current CN8 nomenclature
dest/ori	Destination/origin country	iso2 code. See Tables B4 and B5
nattrans	Nature of transaction	See Table B3
msn	Quantity (kg)	
unispe	Quantity (Physical units)	List of products in nc8_usup.txt
modfrotra	Transportation mode	See Table B1
depexp/depliv	Departement of origin/destination	
valstat	Statistical value	In euros CIF (Imports)/FOB (Exports)
codliv	Incoterm code	
devfac	Invoicing currency	Since 2011

more than 1.5 million transactions per year. But the cumulated value of these flows is naturally negligible, less than .5% of the total value of annual trade.[13]

Figure 3: Number of transactions below 1,000 kilos and 1,000 euros in the extra-EU export data (DAU, *flux* 2)



Note: The vertical line corresponds to the reform in 2010.

The list of the most commonly used variables available in the DAU forms is provided in Table 4.[14] Most of the variables collected under the most stringent level of the DEB procedure are there, together with additional variables such as the invoicing currency. For extra-EU trade, the fiscal and the statistical values are different. The fiscal value is the one declared by the firm, based on its invoice. The statistical value is computed by the Customs using this information and the contractural details provided in the "incoterms". The staistical value is FOB for exports and CIF for imports. It is the value provided to the researchers by the customs, except otherwise stated.

[13] Note that the impact is not equally shared across sectors since some products, such as wine or auto parts are often shipped in batches of less than a thousands euros/kilos.

[14] The complete list of collected information can be found in the DAU form copied at the end of the note.

3 Cleaning data steps

Even though the quality of the data provided by the French customs is high, we recommend a few trimming steps described now. The Stata code implementing these steps is available on the companion website. The objective is to drop invalid and missing data. Table 5 displays the share of the total value of trade which is disregarded at each particular stage.

Table 5: Impact of the trimming: Share of the total value dropped at each step

Trimming step	**% Value**
Imports	
Invalid CN8 code	0.01
Invalid country code	1.3
Unknown or invalid SIREN	0.69
Exports	
Invalid CN8 code	0.01
Invalid country code	0.12
Unknown or invalid SIREN	0.82

Drop invalid French firm identifiers: SIREN numbers are used in the customs dataset to identify French firms. While nowadays, 99% of SIREN numbers are considered valid, older data display a number of identifiers which do not correspond to an existing French firm. The INSEE-SIRENE database can be used to check the validity of a particular siren but is only available for the most recent years. In the absence of such data, a number of codes can nonetheless be dropped, namely numbers that start with at least 5 zeros as well as the few codes listed in Table 6.

Table 6: Special SIREN numbers.

SIREN	**Meaning**
000000000	Missing or unknown SIREN e.g. New born firm
777777777	Group of firms (mostly used in export flows)
222222222 and 202020202	Unknown SIREN for some transactions below the 1000 euros/1000 kgs threshold
888888888	Foreign firms without tax representative in France
999999999	Occasional/non-trader individual (DEB only)
111111111	Monaco (DEB only)

Drop invalid country codes: In the customs data, countries are identified using 2-digit ISO 3166-1 alpha-2 codes (ISO-2). The raw data display a number of transactions with invalid or recoded ISO-2 codes. Table B4 in Appendix provides the list of country codes that are generated by the French customs in specific contexts. Table B5 finally lists a number of changes affecting country codes during the period of data availability. The most important change concerns Belgium and Luxemburg, which trade statistics were grouped under a single country code ("XU") until 1999. When working with longitudinal data, the researcher has no choice but to artificially collapse data observed for Belgium and Luxembourg, after 2000.[15] Invalid country codes such as "FR" as a destination country or "EU", sometimes mistakenly used instead of "US" are dropped as there is no way to recover valid information from such transactions.

As explained in Section 2, the raw DEB files contain two different variables for the country of origin of imports, namely the country of origin ("pyod") and the country from which the good entered France ("payp"). Except if otherwise specified, the French customs provide researchers with a single variable, namely the country of origin which is the relevant definition of the producing country in standard trade analysis. This is the concept of country of origin used in the DAU form.

[15] Note that a slightly more sophisticated solution can be proposed, which exploits the information available in the anonymized VAT numbers for the French firm's partner. Namely, the first two characters of this number correspond to the firm's location, which is either BE or LU for firms located in Belgium and Luxemburg, respectively. As a consequence, one can infer from this information the actual destination of the French firm's exports.

Table 7: Invalid cn8 codes

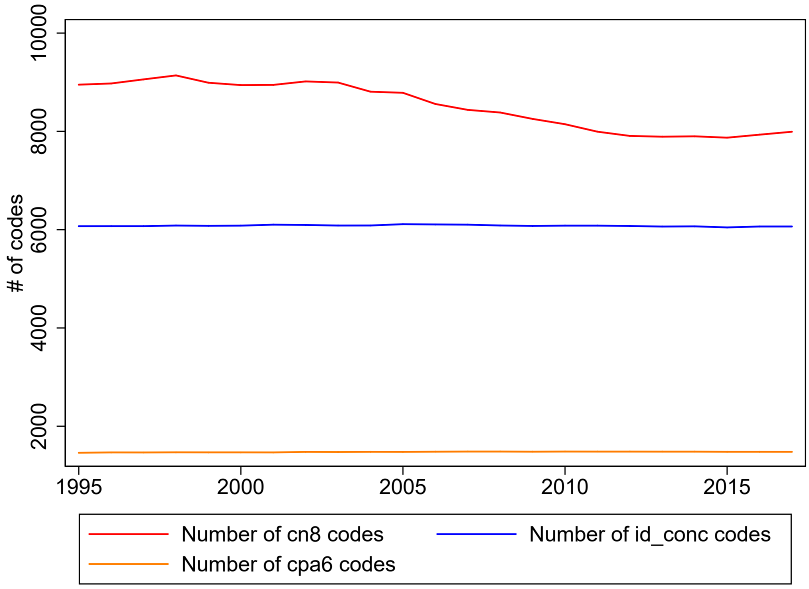
CN8 code	Meaning
9880XX00 - XX being 73, 84, 99 or 85 -	Large manufacturing or commercial firms which import values above 3 million euros and benefit from a simplified code grouping all products together
99050000	Personnal belongings of migrants
99190000	Personal belongings of victims of disasters (coffin...)
9930	Refuelling
9931	Goods sent to offshore infrastructures

Keep valid CN8 codes: A number of special product codes are present in the raw data and need to be dropped to conduct further analysis on French foreign trade at the product level. The full list of these codes is provided in Table 7. The most common invalid code is 99 5000 00. It indicates small transactions of less than 200 euros, in intra-EU trade flows. This code is meant to allow small operators to group transactions realized over several products with the same partner, whenever the cumulated value of these transactions is sufficiently small. This code accounts for more than 95% of dropped products in intra-EU data.

Another, more important, treatment to product codes needs to be performed to conduct panel data analysis at a fine disaggregation level. The reason is that the Combined Nomenclature is regularly revised to take into account changes in the nature of traded products. Major revisions are observed when the Harmonized System is itself updated, in 1996, 2002, 2007, 2012 and 2017. However, more minor revisions to the combined nomenclature also occur each year. These frequent revisions imply that product codes are not always consistently defined over time. Some disappear and new ones are introduced with no one-to-one mapping between the old and new codes. Appendix A extensively discusses this question. We notably describe the algorithm that we recommend, to harmonize product codes over time. The algorithm is taken from Behrens and Martin (2015) and consists in identifying the smallest groups of products that are linked together through one or several revisions of the combined nomenclature. Once identified, these groups of products are assigned a common, time-invariant, new product code which can be applied to individual transactions. In the rest of this paper, this new product code is given the "id_conc" label. Implementing this procedure implies a loss of information in the cross-section since transactions recorded under several product codes can end up grouped together under a single id_conc code. The benefit is to restore the comparability of product categories, over time. Absent this harmonization, a researcher using product-level panel data would overestimate the churning of products, since nomenclature revisions would be wrongly interpreted as a change in the traded product. An alternative would be to use the cpa6 nomenclature, also provided in the customs' data. This nomenclature aggregates products from 1990 to 2018 and is always valid over this time span. However, these 6-digit codes represent significantly more aggregated products and are more difficult to match with external product-level data, which are often available in the harmonized system at the root of the combined nomenclature.[16]

[16] It has to be noted, however, that cpa4 has a direct correspondance with NAF5 rev2 (code APE) and can be matched with NACE rev2, i.e. it can be useful to match trade flows with firms' activity.

Figure 4: Evolution in the number of cn8, cpa6 or id_conc products involved in French exports



Note: Total number of distinct cn8, cpa6 and id_conc categories in export data between 1995 and 2017. id_conc categories are defined by applying the harmonization algorithm over the whole period.

In appendix A, we provide quantitative elements to help the reader evaluate the impact of the algorithm regarding both the upside and downside consequences. As we discuss, the impact varies greatly depending on the period of analysis. Without entering into such details, Figure 4 shows the evolution in the number of product categories in the raw export data, when either the cn8 codes or the time-invariant id_conc and cpa6 categories are used to classify transactions. Working with harmonized product codes implies working with slightly more aggregated data since the number of categories falls to around 6 thousands against between 8 and 9 thousands in the raw cn8 data. However, this number is roughly constant over time which is not the case of the number of cn8 products. The decrease in the number of cn8 codes between 2000 and 2010 explains by more frequent aggregations of cn8 codes in recent revisions. It is thus an artifact of nomenclature updates. Finally, working with cpa6 codes implies aggregating the data significantly more than with id_conc categories, since the number of such codes is around 2,000.

Besides these three trimming steps, the set of do files which we provide in the companion website also contains a short algorithm which is meant to recover some missing cn8 product codes regarding export transactions of firms which are eligible to the less stringent DEB declaration and thus declare the value of the transaction but not the product under study. The idea of the algorithm is to exploit information over firms which pass the threshold during a year. As explained in Section 2, these firms start filling a larger number of variables, including the cn8 code, immediately when they pass the threshold. When these transactions concern the same partner which they were already serving before, one can assume that the relationship has not changed in nature and the cn8 product for the beginning of the year transactions is the same as in the end-of-the-year transactions.[17] We use the same strategy when the firm passes the threshold from one year to the next. This strategy allows to recover 5% of the Firm×Buyer pairs, i.e. transactions, and 12% of the cumulated value of transactions which cn8 code is missing. With this methodology transactions with the product code informed (whether recorded or imputed) represent 88% of the raw number of observed siren× buyer couple and 99% of traded values.

Choose a variable for quantities: As recorded in Tables 3 and 4, the customs files contain two distinct variables for the physical quantities. The first variable is the weight of the traded goods, defined in a uniform unit, namely kilograms.[18] The second variable, which is filled for a subsample

[17] Note that this is not possible when the firm declare exporting two different products to the same partner, in which case we leave the product codes unchanged.

[18] Note that the weight is rounded to the closest kilogram and thus there can be quantities of 0, for shipments of less than 500 grams.

of *cn*8 products represents quantities in some specific physical units, such as pairs of shoes or square meters of fabrics. In 2017, this concerns 2,628 products out of a total of 9,637 categories. The first variable presents the advantage that it is directly comparable across products. However, reported quantities can suffer from important measurement issues. Exporters might not know the exact weight of their goods. They might include packaging, which is not supposed to be included in the declared weight. Finally, some goods have a weight that can vary a lot from one variety to the other without this variation being especially meaningful, e.g. chemical wood pulp. In such examples, the quantity in physical units is probably more accurate, but its coverage is rather limited.

Since the reliability of the weight variable is likely to vary a lot across products, researchers using this variable might want to control for unobserved measurement errors at the product level using fixed effects. Once unobserved heterogeneity across products is controlled for, however, it seems more natural to use the most accurate information available in the data, which is likely to be the quantity in physical units, when requested. As a consequence, the trimming code available online creates a new variable for quantities which corresponds to physical units, when available, and the weight of the traded goods, in the absence of better information. As a side benefit, using this variable allows solving a statistical problem of the Customs data, which is that the coverage of the weight variable varies over time. The reason is that the weight of traded goods was no longer requested for products for which a quantity in physical units is requested, between 2006 and 2010. As a consequence, researchers using the weight as their preferred measure of quantities have observed a substantial drop in the data coverage, around these dates. This problem no longer exists when information on the weight and physical quantities is combined.

Using our preferred measure of course implies that quantities are no longer comparable across products, in levels. While this is not a problem in many cases, this might be for specific research questions. In that case, researchers have no choice but to use the quantity in kilograms as their preferred measure as its coverage is wider. He/she however needs to take into account the decreased coverage between 2006 and 2010. One option is to infer the weight variable using information on the physical units and a self-made correspondance between physical units and weights. Building such correspondance is doable since some firms continue to declare both the physical units and the quantities, after 2006. However, the accuracy of such imputation might be rather limited.

Before concluding, note that our preferred strategy for measuring traded quantities suffers from one caveat, that does not have a straightforward solution. The problem arises when the algorithm used to treat quantities is combined with the harmonization procedure for products. Then, it can happen that two products which quantities are defined in different units end up grouped into the same id_conc category. Since id_conc is the new product category that we argue should be used, keeping this information would amount to comparing apples and oranges. Our conservative solution consists in neglecting any information on quantities in such case.

4 Descriptive statistics

After having applied the trimming steps described in Section 3, we are left with data covering the 1995-2017 period.[19] We now describe the corresponding aggregated sample through statistics over various dimensions of the data. We discuss how the coverage varies depending on the variables of interest.

4.1 Coverage

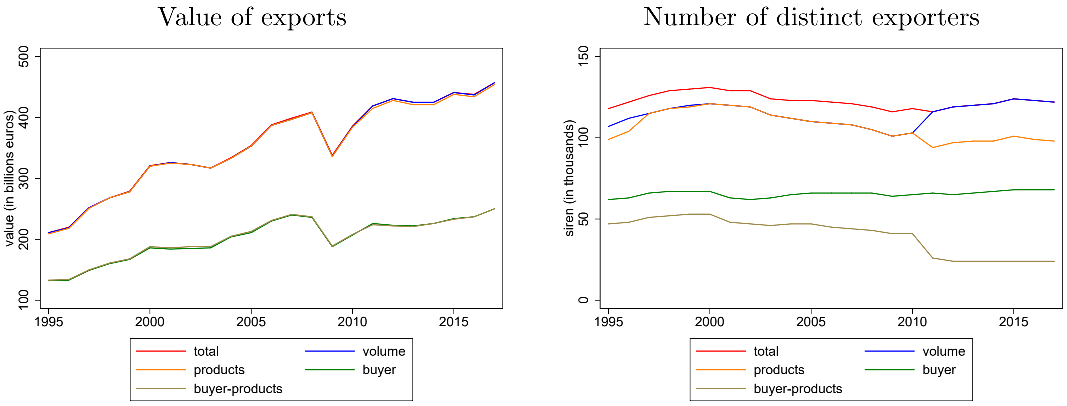
At the most "aggregated" firm×month×year level, data are exhaustive on the export side but do not cover importers below the lowest threshold reported in Table 2. Beyond the value of traded goods, trade economists are often interested in gathering information at the *product* level, which is not possible for export transactions below the lowest declaration threshold (see Table 3). The coverage is further reduced, to firms at stringency levels 1 or 2, when the *volume* of trade is also of interest. Finally, statistical analysis can also be performed at the firm-to-firm level, using the information on the exporting firm *and* its foreign partner available in the DEB form, but the coverage is further reduced to intra-EU exports. Figures 5 and 6 show how the data coverage varies across these various subsamples, for exports and imports, respectively. In these figures, the left panel measures the coverage

[19] Data available to researchers are available from 1993 but the treatment of changes in product nomenclatures is done from 1995 to 2017 which is thus the period of analysis that we consider in this note.

in terms of the value of traded goods and the right panel shows the number of distinct French firms in the various sub-samples.

Consider first Figure 5, left panel. Clearly, the overall value of exports is very comparable in various sub-samples covering the same destination countries, even when the use of product-level or quantity data excludes firms declaring under the less stringent customs regimes. Exploiting the firm-to-firm dimension solely available in the DEB data obviously reduces the overall coverage, since intra-EU export flows represent around two thirds of France's aggregate exports. Without much surprise, differences across sub-samples are more pronounced in terms of the number of distinct exporters covered (right panel). First, the number of intra-EU exporters is about half the total number of French exporters.[20] Second, working with product-level data implies loosing roughly 10-20% of exporters. Finally, the 2011 reform on declaration thresholds induces significant discontinuities in the overall population of exporters. On the one hand, the increased declaration threshold at the most stringent level implies that the number of exporters covered in product-level data reduces signficantly. On the other hand, the simplification into two regimes implies that an increasing number of exporters which are above the new lowest declaration threshold are requested to declare quantities, which improves the coverage of the associated sub-sample.

Figure 5: Value of exports and number of different exporting firms, across sub-samples



Note: This figure shows the value of exports (left panel) and the number of distinct exporters, in various sub-samples of export data. The red line labelled "Total" corresponds to the maximum sample coverage, when the value of exports and the exporter's identity is the only variable of interest.The blue line, labelled "volume" corresponds to the sample obtained when investigating on the traded volume, in kgs or in physical units. The orange line refers to the coverage obtained when focusing on products and thus needing cn8 codes. Both green lines correspond to the sample coverage when the foreign buyer is studied. This dimension is only available for intra-EU exports, thus explaining the relatively low coverage.

Conclusions are roughly similar when comparing sample coverages for import data (Figure 6). Here, things are a bit simpler since no information on firm-to-firm trade is available, and thus only three sub-samples need to be compared. Once again, the overall value of imports is very similar in these sub-samples (left panel) but the *number* of importers varies (right panel). Moreover, none of these covers firms importing from Europe below the lowest declaration threshold (see the discussion in Section 2). As a consequence, the number of importers providing information on the imported products is roughly similar to the total number of importers covered by the data.

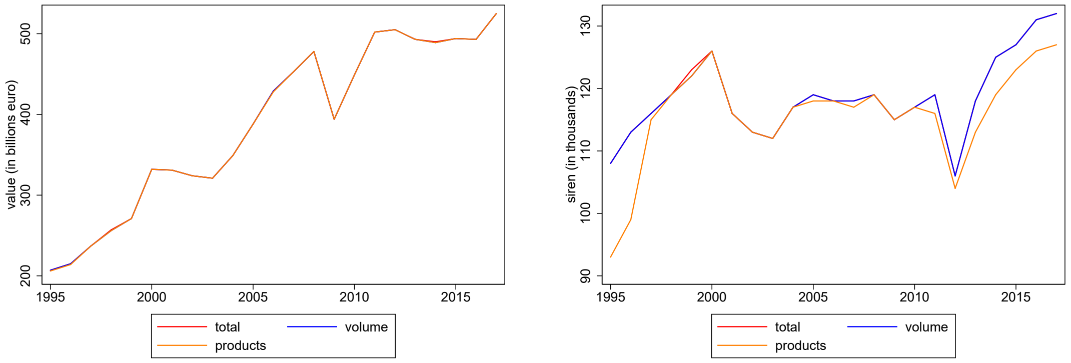
4.2 Additional Summary Statistics

Multi-product firms: Figure 7 provides statistics on multi-product firms, on the export and the import sides, in 2017.[21] Here, products are those defined by the harmonization algorithms and are thus slightly more aggregated than the usually-used cn8 ones. In 2017, around 60% of exporters export more than one product, but represent 98% of the total value of exports. The corresponding numbers are roughly the same when considering multi-product importers (right panel). This is consistent with the view that there is selection into exporting/importing multiple products with larger firms being more likely to be multi-product. The impact of self-selection is further illustrated by the other

[20] Note that this is true even if, on average, firms are more likely to export in Europe than further away (Eaton et al., 2011). The reason is that exporters serving distant destinations are very heterogeneous in terms of which destination(s) they serve and thus the overall population of firms serving at least one non-EU country is relatively large.

[21] See Mayer et al. (2014) for a more systematic analysis of these multi-product firms.

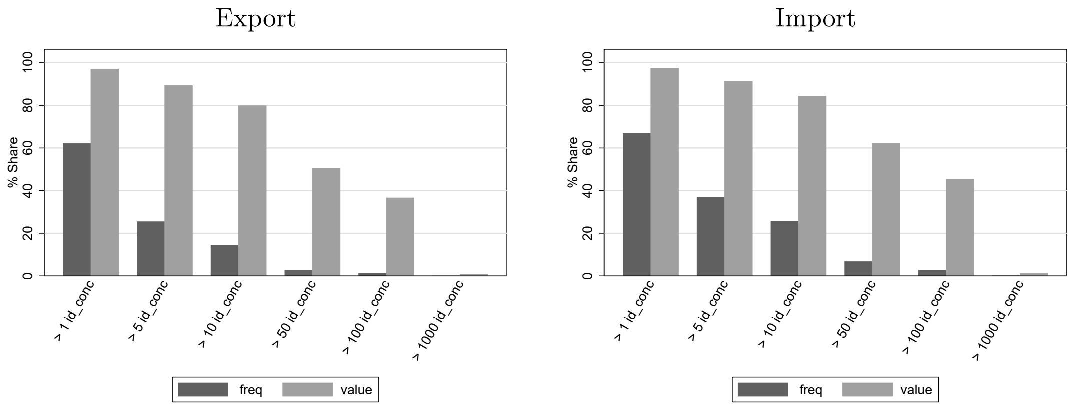
Figure 6: Value of import and number of different importing firms, across sub-samples



Note: This figure shows the value of imports (left panel) and the number of distinct importers, in various sub-samples of import data. The red line labelled "Total" corresponds to the maximum sample coverage, when the value of import and the importer's identity are the only variables of interest. The blue line, labelled "volume" corresponds to the sample obtained when the volum of imports is also of interest. The orange line refers to the coverage obtained when the product dimension is additionally exploited.

bars on the graphs, when the analysis is restricted to firms exporting/importing a larger number of products. The number of such firms decreases very quickly. However, their prevalence in overall trade remains very large, at least up to 10 products. Above this number, both the number of firms and the value of their trade declines more quickly which might be a consequence of the remaining firms being increasingly likely to be intermediaries.

Figure 7: Distribution of multi-product firms in French trade, 2017

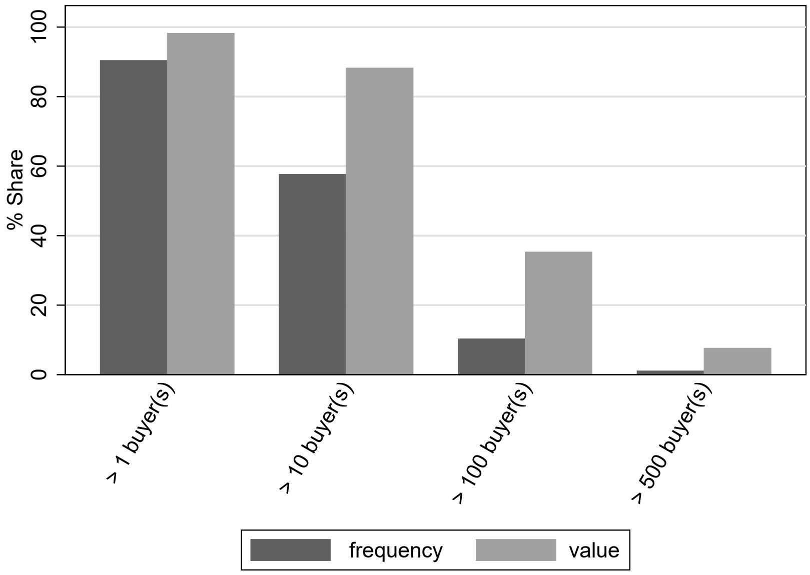


Note: The graph compares the share of transactions (dark grey bars) and the value share of firms exporting/importing an increasing number of products.

Exporters and their partners: Figure 10 further exploits the firm-to-firm dimension to document the extent of heterogeneity across exporters regarding the number of individual partners they serve in Europe, still in 2017.[22] This exploits the information collected under the fiscal framework through the DEB form. When more than 20% of firms have only one partner in Europe, they represent a tiny share of intra-EU French exports. This trend is further confirmed when considering that there are more than 60% of firms with less than 10 buyers but that they represent around 10% of the exported value. Here as well, there is selection into exporting to multiple buyers with larger firms being more likely to have more partners.

[22] See Kramarz et al. (2016) and Lenoir et al. (2018) for a more systematic analysis of this dimension.

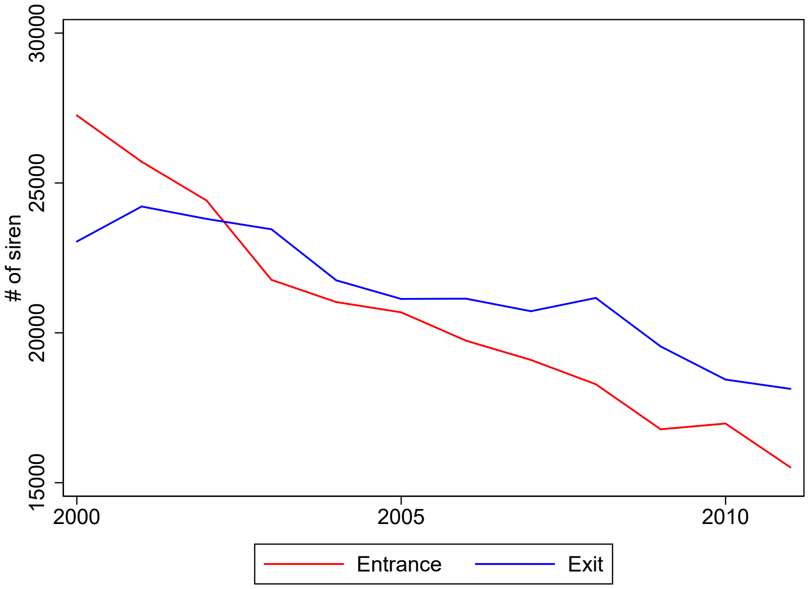
Figure 8: Distribution of the number of buyers of French firms, 2017.



Note: The graph compares the share of transactions (dark grey bars) and the value share of firms sevring to increasing number of foreign partners in Europe.

Entries and exits: An extensively discussed topic in the trade literature concerns the dynamics of exports and the share of it which explains by adjustments at the extensive margin. Figure 9 provides additional statistics on this dimension. Namely, the net entry of firms which can be inferred from Figure 5 is further decomposed in terms of the number of new entrants and the number of firms which have stopped exporting between the previous and the current year. What this figure shows is that there is a lot of churning in the population of exporters. If the net entry of firms is relatively limited, roughly between -3,000 and + 3,000 firms per year, the number of entering *and* exiting firms is substantial. This supports evidence discussed in the previous literature that "sequential exporting" is a quantitatively important phenomenon (Albornoz et al., 2012).

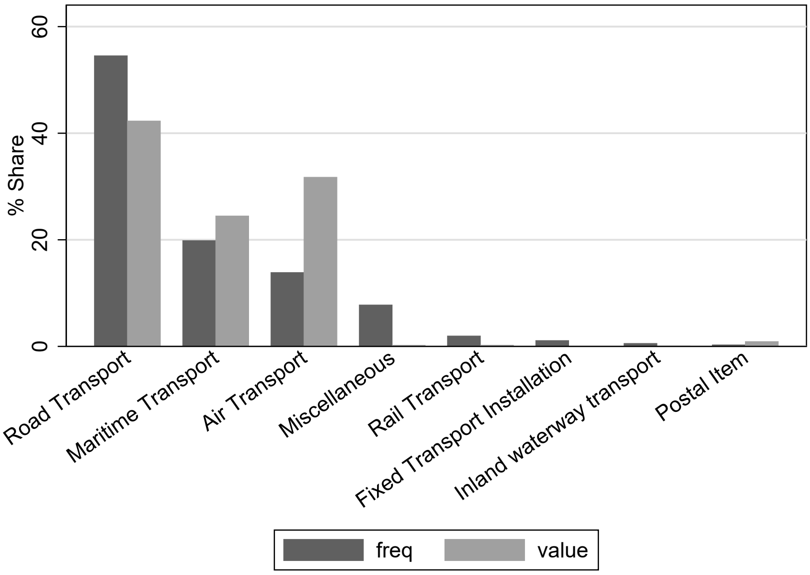
Figure 9: New and exiting French exporters, across years.



Note: The Figure shows the number of new entrants and exiting firms in export data, over time. A new entrant is defined as a French firm which appears in the export data for the very first time. An exiting firm is a firm which was present in the data the year before but no longer is (and will never be afterwards). Left and right censoring is controlled for by focusing on the 2000-2011 sub-period while using the 1995-1999 period as the reference for new entrants and the 2012-2017 period as the reference for exiting firms.

Transportation modes: Figure 10 illustrates the prevalence of various transportation modes, in export data for 2017. The main transportation mode, both in frequency and in value terms, is the road, followed by maritime and air transportation modes. This is explained by the geography of French exports, which is dominated by intra-EU exports, easily served by trucks. Boats and airplanes are also prevalent. Interestingly, air transport is substantially more prevalent in value terms than as a percentage of shipments, which means that high-value goods are more likely to travel by air. Other transportation modes such as trains remain very marginal in nowadays trade.

Figure 10: Distribution of transportation modes for French exports, 2017.



Note: The graph compares the share of firms (dark grey bars) and their value share (light grey bars) by transportation mode.

Figure 11 further shows how this varies depending on the destination country of exports (left panel) and the origin country of imports (right panel). As expected, road transportation is substantially more prevalent in European trade while maritime and air transportation modes are used for trade with the US, China, Japan and other Asia countries. This graph also allows discussing the accuracy of the transportation mode variable. At first view, it might be surprising that a non-zero share of exports to and imports from the US is declared to be transported by raod. There are several explanations for this. First, for multi-mode trade flows, the firm is supposed to report the “most important” one, which might be a bit subjective. For instance, if the good is sold by a firm in Texas, transported by road to an airport on the East coast and shipped to France, the importer in France might consider that the most prevalent transportation mode is the road. Second, for import flows, a substantial share of imports from non-EU origin countries is recovered from the DEB declarations (See Section 2). In such case, it is very likely that the transportation mode which is reported is the one used to transport goods from the EU country where the good entered the EU to France, most likely the road. Finally, there are probably a number of mis-reported transportation modes. This suggests that the information reported in this variable must be used carefully. Still, the relative shares of each mode reported in Figure 11 suggest that the information is, on average, reliable.

References

Albornoz, F., Calvo Pardo, H. F., Corcos, G., and Ornelas, E. (2012). Sequential exporting. *Journal of International Economics*, 88(1):17–31.

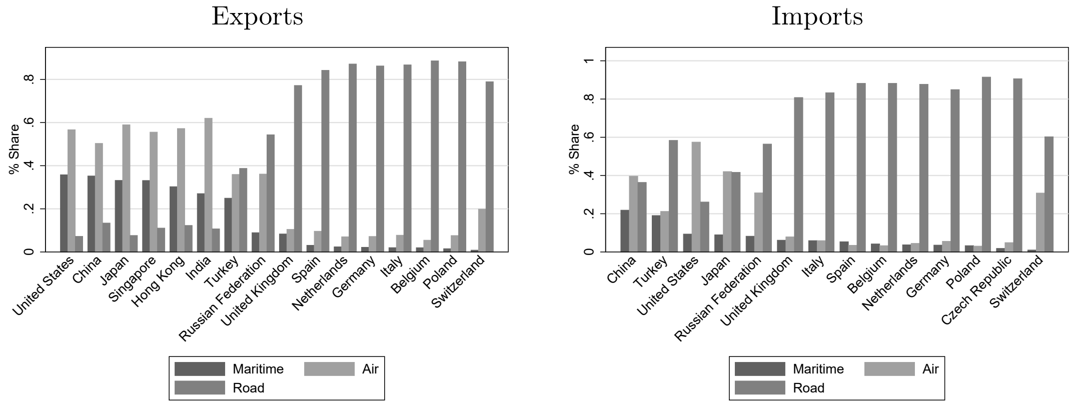
Behrens, K. and Martin, J. (2015). Concording large datasets over time: The C^3 method. Unpublished paper.

Berman, N., Martin, P., and Mayer, T. (2012). How do different exporters react to exchange rate changes? Theory, empirics and aggregate implications. *Quarterly Journal of Economics*, 127(1):437–492.

Bernard, A. B., Jensen, J. B., and Schott, P. K. (2009). Importers, Exporters and Multinationals: A Portrait of Firms in the U.S. that Trade Goods. In *Producer Dynamics: New Evidence from Micro Data*, NBER Chapters, pages 513–552. National Bureau of Economic Research, Inc.

Blaum, J., Lelarge, C., and Peters, M. (2018). The Gains from Input Trade with Heterogeneous Importers. *American Economic Journal: Macroeconomics.*

Figure 11: Probability of a particular transportation mode as a function of the destination/origin country, 2017.



Exports

Imports

Note: The Figure reports the relative prevalence of Maritime, Air and Road transportation modes, by destination country (Left panel) and by origin country (Right panel).

Bricongne, J.-C., Fontagné, L., Gaulier, G., Taglioni, D., and Vicard, V. (2012). Firms and the global crisis: French exports in the turmoil. *Journal of International Economics*, 87(1):134–146.

Crozet, M., Head, K., and Mayer, T. (2012). Quality Sorting and Trade: Firm-level Evidence for French Wine. *Review of Economic Studies*, 79(2):609–644.

di Giovanni, J., Levchenko, A., and Mejean, I. (2018). The Micro Origins of International Business Cycle Comovements. *American Economic Review*, 108(1):82–108.

Eaton, J., Kortum, S., and Kramarz, F. (2011). An Anatomy of International Trade: Evidence From French Firms. *Econometrica*, 79(5):1453–1498.

Fontagné, L., Martin, P., and Orefice, G. (2017). The International Elasticity Puzzle Is Worse Than You Think. Working Papers 2017-03, CEPII research center.

Fontagné, L., Orefice, G., Piermartini, R., and Rocha, N. (2015). Product standards and margins of trade: Firm-level evidence. *Journal of International Economics*, 97(1):29–44.

Kramarz, F., Martin, J., and Mejean, I. (2016). Volatility in the Small and in the Large: The Lack of Diversification in International Trade. CEPR Discussion Papers 11534, C.E.P.R. Discussion Papers.

Lenoir, C., Martin, J., and Mejean, I. (2018). Search Frictions in International Good Markets. Technical report.

Martin, J. and Mejean, I. (2014). Low-Wage Countries' Competition, Reallocation Across Firms and the Quality Content of Exports. *Journal of International Economics*, (1):140–152.

Mayer, T., Melitz, M. J., and Ottaviano, G. I. P. (2014). Market Size, Competition, and the Product Mix of Exporters. *American Economic Review*, 104(2):495–536.

Mejean, I. and Schwellnus, C. (2009). Price convergence in the European Union: Within firms or composition of firms? *Journal of International Economics*, 78(1):1–10.

Pierce, J. R. and Schott, P. K. (2012). Concording U.S. Harmonized System Categories Over Time. *Journal of Official Statistics*, 28(1):53–68.

A Details on the harmonization algorithm

The 8-digit "Combined Nomenclature" is used to classify products in EU international trade. It is based on the international "Harmonized System" with the first six digits corresponding to hs6 products. The so-called cn8 product code is for example used to determine which rate of customs duty applies. As such, it is a requested variable in the DEB and DAU databases.[23] The use of these product codes in longitudinal studies however requires harmonizing the product classification system, over time. Without such harmonization, the same firm can end up selling the exact same product to the same importer in two different years and declare these products under different product codes.

In this Appendix, we explain how to adapt existing algorithms developed to concord nomenclatures over time to the particular case of the EU Combined Nomenclature. The chosen algorithm is based on Behrens and Martin (2015) 'connected components concordance', or C^3 for short, which is itself an improvement over the method proposed by Pierce and Schott (2012) for the US product nomenclature. C^3 uses the graph theory to identify stable and comparable groups of products over time while minimizing the size of each group. The identified groups of products are then assigned to a single, time-consistent, code. Applying the algorithm to the cn8 product categories allows maintaining a high degree of granularity in the definition of products while restoring the time-consistency of product categories. After having explained how to implement the algorithm, we describe the quantitative impact of such implementation. Interested readers can refer to Behrens and Martin (2015) for technical details regarding the C^3 algorithm.

A.1 Implementation of the algorithm

The main input to the C^3 algorithm is a full set of raw data files describing each year-to-year change to the Combined Nomenclature. While the vast majority of changes is concentrated over years when the Harmonized System is updated in 1996, 2002, 2007, 2012 and 2017, a fair number of product code adjustments still take place every year. As a consequence, it is not sufficient to produce concordance tables between various revisions of the Harmonized System. To take into account all of these yearly adjustments, we uploaded the nomenclature and information on year-to-year changes in cn8 product codes from the European Commission's website.[24]

The Stata do-file adapted from Behrens and Martin (2015) and named corres_nc8.do uploads these data for any pre-defined period of analysis, appends them into a long matrix, which is exported to Matlab using a shell.[25] Matlab's "Graph and Network Algorithms" codes are then used to identify the stable groups of product codes, over time. The output is stored into a Stata file called "corres_nc8$firstyear$lastyear.dta" where "$firstyear" and "$lastyear" respectively refer to the global variables defined in the introduction of the code, for the first and last years of the concordance.[26] Adjusting the first and last years of the concordance to the exact period of analysis is important as the average size of harmonized product codes naturally grows when the timespan is further extended. This is illustrated in Section A.2. The Stata code is written in a way that is flexible enough so that the researcher can easily produce a concordance table that exactly fits the period of analysis.

The structure of the output table is fairly simple. It contains three variables allowing to map the year-specific cn8 codes identified by a "cn8" and a "year" variables to the newly created harmonized product category called "id_conc" and stored in the third column. Starting from the raw customs data, it is then straightforward to apply the correspondence table and convert the cn8 trade flows into harmonized categories. The main correspondence table is complemented with a set of additional Stata datasets called "corres_idconc_nc8YEAR" for YEAR between $firstyear and $lastyear which are meant to allow the researcher "reverse-engineer" the concordance. Namely, for each year, the corresponding "corres_idconc_nc8YEAR" table allows going back from the "id_conc" to the cn8 codes. These tables are of particular use when the researcher wants to map the harmonized customs data with external information on products, which is defined in a particular version of the combined nomenclature / harmonized system.

Table A1 provides an example of the outcome of corres_nc8.do. In this simple example, the 1995

[23] As indicated in Table 5, the product code is not compulsory for firms filling the DEB form under the fourth level of the customs procedure.

[24] See http://ec.europa.eu/eurostat/ramon/nomenclatures. From this website, we recovered the year-to-year cn8 change, in .txt format, as well as the lists of all product codes active in a given year, in .csv.

[25] The algorithm was initially run on Stata 14 and Matlab 2015, this is the recommended configuration.

[26] By convention, the first set of product code adjustments is between $firstyear and $firstyear+1. Likewise, the last set is between $lastyear-1 and $lastyear.

Table A1: Example of a concordance produced by the C^3 method (based on corres_nc819951998.dta)

cn8 codes	Year	id_conc
44 1810 00	1995	126
44 1810 10	1996	126
44 1810 50	1996	126
44 1810 90	1996	126
44 1810 10	1997	126
44 1810 50	1997	126
44 1810 90	1997	126
44 1810 10	1998	126
44 1810 50	1998	126
44 1810 90	1998	126

cn8 product number 44 1810 00 was split into three products after the 1996 revision of the Harmonized System. When working with data starting in 1995, the researcher has no choice but to collapse the post-1996 information available for the three newly defined product codes into a single category, which is comparable in scope with the 1995 code. This is exactly what C^3 is doing, within the newly defined category number 126.[27]

In this particular case, the concordance is pretty straightforward. However, the succession of nomenclature revisions means that the algorithm has to deal with more complex graph structures in which year-specific categories can be merged together, before being split into two different codes that will eventually be further merged with other product categories afterwards. In such situation, the solution consists in identifying the isolated component of the graph that links all these year-specific products together through various nomenclature updates, before matching the products to a newly defined broader category. While the majority of product code adjustments concerns situations similar to the example in Table A1, it is important to treat more complex situations as well, especially when the period of analysis starts running over several major HS revisions.

Having described the implementation of the algorithm, we now provide summary statistics over the quantitative impact of the concordance. These statistics are based on various concordance tables computed using C^3 and several sub-periods between 1995 and 2018.

A.2 Impact of the harmonization

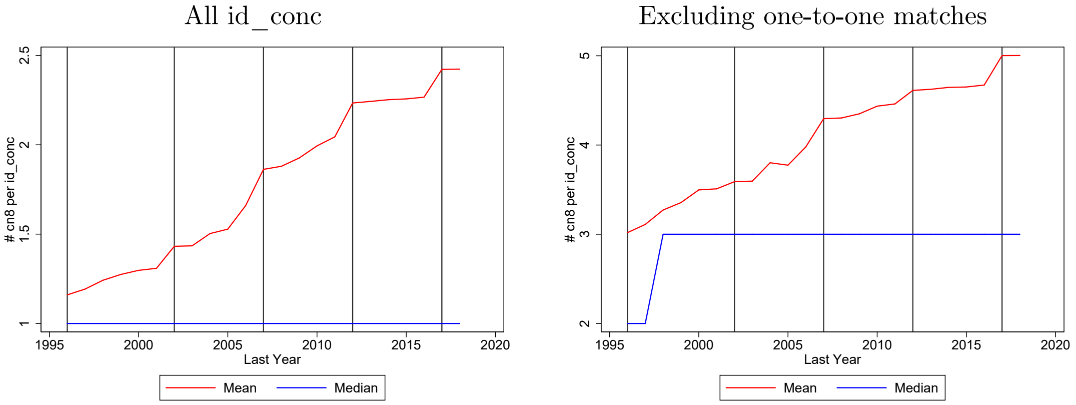
A.2.1 Impact on the average product

The outcome of corres_nc8.do is a mapping between year-specific product categories and harmonized "id_conc" groups. Obviously, all nc8 products are not affected by nomenclature changes. As a consequence, there is a substantial number of "id_conc" categories that map one for one with a single nc8 code, over the whole period of observation. An example of such a product would be coil compression springs (nc8 7320 20 81) which code never changes between 1995 and 2018. For such products, implementing the C^3 algorithm is innocuous. At the other side of the spectrum, the successions of revisions affecting the same product categories over and over can induce the creation of harmonized product categories composed of many different year-specific product codes. Since implementing the harmonization algorithm to actual trade data will eventually imply collapsing the information over these many product categories into a single observation per year, it is important to quantify the size of these large groups.[28] For example, in the largest group created over the 1995 - 2018 time period, a code identifying borides, a chemical compound, in the 1997 cn8 nomenclature is grouped with a code identifying waste and scrap of tinned iron or steel in 2017. While these goods do not share a lot of characteristics, a succession of nomenclature revisions implies that they are indirectly "linked" through at least one product.

[27] It has to be noted that the exact product categories in "id_conc" do not have a particular meaning. However, the algorithm is written in a way that the smallest values of "id_conc" correspond to groups that contain the largest numbers of time-varying product categories.

[28] The researcher may want to minimize the size of these largest components which can be done by simply dropping the corresponding id_conc from the analysis. Another solution consists in adjusting the period of analysis to avoid major product nomenclature revisions, which naturally inflate the size of the largest components.

Figure A.1: Statistics on the mean and median number of cn8 codes per id_conc, over time



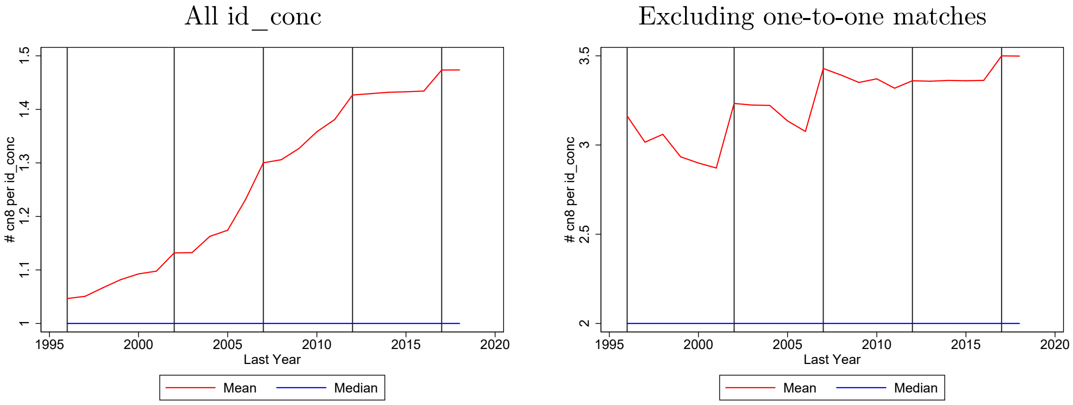
Note: Statistics based on the outcome of the C^3 algorithm applied to cn8 product changes between 1995 and 2018. The y-axis is the mean/median number of distinct nc8 codes composing a given id_conc, over time. The x-axis shows how statistics are affected by extending the time-period between 1996 and 2018. The left panel is based on all id_conc categories. The right panel is restricted to id_conc categories that encompass at least one product change, i.e. id_conc categories that are mapped to at least two cn8 codes over time. Vertical lines correspond to major revisions in the hs nomenclature.

Figure A.1 shows the average impact of the harmonization procedure, calculated across all newly created harmonized product categories. Consider first the left panel, which is based on all id_conc categories. Whatever the period on which the algorithm is implemented, the median number of cn8 products within an id_conc category is one, meaning that more than 50% of product categories have their scope unchanged between 1995 and 2018. While the median stays at one, the mean number of product categories grouped together is mechanically above one and grows when the harmonization procedure is extended to a longer period. Remarkably, even under the maximum time period between 1995 and 2018, the mean number of cn8 products grouped into a single id_conc category stays low, at 2.5. On average, the impact of harmonizing product codes over time is limited.

As illustrated in the right panel of Figure A.1, the limited impact of the harmonization is not entirely driven by those product categories that never change over the period of observation. Even when restricting the sample of harmonized categories to the sub-sample of products that are actually impacted by product changes, the mean number of nc8 codes per id_conc category is small, below 5. In more than half the cases, a single id_conc category corresponds to just three cn8 codes.[29]

[29] Having three distinct cn8 products into an id_conc group means that i) a single product code has been renamed twice over the period of observations, or ii) the product code has been split into two sub-categories, or iii) two original codes have been merged together at a point in time.

Figure A.2: Statistics on the mean and median number of cn8 codes per id_conc, in the 1995 cross-section


All id_conc

Excluding one-to-one matches

Note: Statistics based on the outcome of the C^3 algorithm applied to cn8 product changes between 1995 and 2018. The y-axis is the mean/median number of nc8 codes from 1995 that each id_conc category active that particular year encompasses. The x-axis shows how statistics are affected by extending the time-period between 1996 and 2018. The left panel is based on all id_conc categories. The right panel is restricted to id_conc categories that encompass at least one product change, i.e. id_conc categories that are mapped to at least two cn8 codes over time. Vertical lines have been added whenever a major change in nomenclature occurred.

While Figure A.1 is indicative on the number of nc8 codes that a given id_conc category encompasses over time, Figure A.2 conveys information on the potential loss of information, in the cross-section. Since harmonizing the data over time implies collapsing the cross-sectional information relative to several product codes which are linked together through a subsequent/past nomenclature change into a single category, the "cost" of the harmonization can be assessed through the size of the cross-sectional collapse. This is what Figure A.2 is meant to capture. Here as well, mean and median statistics reveal a rather limited impact of applying the concordance algorithm. For more than 50% of id_conc categories, there is no loss of information, in the cross-section, since the group encompasses a single cn8 code of the original (1995) nomenclature (see the Median line of the left panel in Figure A.2). As the right panel reveals, id_conc categories that actually correspond to a loss of information usually collapse the information over just two nc8 products (and 3.5 products on average). Again, this is consistent with the idea that product changes over time are not pervasive in the actual data.

Having shown that, on average, the harmonization procedure does not massively affect the classification of cn8 product codes, we now turn to the few groups of products that are the most strongly affected by the harmonization procedure. These are indeed the id_conc categories that the researcher might want to follow more closely.

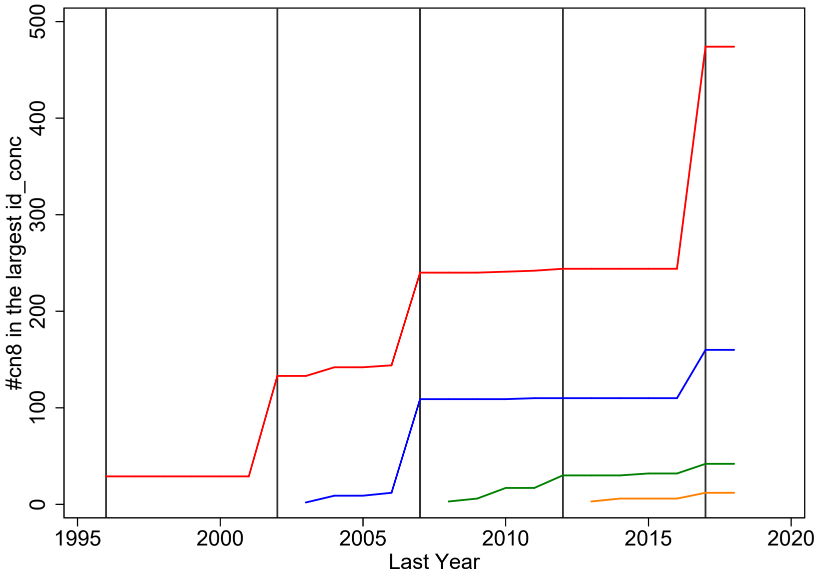
A.2.2 Impact on the most-varying product categories

While the previous section has established that the vast majority of cn8 codes are little affected by the harmonization procedure, the structure of cn8 changes and the chosen harmonization strategy implies that a few harmonized id_conc categories can end up grouping a very large number of cn8 codes, both over time and in the cross-section. This is illustrated in Figure A.3 which reports statistics on the size of the largest component. Here, the size of an id_conc category is defined as the number of cn8 codes entering it in a cross-section and is thus indicative of the extent of information which will eventually be collapsed once the algorithm is applied to the trade data.

As explained before, the very structure of the algorithm implies that the largest component, which corresponds to the newly created product category encompassing the most cn8 product categories, is growing steadily when more nomenclature updates are taken into account. This is the reason why extending the timespan of the analysis makes the problem of changes in product categories more serious. Moreover, the timing of nomenclature adjustments displays discontinuities since major revisions in the Harmonized System in 1996, 2002, 2007, 2012 and 2017 induce a substantially larger number of product changes over these particular years.

These properties are illustrated in Figure A.3. Here, we show the size of the largest component, as measured by the maximum number of cn8 codes that a given group encompasses, as a function of the starting and end years. For instance, the red line corresponds to the algorithm being implemented

Figure A.3: Size of the largest id_conc category in the cross-section



Note: The figure displays the number of cn8 codes in the largest id_conc category, as a function of the last year of the considered time span. Each colored line corresponds to a different starting period and the different points on the line various end years. By convention, the size of the id_conc category is measured as the number of cn8 codes which are mapped to this category, in the original cross-sectional nomenclature. Vertical lines correspond to major changes in the cn nomenclature.

from 1996 up to 2018, i.e. a maximum time period of 24 years. When the harmonization algorithm is restricted to changes occuring between 1995 and 1996, the largest id_conc category contains 29 original cn8 codes. This number hardly changes when the time period is further extended until 2001.[30] When one more year of data is added, and the 2002 hs revision is taken into account, the largest component more than triple in size, reaching 133 original cn8 products. This massive impact of the hs revision is also observed for revisions taking place in 2007 and 2017. Instead, the 2012 period does not massively affect the size of the largest component. All in all, working on the largest time period, between 1995 and 2018, implies that the largest component encompasses almost 500 different original cn8 codes. This is arguably sufficiently large a number for the problem to be seriously taken into account, either by dropping the largest components or by adjusting the sample period so as to reduce its size. As illustrated by the other lines in Figure A.3, working on the 2007-2018 period is for instance an option, since major HS revisions occuring during this period do not inflate the size of the largest component too much.

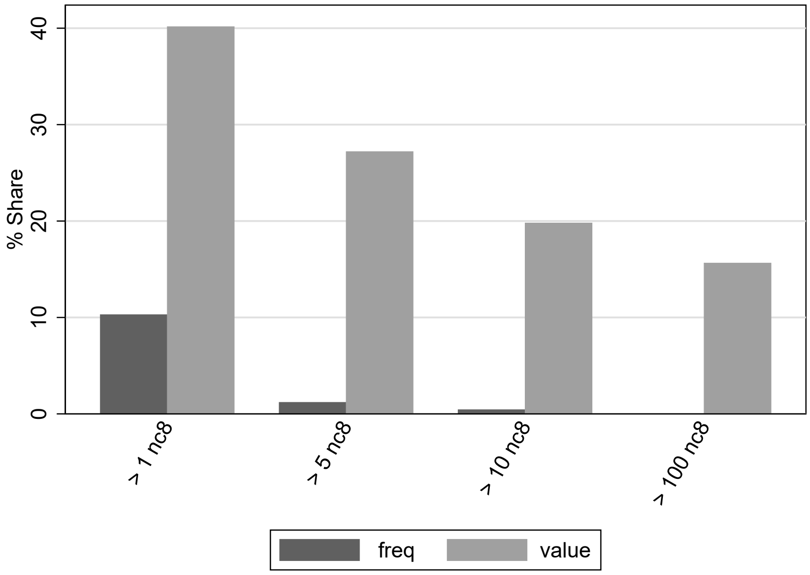
A.2.3 Discussion

Once the individual cn8 are mapped into broader id_conc categories, the question of efficiency and relevance of the results remains. As this appendix should make it clear, the use of the C^3 method generates a loss of information in the cross-section since cn8 codes are mapped into broader categories. This information loss has to be balanced with the benefit of the concordance, which is to restore the comparability of product categories, over time.

The cross-sectional impact of the harmonization can be assessed through the Figures in section A.2.1. However, these are based on the concordance tables obtained when implementing the algorithm, without any reference to the trade data to which the concordance will ultimately be applied. From this point of view, Figure A.4 is more informative. Namely, it shows the share of products and of the traded value which ends up mapped with id_conc categories that encompass more than one cn8 code. The share of such products is around 10%. However, it represents about 40% of the value of exports. Even more worrying is the share of overall trade collapsed in very large id_conc categories, almost 15% in the few groups made of more than 100 cn8 products. This clearly shows that the information loss can be substantial.

[30] Note that this is true even though some cn8 adjustments occur between 1996 and 2001. While they do not modify the size of the largest connected component, they do increase the size of smaller id_conc categories.

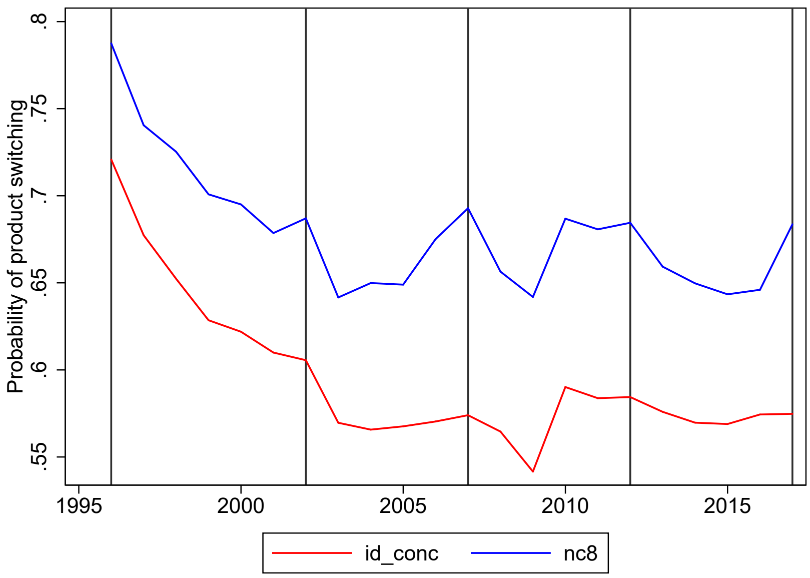
Figure A.4: Share of French exports in multi-product id_conc categories, 2017



Note: The figure shows the share of products (dark bars) and the value of exports (light grey bars) collapsed into id_conc categories of more than one cn8 product. Based on French export data of 2017.

While the use of id_conc instead of cn8 codes induces a loss of information at the product-level, this bias has to be compared with the benefit induced by the harmonization of product codes, over time. To illustrate this benefit, Figure A.5 shows the probability for a firm to start exporting a new product, when products are either based on cn8 codes or on id_conc categories. The measured probability is systematically found larger when using the raw cn8 product codes than when accounting for changes in the product nomenclature. Even a simple one-to-one change in the nomenclature would indeed induce the firm to declare a different product in years t and $t + 1$ which the researcher would inaccurately interpret as a product switching. Instead, the harmonized data would record both products into the same id_conc category. This corrects the measurement bias in switching probabilities. What Figure A.5 further illustrates is that the bias is not uniformally distributed over time. Instead, significant increases in the switching probability based on raw nc8 data are systematically observed at the time when a major HS revision occurs, most notably in 2007.

Figure A.5: Probability of exporting a new product, in raw and harmonized data

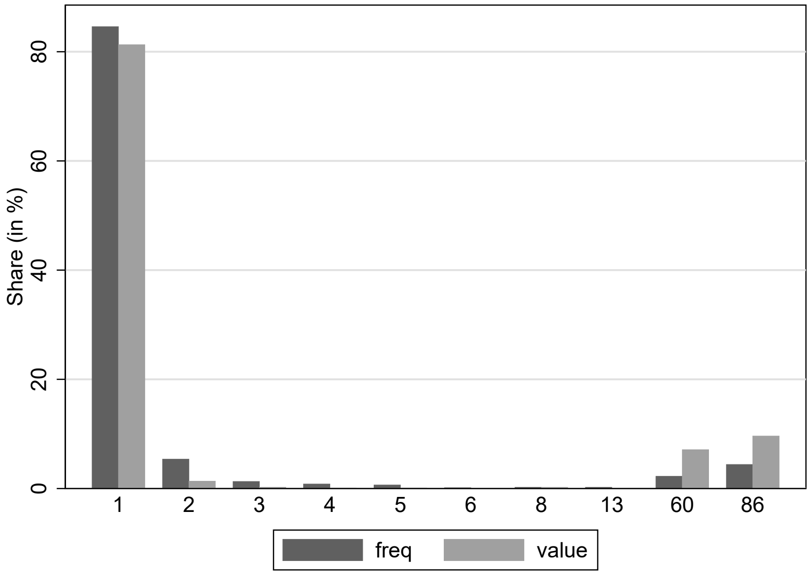


Note: This figure shows the evolution in the probability that a firm starts exporting a new product in terms of cn8 and id_conc categories, between 1995 and 2018. Vertical lines correspond to major changes in the HS nomenclature.

We close this section with a refinement of the previous discussion regarding the cross-sectional information loss induced by the product harmonization. Intuitively, the information loss might not be especially important when the products collapsed into id_conc categories are very similar. This is the case in Table A1, for instance. Here, the product change involves French windows and their frames (NC8 code 44 1810 00), which are further decomposed into sub-categories depending of the essence of the wood used to build the frame, starting in 1996. In most cases, the researcher is not especially interested by the exact essence of the wood and the information loss is thus innocuous.

To evaluate the extent to which this example is representative of the data at hands, we classified id_conc categories according to their scope as measured by the variety of HS4 categories they encompass. Figure A.6 shows the results. More than 80% of products and of the value of exports end up mapped with id_conc categories that do not cover more than one HS4 product. These categories are thus made of similar nc8 products and collapsing the information over each single product transaction into a broader value flow might not be especially costly. In a few cases, however, id_conc categories are made of nc8 products that belong to completely different HS4 families. A trade flow obtained from collapsing the information over these very different products might not be very insightful. In such cases, the researcher may want to discard the information.

Figure A.6: Distribution of exports by id_conc categories encompassing various numbers of HS4 codes



Note: The figure shows the distribution of exports (light grey bars) and of exported products (dark grey bars) mapped into id_conc categories encompassing various numbers of HS4 codes. The raw data cover French exports between 1995 and 2017.

B Additional Tables

Table B1: Details on transportation modes

Transportation Code	Transportation Type
1	Maritime Transport
2	Rail Transport
3	Road Transport
4	Air Transport
5	Postal item
7	Fixed transport installation
8	Inland waterway transport
9	Miscellaneous

Notes: For export flows, this corresponds to the main transportation mode after the good has exited France. For imports, it is the transportation mode up to the French frontier.

Table B2: Special departement codes

Codes	**Departement**
9A	Guadeloupe
9B	Martinique
9C	Guyane
9D	La Réunion
9E	Mayotte
99	Foreign - including Monaco and TOMs.
00	Under-threshold DEB/ Before 2012: DEB level of obligation 3 or 4.
98	Unknown departement

Notes: The “departement” variable indicates the region that the good leaves (export flows) or where it is delivered (import flows).

Table B3: Nature of the transaction

1	Transactions involving actual or planned transfer of ownership against compensation (financial or otherwise)
	Including:
11	Binding purchase/sell
12	Delivery for the purpose of a trial, for consignment or with the intermediary of a broker
13	Barter
14	Financial leasing
19	Other
2	Return of property after registration of the original operation under code 1 ; Free replacement of goods
	Including:
21	Return of goods
22	Replacement of returned goods
23	Replacement of non-returned goods
29	Other
3	Non-temporary transactions involving transfer of ownership without any compensation (financial or otherwise)
	Including:
30	
4	Operation before a tolling agreement
	Including:
41	Good sent to be redirected to the original country
42	Good sent to be redirected to another country
5	Operation after a tolling agreement
	Including:
51	Good redirected to the original State
52	Good redirected to another State
6	Transfer of product under inward and outward processing arrangements, with DAU waiver
	Including:
60	
7	Operations for joint defence projects or other joint intergovernmental production programs
	Including:
70	
8	Supply of material and equipment in the context of a general construction or civil engineering contract
	Including:
80	
9	Other
	Including:
91	Lease, loan and operational leasing for a period of more than 24 months
99	Other

Source: *Bulletin officiel des douanes*, 2016.

Table B4: Specific country codes

iso2 code	Country	Beginning	End
AA	Unknown	1900	
QP	High seas, outside territorial waters	2013	
QQ	Autocorrections	1993	
QR	Refuelling in intra-EU exchanges	2005	
QS	Refuelling in extra-EU exchanges	2005	
QT	Refuelling	1997	2000
QU	Non-identified country	1900	1999
QU	Unknown country or territory	2012	
QV	Unknown country or territory: intra-EU exchanges	1900	2011
QW	Unknown country or territory, extra-EU exchanges	1900	2011
QY	Indeterminate country	1900	2009
QZ	Indeterminate country	1900	2009
XA	American Samoa	1992	2000
XB	Saint Barthélemy	1900	
XC	Ceuta	1999	
XE	Saint-Eustatius	1900	2005
XF	Canary Islands	1986	1996
XG	Ceuta and Melilla	1900	1999
XI	Saint-Martin, meridional part	1900	2005
XK	Kosovo	2005	
XL	Melilla	1999	
XM	Montenegro	2005	2006
XN	Bonaire	1900	2005
XO	Australian Oceania	1900	2000
XP	West Bank - Gaza Strip	1995	2000
XQ	South Africa Republic - Namibia	1900	1989
XR	Polar regions	1900	2000
XS	Saba	1900	2005
XS	Serbia	2005	
XT	Saint-Martin, septentrional part	1900	2012
XV	Curaçao	1900	2005
XW	British Virgin Islands, Montserrat	1900	1997
XX	Undefined countries	1997	1998
XZ	New Zealand Oceania	1900	2000
ZZ	Various	1900	1998

Table B5: Changes among iso2 codes

iso2 code	Country	Beginning	End
CS	Serbia and Montenegro	2004	2005
SJ	Svalbard and Jan Mayen	1995	1996
TP	East Timor	2001	2002
XU	UEBL (Belgium-Luxembourg Economic Union)	1900	1998
YU	Serbia - Montenegro	1993	2003