Identifying Foreign Suppliers in U.S. Merchandise Import Transactions [*]

Fariha Kamal[†] Ryan Monarch[‡]
U.S. Census Bureau Federal Reserve Board

May 4, 2016

Abstract

Relationships between firms and their foreign suppliers are the foundation of international trade, but data limitations and reliability concerns make studying such relationships challenging. We evaluate and address these concerns using U.S. import data, and present new facts about U.S. buyer- foreign supplier relationships. The pattern of U.S. imports changes substantially by tracing trade back to the original supplier's location. Ranking cities by the number of U.S. buyer-foreign supplier relationships, nine of the top ten cities in 2011 are Chinese. Related-party relationships have more trade, while richer countries and more timely products tend to have more relationships.

[*] Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the U.S. Census Bureau, the Board of Governors of the Federal Reserve System, or of any other person associated with the Federal Reserve System. All results have been reviewed to ensure that no confidential information is disclosed. We thank Kyle Handley, C.J. Krizan, Javier Miranda, and Christian Volpe Martincus for valuable comments. We have benefitted immensely from conversations with David Dickerson and Glenn Barresse of the U.S. Census Bureau Economic Statistical Methods Division, Kristen Nespoli of the U.S. Census Bureau International Trade Management Division, and Diana Wyman from Statistics Canada. Clint Carter and William Wisniewski were extremely helpful with data requests and disclosure processes. All errors are ours.

[†] Contact: fariha.kamal@census.gov.

[‡] Corresponding Author. Contact: ryan.p.monarch@frb.gov.

1 Introduction

Every international trade transaction is an agreement between two firms, an importer (buyer) and an exporter (supplier), located in two different countries. For this reason, the recent availability of databases that provide the identity of both importers and exporters for individual transactions has fundamental appeal for the field of international trade. Indeed, the existence of such "two-sided" data has the potential to establish novel facts about traders that can augment the heterogeneous firm framework widely used throughout the literature (Melitz (2003)). To the best of our knowledge, two-sided trade transactions data has been analyzed for Colombia (Benguria (2014)), Chile and Colombia (Blum et al. (2013)), Costa Rica, Ecuador, and Uruguay (Carballo et al. (2013)), Norway (Bernard et al. (2014)), and the United States (Pierce and Schott (2012); Dragusanu (2014); Eaton et al. (2014); Kamal and Sundaram (2013); Monarch (2014); Heise (2016); Monarch and Schmidt-Eisenlohr (2016)).

That said, one of the primary concerns about such data is reliability: in order to have individual transactions that include both importing and exporting entities, one data source must identify individual traders in both countries. While it may be in the best interest of governments to collect reliable information about firms located in their jurisdiction for taxation purposes, it is not obvious that the same governments would have the incentive, or even the authority, to maintain accurate statistics on firms located outside its national borders. Subsequently, two-sided trade data will by definition be more susceptible to issues related to the identification of "foreign" buyers or suppliers. This paper describes an enhancement of data representing foreign suppliers to the U.S. to address potential concerns about the quality of relationship-level data, and uses this improved data to present new findings about relationships between U.S. buyers and their foreign suppliers.

We first describe the method for identifying foreign suppliers in U.S. import transactions. U.S. importing firms with shipments above $2,000 are required to complete U.S. Customs and Border Protection (CBP) Form 7501, part of which entails forming a code- known as the *Manufacturer ID* or *MID*- for the foreign supplier of the transaction. We explore some of the potential errors that may arise in completing this code, and demonstrate that the MID is widely used by both the U.S. and Canadian governments for official purposes. Additionally, we show using external data that following the rules of MID creation tends to generate unique identification of suppliers.

After this investigation, we describe our efforts to update the database of U.S. merchandise import transactions.[1] We propose a major refinement of the MID, using the bigram method to collapse very similar MIDs into one. In addition, we perform some common-sense cleaning methods for correcting potential errors that may arise as importers construct it, including attempting to identify and eliminate potential intermediaries from the supplier database. We demonstrate that our "foreign supplier" identifier improves the reliability of subsequent statistics on relationships.

In the last part of the paper, armed with our refined data, we present five empirical discoveries that come from examining relationships between U.S. importers and their suppliers. First, there are sizable discrepancies between the "exporting country" recorded on a customs form and a supplier's location, and we show that the pattern of U.S. imports would change significantly if exports were traced to the original location of production. Second, by using the MID to generate a sub-national database of exports to the U.S., we show that nine of the top ten cities (by number of suppliers) exporting to the U.S. are from China or Hong Kong. The city of Calgary in Canada has the highest exports to the U.S. by a wide margin. Third, we explore buyer and seller margins: mirroring results from other studies, U.S. importers tend to buy from many suppliers, while foreign suppliers tend to sell to few U.S. buyers. Fourth, although related party trade is about 40% of total imports in 2011, we show related-party relationships occupy a much smaller share of overall relationships. Finally, we examine the characteristics that lead to more relationships: richer countries tend to have more relationships with U.S. importers, while all else equal, countries exporting more to the U.S. have fewer relationships. We also find that time-sensitive products generally have more relationships.

The paper proceeds as follows. Section 2 describes the MID in greater detail, as well as the reasons it is included on customs forms. Section 3 presents our cleaning methodology, and examines those aspects of the MID that are potentially worrisome for its reliability. Section 4 uses the updated data to formulate our core set of stylized facts. Section 5 concludes.

[1] The Linked Longitudinal Foreign Trade Transaction Database (LFTTD) is maintained by the U.S. Census Bureau. See `http://www.census.gov/ces/dataproducts/datasets/imp.html` for further description.

2 Background and History

2.1 MID Creation

U.S. importers are required to fill out CBP Form 7501 in order to complete importation of goods into the United States (see Figure 1). Importing firms must record information about the value, quantity, and 10-digit HTSUS product category of the imported merchandise, as well as, in Box 13, the "Manufacturer ID" (MID) for each product. This field will contain information about the identity of the plant that produced the exported good. In general, CBP requires that the Manufacturer ID constitute the supplier, not trading companies or other trading agents:[2]

"For the purposes of this code, the manufacturer should be construed to refer to the invoicing party or parties (manufacturers or other direct suppliers). The name and address of the invoicing party, whose invoice accompanies the CBP entry, should be used to construct the MID." (U.S. Department of Homeland Security (2012)).

Customs Directive No. 3550-055 lays out the current method for deriving the MID metric for manufacturers and shippers.[3] The MID consists of an alphanumeric code that is constructed according to a pre-specified algorithm, using information on the seller's name and address from the importer's official invoice. The derivation (known as "keylining") is as follows: the first two characters of the MID must contain the two-digit ISO country code of the supplier, the next three characters the start of the first word of the exporter's name, the next three characters the start of the second word, the next four characters the beginning of the largest number of the street address of the foreign exporter, and the last three characters the start of the foreign exporter's city (see Table 1 for stylized examples).[4] The MID has a maximum length of fifteen characters.

The multi-step process for constructing the MID described above may cause concerns about its reliability as a usable identifier, or the susceptibility of the MID to erroneous data entry. We first note that 96% of all entries filed with CBP are filed electronically through the CBP's

[2] Due to strict rules-of-origin requirements, the MID for textile shipments represents "the entity performing the origin-conferring operations", based on Title 19 Code of Federal Regulations (CFR). See `http://www.gpo.gov/fdsys/pkg/CFR-2011-title19-vol1/pdf/CFR-2011-title19-vol1-sec102-23.pdf`. Textile products include both textile or apparel products as defined under Section 102.21, Title 19, CFR.

[3] See `http://www.cbp.gov/document/directives/3550-055-instructions-deriving-manufacturershipper-identification-code`.

[4] See page 7 at `http://forms.cbp.gov/pdf/7501_instructions.pdf` for a description of the MID and Appendix 2 for more detailed instructions on constructing MIDs.

Automated Broker Interface, which already reduces somewhat the probability of misspellings, illegibility or incorrectly filed MIDs. Second, it is very common to either employ in-house licensed customs brokers to facilitate the import process or use outside customs brokerage service providers to handle the shipment clearance process. In fact, Customs Broker License Examinations administered by CBP (passage of which is required if transacting customs business on behalf of others) typically include a question about MID construction.[5] Thirdly, customs brokers utilize specialized software that includes validation checks on entry data to prepare and transmit invoices electronically to CBP, such as SmartBorder.[6] In particular, SmartBorder software can store customer information that auto-populates, thereby further reducing errors due to manual data entry. Together these details should allay concerns about the potential for misspellings leading to errors in the construction of the MID.

2.2 Official Uses of the MID

Why does the MID exist? We have found that the MID field was included on U.S. CBP forms pursuant to the program of exchanging trade data for statistical purposes between the U.S. and Canadian governments: Canada uses the MID to augment its domestic data on establishment activity with export information. The Government of Canada does not independently measure exports to the U.S.- instead, they rely on U.S. import data officially transferred to them by the U.S. Census Bureau. Based on discussions with employees at the U.S. Census Bureau and Statistics Canada (the statistical bureau of Canada), we believe that such an exchange was the main impetus for the generation of the MID- Statistics Canada links (via Canadian supplier MIDs) export information from U.S. import data to Canadian establishment-level data. Filling out an MID was then made a requirement for imports from any country.

What does the U.S. government use the MID for, and why would it have the incentive to ensure U.S. firms are writing down the identity of their foreign partners correctly? According to U.S. law, there are two apparent reasons. First, the MID is utilized in national security programs such as the Customs-Trade Partnership Against Terrorism (C-TPAT). An active MID is required to be qualified for the program. Companies that join C-TPAT "sign an agreement to work with CBP to protect the supply chain, identify security gaps, and implement specific security

[5] See `http://www.cbp.gov/trade/broker/exam/announcement` for details about the exam. `http://www.cbp.gov/document/publications/past-customs-broker-license-examinations-answer-keys` includes sample exam questions and answer keys. Questions 5 and 12 on the April 2014 examinations ask about MID construction.

[6] See `http://www.smartborder.com/newsb2/ProductsSmartBorderABI.aspx`.

measures and best practices.[7] C-TPAT members are less likely to be subject to examinations at the port since they are considered "low-risk". The CBP reports that the program covers about 10,000 companies, accounting for over 50 percent of U.S. import value.

Second, the United States is clearly interested in enforcing trade-related regulatory requirements that relate to the identity of foreign suppliers to the U.S. For instance, anti-dumping measures are foreign-firm specific in nature. Furthermore, it is clear from U.S. regulations that the Manufacturer ID is used to track compliance with U.S. restrictions for textile shipments. MID criteria for textiles are more stringent than those for other products, since non-textile products typically do not have the rule-of-origin restrictions that exist for textile and apparel products. If an entry filed for such merchandise fails to include the MID properly constructed from the name and address of the manufacturer, the port director may reject the entry or take other appropriate action. The above discussion highlights the regulatory imperatives to provide an accurate MID and thereby lends credence to the idea that U.S. importers have incentives to accurately identify the foreign manufacturers from which they are importing.

2.3 External Validity

Even if U.S. importers are completing the MID correctly, there is still the concern that the amount of information collected is too limited to separately identify different suppliers. To check this, we use firm names and address information from external data to make "Pseudo-MIDs" and determine how uniquely they identify exporters.[8] We do this using exporter names and addresses from Chinese firm level production data, following the algorithms set forth by CBP and described above. We can then evaluate the uniqueness of the constructed MIDs using the source country data, allowing us to quantify how commonplace the problem of two firms having the identifier is. Second, we assess how often different cities have the same city code from their Pseudo-MID. Monarch (2014) undertakes this exercise with Chinese firm-level data collected by the Chinese National Bureau of Statistics (NBS), creating MIDs for exporting firms within particular Chinese Industrial Classification Codes using the firm name, city and address, with Chinese characters romanized according to the Hanyu Pinyin system.[9]

[7] `http://www.cbp.gov/border-security/ports-entry/cargo-security/c-tpat-customs-trade-partnership-against-terrorism`

[8] Note this exercise is just to check how well the MID coding procedure can identify firms. The outside data need not match U.S. import data.

[9] The exercise is not perfect, as the observations of Chinese production data are at the firm level, while the MID is meant to capture actual production locations, or plants. Since there is no matching or comparing between datasets, the exercise should be construed simply as a general test of MID rules.

Table 2 reproduces the tables in Monarch (2014). Panel A, column 2 shows the number of Chinese exporters within each of five industries calculated using NBS firm level data. Column 3 shows the number of Pseudo-MIDs that could be created using the name and address information in the same dataset. The final column lists the percentage share of Pseudo-MIDs in the total number of exporters. The very high percentages (ranging from 97 to 100 percent) indicate that the algorithm used to generate MIDs often produces unique identifiers within an industry for an exporter. Panel B shows results from an identical exercise using city information. Column 2 shows the number of cities with at least one exporter within each industry using NBS firm level data. Column 3 shows the unique number of cities generated using the last three digits of the Pseudo-MIDs. Again, the higher percentages in the final column indicate that the three digit codes in the MID representing the city of the exporter tends to match the actual number of cities quite well. Taken together, the results in this table are another demonstration that U.S. importers constructing MIDs according to the rules are likely to generate reasonably unique identifiers of foreign exporting firms, especially within industry categories.

3 Cleaning Methodology and Summary Statistics

For the reasons laid out above, we believe that even in its raw form, the MID is likely to provide a useful foundation for identifying foreign suppliers to the U.S. Nonetheless, we undertake both probabilistic matching methods and basic checks in order to make the data as reliable as possible. In this section, we describe our methodology for cleaning the MID and offer some summary measures of the resulting *relationship* level data, where the term relationship will refer to the two-way combination of a U.S. firm and its foreign supplier, unless otherwise noted. All of our analysis will take place using only the most recent year of available data, 2011.

Before beginning any cleaning procedures, we first note that MIDs are missing in 1.9% of the 59 million import transactions in 2011, a sizable number. Why might an MID be missing? If U.S. companies import through a *foreign-trade zone*- a designated location in the United States where companies are allowed to delay or reduce duty payments on foreign merchandise and have access to streamlined customs procedures- they are not required to fill out an MID.[10] Foreign-trade zone status must be noted on Form 7501 (Box 2), meaning we can see how often

[10] There are about 250 such zones in the United States. See also `http://enforcement.trade.gov/ftzpage/index.html`.

this is the explanation: we find that 98.8% of missing MIDs are associated with foreign-trade zone transactions.

3.1 Bigram Matching

As our baseline, we use a character matching protocol known as bigram matching to combine very similar MIDs. A bigram is an approximate string comparator computed from the ratio of the number of common two consecutive letters of the two strings and their average length minus one. We use the STATA-based bigram matching algorithm developed by Wasi and Flaaen (Forthcoming)- such that all possible MID pairs are assigned a field-similarity score- to set a standard for determining if any Manufacturer ID is "similar enough" to another Manufacturer ID.[11] Appendix A provides examples of pairs and their associated field-similarity score.

How similar should two MIDs be in order to consider them the same supplier? For the 15 character Manufacturer ID, we identify a few rules of thumb for field-similarity (where 1 means a 100% match): a score of 0.98 or higher tends to match MIDs with 1-2 characters being different, while scores of 0.97 or higher tend to match to those MIDs that are identical in all aspects, other than one has a numeric address field and the other has none. For our results, we decide to adopt a field-similarity score of 0.98, meaning that we are likely to pick up simple typographical errors such as missing one character or only using the first name of a company, but we will count similar MIDs with different addresses as different suppliers. We believe this standard is sufficiently conservative, so as to allow for the possibility of simple coding errors, while still being stringent enough to not lump together two different suppliers.

The implementation procedure is as follows: for each origin country (the location of the supplier) in 2011, we match each MID to every other MID, producing a field similarity score. If the field similarity score for a match is 0.98 or above, then we will consider those MIDs to be the same. If multiple MIDs are found to be similar to the same MID, then all of those MIDs will be considered to be the same supplier.[12] We then provide a "best MID" variant for each MID in the underlying data, which enables us to generate relationships and other supplier-specific

[11] Other papers that use bigram matching include Anderson et al. (2015), Ernstberger and Grüning (2013), Flaaen (2014), Green and Jame (2013),Chodorow-Reich (2014), and Braun and Raddatz (2010).

[12] For example, if supplier A and supplier B are both similar to supplier C, then we consider supplier A, B, and C to be the same supplier, even if A and B are not found to be similar to each other (a situation that is exceedingly rare). In this work, we are agnostic about which variant of the MID (in this example, A,B, or C) should be retained, choosing randomly.

variables (such as size) at the best MID level.[13]

3.2 Additional Cleaning

There are a few other common sense adjustments to the MID that we make. We drop any MID that does not conform to the algorithm outlined in the CBP Form 7501 Instructions, including MIDs that are a series of numbers, MIDs that do not have three letters for the city code (one common mistake is for suppliers from New Territories, Hong Kong to have their city code written NT, resulting in a misspecified city code), and the like. We also drop any MID that has a country code corresponding to no known ISO2 code. All told, these changes together with the above methodology end up reducing the total number of MIDs in 2011 from 1,287,630 to 911,765, a drop of 29%.

There is a straightforward way to show that these changes affect the reliability of the relationship data, using an additional field on CBP Form 7501 that U.S. importers have to complete for each transaction. U.S. firms are required to write down (in Column 32C) whether the transaction was between "related parties" according to Section 105.102(g), Title 19 CFR, meaning one party has a 5% controlling interest in the other, or the parties have an employer/employee relationship, share offices or directors, or are family members or partners. In theory (excluding within-year ownership changes), a U.S. firm and its supplier should either have all of their transactions classified as related, or none. From the raw data to the cleaned data, the share of relationships that mix related and non-related transactions falls somewhat, from 5.8% to 5.5%.

3.3 Summary Statistics

We next illustrate some of the properties of our updated MID. After implementing our cleaning procedures, the minimum length of any MID in the data is 11 characters, and the maximum is 15 characters. The longer the MID is, the more likely it is to distinguish between suppliers. Table 3 shows that MIDs are split about evenly between 11,12,13,14 and 15 characters. 19% of these codes are the maximum length allowable- 15 characters. Table 4a shows how often the address component of the MID is populated: the vast majority of MIDs (89%) do have at least some address information included.

[13] The related party status of a cleaned MID relationship with both related and non-related party transactions will be random.

A worrisome issue concerning the address component of the MID is the presence of non-numeric address conventions in Latin America- according to a 2007 Los Angeles Times article, "most Costa Rican address are expressed in relation to the closest community landmark".[14] Theoretically, this could result in fewer fully-populated address codes.[15] Table 4b shows that South American and "Mexico and Central America" do not actually have a major lack of numeric address components compared to other regions- Europe, Asia, and Africa all have larger fractions with no address information. However, Costa Rica (Table 4c) is an exception, with about 18% of Costa Rican MIDs empty of address information. Table 4b also shows that North American MIDs (predominantly Canada) have full address information for almost half of all MIDs, not surprising given that Statistics Canada continues to successfully match MIDs to their domestic establishment database.

An additional worry is that the direct supplier of the good is not being used to generate the MID, with the U.S. importing firm instead simply writing down an MID corresponding to its intermediary or trading firm. Even though CBP expressly warns against doing so, we know that intermediaries play an integral role in facilitating international trade, so there is certainly some possibility of it occurring. One way to assess this is to examine the number of product or industry categories an MID-identified supplier is shipping. Intermediaries are more likely to export products spanning different industries (Ahn et al. (2011)), while manufacturers are more likely to possess a core competency- there may be few benefits from producing apples, socks, and vacuum cleaners at the same facility. Table 5 shows that 96% of MIDs export 5 or fewer HS2 codes, and 97% of MIDs export 10 or fewer HS10 codes. We decide to drop those with more than 10 HS2 codes from our data, resulting in a further reduction in the total number of MIDs by 1%.

4 Findings from Relationship-Level Trade Data

After undertaking this large-scale upgrade of the U.S. merchandise import data in order to make the foreign supplier identifier as reliable as possible, we are left with a total of 1,579,983 buyer-supplier relationships. Below we present five sets of stylized facts, relying on our cleaned

[14] "With Costa Rica's mail, it's address unknown", by Marla Dickerson. November 5, 2007 `http://articles.latimes.com/2007/nov/05/business/fi-crmail5`.

[15] Some of the examples from the article- such as "125 meters west of the Pizza Hut" or "200 meters south of the cemetery, cross the train tracks, white two-story house"- do have numeric characters, though it is impossible to tell if suppliers actually include this information on their invoice.

MID variable.

4.1 Exporting Country can differ from Producer Country

Returning again to the Form 7501 shown in Figure 1, note that in addition to the Manufacturer ID (Box 13), importers also have to complete a field for the exporting country of a product (Box 14). We find that in 17% of relationships (accounting for 29% of total U.S. imports), the exporting country does not actually match the supplier's country of origin as denoted by the first two characters of the MID.

Why might the exporting country differ from the country of the MID? CBP Instructions read "The country of exportation is the country of which the merchandise was last part of the commerce and from which the merchandise was shipped to the U.S. without contingency of diversion." (U.S. Department of Homeland Security (2012)). In practice, based on discussions with U.S. Census Bureau staff, what a discrepancy between these two likely means is that the "exporting country" is re-exporting the goods. In other words, if already produced goods were not substantially transformed, but instead were repackaged or re-sold from a second country, then the second country would be listed as the official exporting country.

Given that aggregate trade statistics for the U.S. are calculated using the exporting country, rather than the "country of origin" derived from the MID, one can see how different U.S. trade patterns may look if goods were traced all the way back to their actual production location. Table 6 presents the top 10 exporters to the U.S. by both of these measures in 2011. Interestingly, though China is the top source by either measure, its share of total U.S. imports drops when measured by the country of origin. This fits with the general intuition laid out above, as China is generally thought of as a major re-exporter with a nontrivial share of its exported products having little domestic value added. It is also apparent that more exports to the U.S. originate in Mexico than indicated by aggregate data, while the reverse is true for Canada.

4.2 Building a Sub-National Export Database

The city code embedded in the MID presents an opportunity to explore the geography of exports to the U.S. at a sub-national level. As discussed above, the MID is meant to capture the establishment-level supplier of the traded good, meaning the city code should correspond to the actual location of production. However, a limit of three characters in the city code can in many cases make definitive identification of the exporting city difficult. As just one example, for suppliers from China with city code "SHA", the city could plausibly refer to either to Shanghai (a city of 24 million people) or Shantou (6 million).[16] The use of the city code thus depends greatly on the research question, and in particular, aggregating exports to the city level should only be done with some caution.[17]

Taking this caveat in mind, we present Table 7, which presents the top city codes in terms of export value and number of relationships to the U.S. One striking fact is the widespread prevalence of Chinese cities in the list: going by the actual number of relationships, nine of the top 10 exporting cities to the U.S. in 2011 were in China or Hong Kong. Note that even if one were to split "SHA" into two equally sized cities, both would still be counted among the top 10 exporting cities to the U.S. by value and by the number of relationships. Thus, creating export data at the sub-national level reveals to an even greater extent the massive export powerhouse that China has become.

A second interesting pattern is the importance of geographic proximity of cities and total export value to the U.S. The city of Calgary in Canada is by far the largest city of exports to the U.S by value- perhaps as a result of U.S. oil imports (of which Canada represents 10%) from the commodity-rich province of Alberta. Two Mexican cities also appear on the top cities by value to the U.S.

4.3 Buyer, Seller, and Product Margins

We can also compare results from our data to the work of Bernard et al. (2014) on two-sided heterogeneity and matching in international trade. We find that U.S. importers have an average of 12 suppliers, while suppliers to the U.S. have an average of 4 importers. The corresponding figures in Bernard et al. (2014) using Norwegian import data exhibit the same pattern of buyers matching to many sellers, but sellers matching to few buyers: Norwegian buyers have an average of 9 exporters, while exporters to Norway have an average of 2 buyers.[18] Diving deeper into the HS10 product level (and continuing to drop suppliers with over 10 HS10 products), we find that

[16] There are other possible cities in China "SHA" could refer to; the two listed are the two most populous. The same metric for determining likely cities is used below.

[17] Though importers do have to include the "Foreign Port of Lading" in Box 19 of Form 7501, we have not found this field to be widely populated in the LFTTD.

[18] Unlike the LFTTD, buyers in the Norwegian buyer-seller data may be entities other than firms, such as individuals.

the average relationship spans 7 products, while the average supplier exports 3 HS10 products. The average number of buyer-supplier-product relationships across source country-product bins is 8, while the median is 3.

4.4 Related Party Relationships

According to official Census Bureau data, trade within related parties typically accounts for about 40% of all U.S. annual imports. Since we can use the MID to actually identify related-party relationships in the data, we can contrast them to arm's-length relationships. In fact, related-party relationships occupy a very small share of total relationships, only 6.6%. In order for such a small share of total relationships to account for a much larger share of trade, it must be the case that these relationships have higher values. Indeed, a simple regression with product and source country fixed effects shows that related party relationships- at the buyer-supplier-product level- do trade more than non-related parties (Table 8 Column 1). We also find that related party relationships tend to have higher unit values (Table 8 Column 2). This effect is more precisely estimated than is typical for related parties, as we use trade and unit values at the *relationship-level*, rather than the firm level.

4.5 Relationships and Country/ Product Characteristics

Using our refined MID measure, Table 9 shows which countries have the most supplier relationships with the U.S. in 2011. Over a quarter of all trade relationships in 2011 were between U.S. buyers and mainland Chinese suppliers, a share that bumps up to one-third of all relationships by also including Hong Kong.

We next examine which country and product characteristics are associated with more relationships. To do so, we take the number of buyer-supplier-product relationships, and regress it (together with product-level fixed effects) on a number of different country characteristics. Table 10 shows that higher-income countries (measured by per-capita GDP) tend to have more relationships. Perhaps surprisingly, the number of relationships within a country is negatively correlated with log exports to the U.S. in 2011, meaning that all else equal, countries with higher exports to the U.S. (after accounting for product characteristics) actually have fewer relationships. This implies that trade between the U.S. and its major exporting partners is dominated by a relatively small number of relationships.

We carry out a similar exercise with a product-level measure of substitutability and the number of relationships within a country-product pair. Our hypothesis is that more substitutable products are more likely to have thicker markets and thus will have more suppliers and thus greater opportunities to form relationships. To check this, we follow work by Hummels and Schaur (2013) to calculate the timeliness of products by constructing the share of air imports over total import value in an HS10 category.[19] Their intuition is that consumers are more likely to switch from lengthy ocean shipping to quicker (and more expensive) air shipping when products are closer substitutes. Since products heavily reliant on air travel tend to be closer substitutes (for example, products in the Automotive and Foods and Beverages category have high air shares), we expect a positive correlation between this measure and the total number of relationships in a product category. Again taking the number of relationships as a dependent variable for a regression with source country fixed effects, Table 11 shows that more timely products tend to have more relationships.

5 Summary

This paper investigates the properties of the Manufacturer ID variable that identifies the foreign supplier in a U.S. merchandise import transaction, and uses it to generate a number of stylized facts about U.S. importer- foreign exporter relationships. We document the rules and laws that govern the generation of the MID, noting that the MID is primarily meant to capture the origin-conferring entity in a merchandise import transaction. Next, we present a set of cleaning algorithms and procedures meant to make the MID as usable as possible, improving the underlying reliability of the variable. This includes collapsing very similar MIDs into one, as well as common-sense checks for suspicious entries. Finally, we illustrate new findings about buyer-supplier relationships in international trade permitted by the availability of the MID. Adjusting aggregate U.S. exports to their actual supplier's country of origin has intuitive effects on the overall pattern of trade, and by examining the sub-national sources of U.S. imports, we show that Chinese cities are major export sources to the U.S. Richer countries have more relationships, as do more timely products.

In any national dataset attempting to measure information on foreign firms, there are bound

[19] We also tried to use estimates of product substitutability over the years 1990-2001 from Broda and Weinstein (2006), but changes in HS product codes between 2001 and 2011 mean that the usable sample of relationships shrinks dramatically from 3,600,000 to only 1,600,000.

to be questions about the underlying reliability. The results of our study indicate that when used appropriately, the Manufacturer ID can be an important part of deeper investigations of buyer and supplier relationships in international trade. Our findings offer the first set of systematic evidence in identifying potential issues with using the MID and methods to modify the MID in order to address pertinent concerns. One aspect we have not addressed in this paper is the dynamic nature of buyer-supplier relationships: combining similar MIDs into one is relatively straightforward in a single year, but becomes extremely challenging when trying to implement the procedure over time. We see this as the next step in continuing to refine and improve foreign supplier identification in U.S. merchandise import data.

References

Ahn, JaeBin, Amit K Khandelwal, and Shang-Jin Wei, "The role of intermediaries in facilitating trade," *Journal of International Economics*, 2011, *84* (1), 73–85.

Anderson, Michael A, Martin H Davies, José E Signoret, and Stephen LS Smith, "Firm Heterogeneity and Export Pricing in India," 2015.

Benguria, Felipe, "Production and Distribution in International Trade: Evidence from Matched Exporter-Importer Data," 2014. Mimeo.

Bernard, Andrew B., Andreas Moxnes, and Karen Helene Ulltveit-Moe, "Two-sided Heterogeneity and Trade," Working Paper 20136, National Bureau of Economic Research 2014.

Blum, Bernardo S., Sebastian Claro, and Ignatius J. Horstmann, "Occasional and Perennial Exporters," *Journal of International Economics*, 2013, *90* (1), 65–74.

Braun, Matías and Claudio Raddatz, "Banking on politics: when former high-ranking politicians become bank directors," *The World Bank Economic Review*, 2010, pp. 1–46.

Broda, Christian and David E Weinstein, "Globalization and the Gains From Variety," *The Quarterly Journal of Economics*, 2006, *121* (2), 541–585.

Carballo, Jerónimo, Gianmarco IP Ottaviano, and Christian Volpe Martincus, "The Buyer Margins of Firms' Exports," Discussion Paper 9584, CEPR 2013.

Chodorow-Reich, Gabriel, "The employment effects of credit market disruptions: Firm-level evidence from the 2008–9 financial crisis," *The Quarterly Journal of Economics*, 2014, *129* (1), 1–59.

Dragusanu, Raluca, "Firm-to-Firm Matching Along the Supply Chain," 2014. Harvard University, mimeo.

Eaton, Jonathan, Marcela Eslava, Cornell J Krizan, Maurice Kugler, and James Tybout, "A Search and Learning Model of Export Dynamics," 2014.

Ernstberger, Jürgen and Michael Grüning, "How do firm-and country-level governance mechanisms affect firms disclosure?," *Journal of Accounting and Public Policy*, 2013, *32* (3), 50–67.

Flaaen, Aaron, "Multinational Firms in Context," Working Paper, University of Michigan 2014.

Green, T. Clifton and Russell Jame, "Company name fluency, investor recognition, and firm value," *Journal of Financial Economics*, 2013, *109* (3), 813–834.

Heise, Sebastian, "Firm-to-Firm Relationships and Price Rigidity: Theory and Evidence," 2016.

Hummels, David and Georg Schaur, "Time as a Trade Barrier," *American Economic Review*, 2013, *103* (7), 2935–2959.

Kamal, Fariha and Asha Sundaram, "Buyer-Seller Relationships in International Trade: Do Your Neighbors Matter?," 2013. Mimeo.

Melitz, Marc, "The Impact of Trade on Intra-Industry Reallocations and Aggregate Industry Productivity," *Econometrica*, 2003, *71* (6), 1695–1725.

Monarch, Ryan, "It's Not You, It's Me: Breakups in U.S.-China Trade Relationships," Working Paper 14-08, U.S. Census Center for Economic Studies 2014.

_ **and Tim Schmidt-Eisenlohr**, "Learning and the Value of Trade Relationships," 2016.

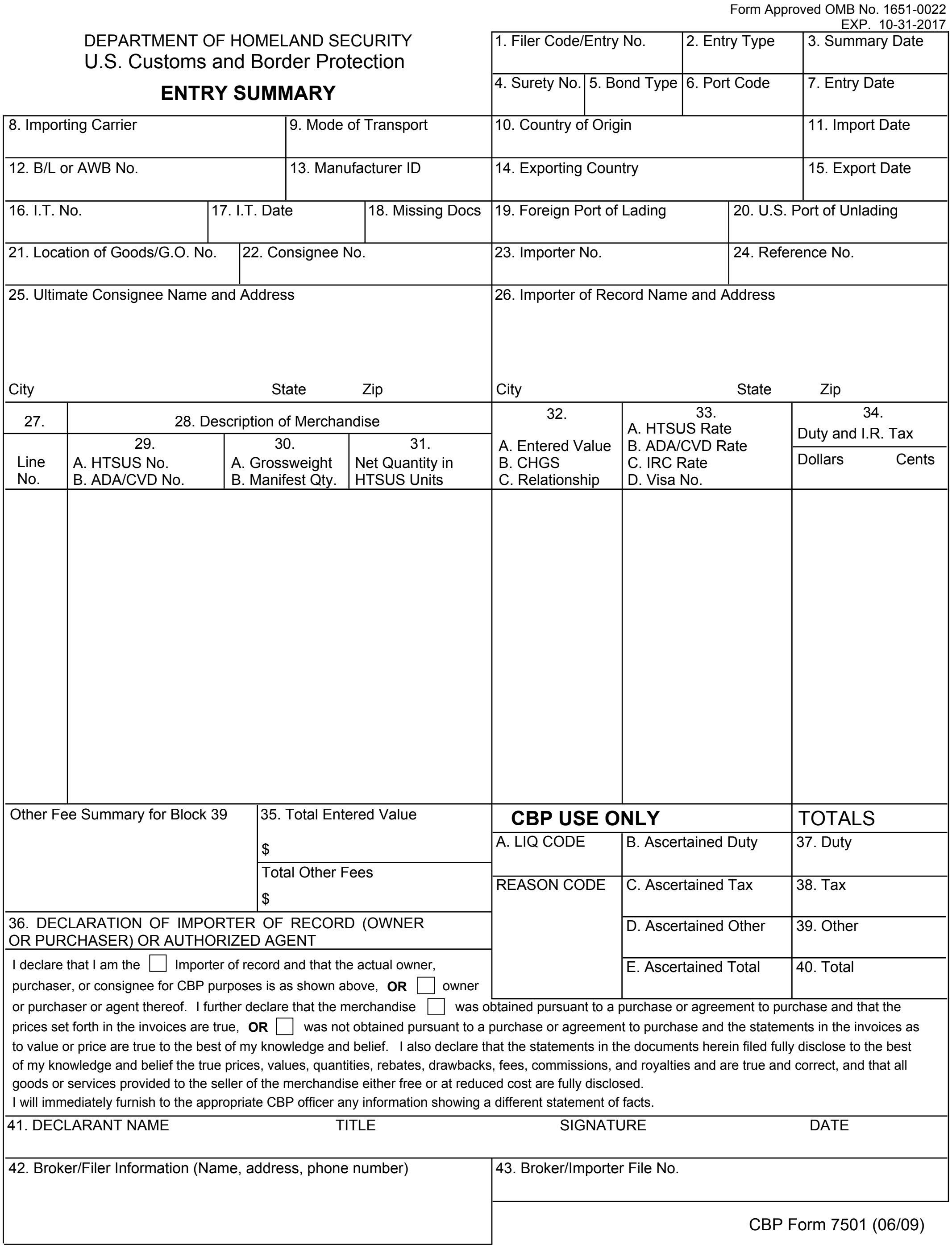
Pierce, Justin R. and Peter K. Schott, "The Surprisingly Swift Decline of U.S. Manufacturing Employment," Working Paper 18655, National Bureau of Economic Research 2012.

U.S. Department of Homeland Security, "CBP Form 7501 Instructions," 2012.

Wasi, Nada and Aaron Flaaen, "Record Linkage using STATA: Pre-processing, Linking and Reviewing Utilities," *The Stata Journal*, Forthcoming.

Figures and Tables

Figure 1: CBP Form 7501



Form Approved OMB No. 1651-0022
EXP. 10-31-2017

DEPARTMENT OF HOMELAND SECURITY
U.S. Customs and Border Protection

ENTRY SUMMARY

1. Filer Code/Entry No.	2. Entry Type	3. Summary Date		
4. Surety No.	5. Bond Type	6. Port Code	7. Entry Date	
8. Importing Carrier	9. Mode of Transport	10. Country of Origin	11. Import Date	
12. B/L or AWB No.	13. Manufacturer ID	14. Exporting Country	15. Export Date	
16. I.T. No.	17. I.T. Date	18. Missing Docs	19. Foreign Port of Lading	20. U.S. Port of Unlading
21. Location of Goods/G.O. No.	22. Consignee No.	23. Importer No.	24. Reference No.	
25. Ultimate Consignee Name and Address City State Zip		26. Importer of Record Name and Address City State Zip		

27.	28. Description of Merchandise			32.	33.	34.
Line No.	29. A. HTSUS No. B. ADA/CVD No.	30. A. Grossweight B. Manifest Qty.	31. Net Quantity in HTSUS Units	A. Entered Value B. CHGS C. Relationship	A. HTSUS Rate B. ADA/CVD Rate C. IRC Rate D. Visa No.	Duty and I.R. Tax Dollars Cents

Other Fee Summary for Block 39	35. Total Entered Value $	**CBP USE ONLY**		TOTALS
		A. LIQ CODE	B. Ascertained Duty	37. Duty
	Total Other Fees $	REASON CODE	C. Ascertained Tax	38. Tax
36. DECLARATION OF IMPORTER OF RECORD (OWNER OR PURCHASER) OR AUTHORIZED AGENT			D. Ascertained Other	39. Other
			E. Ascertained Total	40. Total

I declare that I am the □ Importer of record and that the actual owner, purchaser, or consignee for CBP purposes is as shown above, **OR** □ owner or purchaser or agent thereof. I further declare that the merchandise □ was obtained pursuant to a purchase or agreement to purchase and that the prices set forth in the invoices are true, **OR** □ was not obtained pursuant to a purchase or agreement to purchase and the statements in the invoices as to value or price are true to the best of my knowledge and belief. I also declare that the statements in the documents herein filed fully disclose to the best of my knowledge and belief the true prices, values, quantities, rebates, drawbacks, fees, commissions, and royalties and are true and correct, and that all goods or services provided to the seller of the merchandise either free or at reduced cost are fully disclosed.
I will immediately furnish to the appropriate CBP officer any information showing a different statement of facts.

41. DECLARANT NAME	TITLE	SIGNATURE	DATE
42. Broker/Filer Information (Name, address, phone number)		43. Broker/Importer File No.	

CBP Form 7501 (06/09)

Table 1: Stylized Examples of Manufacturer ID

Country	**Exporter Name**	**Address**	**City**	**MID**
Bangladesh	Red Fabrics	1234 Curry Road	Dhaka	BDREDFAB1234DHA
France	Green Chemicals	555 Baguette Lane, #1111	Paris	FRGRECHE1111PAR
Republic of Korea	Blue Umbrellas	88 Kimchi Street	Seoul	KRBLUUMB88SEO

Note: The above examples are based on fictitious names and addresses.

Table 2: Analysis of MIDs as Constructed from China Industrial Production Data

(a) Uniqueness of the "MID", 2005

Industry (CIC)	# of Exporters	# of "MID"s	%
CIC 3663	39	38	97.4
CIC 3689	27	26	97.3
CIC 3353	37	37	100
CIC 3331	35	35	100
CIC 4154	74	73	98.6

Note: This panel uses name, address, and city information from China NBS firm data to construct a "MID" for each firm, according to the rules laid out in U.S. CBP Form 7501. In constructing the name of the firm in English, the Hanyu Pinyin romanization of Chinese characters, with two to three characters per word of the English name, is used. The second column states the number of firms with positive export values in the given industry in 2005. The third column states the number of unique constructed "MID"s.

(b) Uniqueness of the City Code

Industry (CIC)	# of Cities	# of City Codes	%
CIC 3663	22	21	95.5
CIC 3689	15	14	93.3
CIC 3353	28	24	85.7
CIC 3331	15	13	86.7
CIC 4154	19	18	94.7

Note: This panel uses city information from China NBS firm data to construct city information as found in the MID, where only the first three letters of city are given. The second column states the true number of cities with at least one exporting firm in the data from 2005, while the third column states the number of unique city codes. Source: China National Bureau of Statistics, Monarch (2014).

Table 3: Distribution of MID Lengths

11	12	13	14	15
14%	18%	26%	23%	19%

Note: The maximum length of an MID is 15 characters. Our cleaned sample of MIDs has a minimum of 11 characters.

Table 4: MID Address Field

(a) All Countries

None	1	2	3	4
11%	15%	27%	24%	23%

(b) By Region

	None	1	2	3	4
North America (ex. Mexico)	1%	3%	13%	34%	49%
Central America and Mexico	10%	10%	21%	34%	24%
South America	9%	6%	16%	37%	33%
Europe	13%	22%	37%	14%	14%
Asia	12%	13%	24%	27%	24%
Oceania	6%	13%	33%	28%	20%
Africa	16%	14%	27%	20%	22%

(c) Costa Rica

None	1	2	3	4
18%	12%	16%	34%	19%

Note: MIDs can have anywhere from 0-4 numeric characters in the address field, taken from the largest number in the address on the supplier's invoice.

Table 5: Distribution of MIDs, by Number of Exported Products/Industries

(a) HS10 Products

1-5	6-10	11-20	21-50	More than 50
84%	13%	3%	0.6%	0.1%

(b) HS2 Industries

1-2	3-5	6-9	10-20	More than 20
84%	12%	3%	0.9%	0.1%

Note: This table shows the distribution of suppliers to the U.S. identified by the MID by the number of products or industries exported.

Table 6: Top 10 Export Countries to the U.S., 2011

(a) By "Exporting Country"

Country	Share
China	18%
Canada	14%
Mexico	12%
Japan	6%
Germany	5%
South Korea	3%
Great Britain	2%
Saudi Arabia	2%
Venezuela	2%
Taiwan	2%

(b) By MID "Country of Origin"

Country	Share
China	15%
Mexico	13%
Canada	12%
Japan	9%
Germany	5%
Taiwan	4%
South Korea	3%
Great Britain	3%
Hong Kong	3%
Switzerland	3%

Note: The "exporting country" can differ from the "country of origin" of a trade transaction, and typically the "exporting country" is the last stop without significant origin-conferring operations. The left panel utilizes publicly available import data from the U.S. Census Bureau.

Table 7: Top 10 Export Cities to the U.S.

(a) By Number of Relationships

Location	City Code	Total Relationships	Likely City/Cities
Taiwan	TAI	50,196	Taipei
Hong Kong	HON	46,187	Hong Kong
China	SHA	45,385	Shanghai, Shantou
China	GUA	42,285	Guangzhou
China	SHE	38,064	Shenzhen, Shenyang
China	DON	29,602	Dongguan
China	JIA	24,177	Jiangmen, Ji'an
China	ZHE	20,815	Zhenyang, Zhejiang (Province)
Hong Kong	KOW	20,491	Kowloon
China	NIN	16,221	Ningbo

(b) By Value

Location	City Code	Total Trade (in billions of USD)	Likely City/Cities
Canada	CAL	43.0	Calgary
China	SHA	29.0	Shanghai, Shantou
Singapore	SIN	28.6	Singapore
Taiwan	TAI	24.3	Taipei
China	SHE	22.3	Shenzhen, Shenyang
Mexico	MEX	21.6	Mexico City
Hong Kong	HON	17.6	Hong Kong
China	JIA	15.0	Jiangmen, Ji'an
Mexico	CDJ	14.9	Ciudad Juarez
China	GUA	14.3	Guangzhou

Note: It is possible to use the three-character city code from the MID to rank the top exporting cities to the U.S. A relationship is a U.S. importer-foreign supplier combination. We have a total of 1,579,983 relationships in the data. "Likely cities" are determined from population-ranked cities within a country.

Table 8: Related Party Relationships

	Log Trade	Log Price
Related	0.149***	0.107***
	(0.002)	(0.003)
Country FE	Yes	Yes
Product FE	Yes	Yes
N	4,440,000	3,110,000

Note: In the U.S. import data, two parties are considered to be related by ownership if one owns 5% or more of the other. Other possibilities for related party affiliation are family ties, an employer/employee relationships, or shared leadership. Log Trade refers to the logged total value of trade within the relationship (importer-exporter-product combination) in 2011, while the Log Price is the total value in the relationship divided by the total quantity. Observations are at the buyer-supplier-product level. Observation counts are rounded for disclosure purposes. Coefficients are significant at the 1% level.

Table 9: Top 10 Export Countries to the U.S.

By Number of Relationships

Country	Share
China	27%
Canada	7%
Hong Kong	6%
Italy	6%
Taiwan	5%
Germany	5%
Great Britain	4%
India	4%
Japan	3%
Korea	3%

Note: This table ranks U.S. export partners by the total number of importer-exporter relationships.

Table 10: Source Country Characteristics and Relationships

	Number of Relationships
Log GDP Per Capita	1.10***
	(0.007)
Log U.S. Exports	-0.097***
	(0.004)
Product FE	Yes
N	3,500,000

Note: This is a regression of the number of relationships within a source country-HS10 product group on source country characteristics. The observation count is rounded for disclosure purposes. Log GDP Per Capita in 2011 comes from the World Bank World Development Indicators. Log U.S. Exports are from publicly available U.S. Census Bureau totals. Coefficients are significant at the 1% level.

Table 11: Product Characteristics and Relationships

	Number of Relationships
Air Share	1.06^{***}
	(0.025)
Country FE	Yes
N	3,600,000

Note: This is a regression of the number of relationships within a source country-HS10 product group on product characteristics. The observation count is rounded for disclosure purposes. Air Share is the share of total U.S. imports in an HS10 category that comes via air. Coefficient is significant at the 1% level.

A Examples of the Bigram Matching Program

In Section 3.1, we describe the procedure whereby we collapse "similar" Manufacturer IDs into a single Manufacturer ID, where "similar" is defined as a score, calculated according to the number of matching bigrams within the Manufacturing ID. The procedure follows Wasi and Flaaen (Forthcoming) in order to calculate such a score. We have described rules of thumb to choose bigram matching scores in order to "clean" the MIDs. Here, we provide detailed examples of matches between MIDs and the associated scored, using hypothetical MIDs. Consider the following hypothetical firm name and address:

Quan Kao Company
1234 Beijing Lane
Beijing, China

Following the rules described in Section 2, the Manufacturing ID for this firm would be: CNQUAKAO1234BEI. Below we present seven permutations of this Manufacturer ID, along with their accompanying bigram matching score.

As can be seen from the table, the closer the two strings are, the higher is the associated match score. Furthermore, our criterion of consolidating similar firms if the two codes have similarity indices of over 0.98 seems reasonable according to the above standards: while some simple coding errors (such as missing one character in the name) might be reasonable to assume as potentially occurring in the data, errors on the scale of wholly different addresses or firm names are certainly likely to be much less common.

Table A1: Hypothetical MIDs and Bigram Matching Scores

Raw MID to be Matched	**Possible Matches**	**Difference**	**Score**
CNQUAKAO**1234**BEI	CNQUAKAO**123**BEI	One Character Missing	0.9951
CN**QUAKAO**1234BEI	CN**QUAKAU**1234BEI	One Character Different	0.9917
CN**QUAKAO**1234BEI	CN**QUA**1234BEI	Second Word Missing	0.9830
CNQUAKAO1234**BEI**	CNQUAKAO1234**SHA**	Different City	0.9802
CN**QUAKAO1234**BEI	CN**QUAKAO**BEI	No Number	0.9723
CNQUAKAO**1234**BEI	CNQUAKAO**5555**BEI	Different Number	0.9381
CN**QUAKAO**1234BEI	CN**JIACHA**1234BEI	Different Name	0.5321