Received: 4 May 2016 | Revised: 5 July 2017 | Accepted: 6 July 2017

DOI: 10.1111/roie.12306

ORIGINAL ARTICLE

WILEY

Identifying foreign suppliers in U.S. import data

Fariha Kamal[1] | **Ryan Monarch**[2]

[1] U.S. Census Bureau, Washington DC

[2] Board of Governors of the Federal Reserve System, Washington DC

Correspondence
Ryan Monarch, Board of Governors of the Federal Reserve System, 20th Street and Constitution Avenue N.W., Washington DC 20551.
Email: ryan.p.monarch@frb.gov.

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 enhance supplier information in U.S. import data and present new facts about importer–exporter relationships. Count of foreign exporters from U.S. import data tends to exceed those from source country data, especially from China. The pattern of U.S. imports from origin countries changes substantially by tracing trade back to the supplier's location instead. Related-party relationships trade more, while larger countries have more relationships.

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 availability of datasets 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, Claro & Horstmann, 2013), Costa Rica, Ecuador, and Uruguay (Carballo, Ottaviano, & Martincus, 2013), Norway (Bernard, Moxnes, & Ulltveit-Moe, 2014), and the United States (Pierce & Schott, 2012; Dragusanu, 2014; Eaton, Eslava, Krizan, Kugler, & Tybout, 2014; Monarch, 2014; Kamal & Sundaram, 2016, 2017; Heise, 2016; Monarch & Schmidt-Eisenlohr, 2016).

One of the primary concerns about two-sided trade transactions 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 their 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 data representing foreign suppliers to the United States, discusses potential concerns about

the quality of the data as well as some suggested refinements, and presents new findings about relationships between U.S. buyers and their foreign suppliers.

We first describe the method for identifying foreign suppliers in U.S. merchandise import transactions.[1] U.S. importing firms with shipments above 2000 U.S. dollars are required to complete U.S. Customs and Border Protection (CBP) Form 7501, part of which entails constructing and reporting a code—known as the Manufacturer ID or "MID"—for the foreign supplier in the transaction. The MID is widely used by both the U.S. and Canadian governments for official purposes. We explore the potential for errors that may arise in completing the MID, and note that 13% of U.S. import value is associated with transactions with no MID. Additionally, we show using external data that following the rules of MID creation, as outlined by CBP, tends to generate unique identifiers for suppliers within sectors.

After this investigation, we describe our efforts to update the MID in U.S. merchandise import transactions. We correct for possible clerical errors that may arise as importers construct this variable. Then, we collapse very similar MIDs into a single MID using string similarity scores. Following this, we perform various "stress tests" of our changes, showing both that our foreign supplier identifier improves the reliability of related-party relationships, and that MIDs we group together share very similar characteristics, such as sectors or buyers.

In the last part of the paper, we present five empirical patterns derived from our refined data on foreign suppliers selling to U.S. importers. First, U.S. import data tends to identify more exporters to the U.S. than foreign export data (especially in the case of China) on average. However, exporter counts match well within many broad sectors. Second, there is significant churning in the population of suppliers to the U.S., with rampant exit each year. Third, 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 were exports assigned to the original location of production. Fourth, related-party relationships exhibit higher trade volumes and higher prices. Finally, we find that larger countries, as well as countries in a trade agreement with the United States, tend to have more relationships with U.S. importers and higher value per relationship.

The paper proceeds as follows. Section 2 describes the MID in greater detail, including the institutional reasons it is included on customs forms and assessing its uniqueness. Section 3 presents our grouping methodology and related stress tests. Section 4 uses the updated data to establish our core set of stylized facts, and the last section concludes.

2 | BACKGROUND AND HISTORY

2.1 | MID creation

U.S. importers are required to fill out CBP Form 7501 (see Figure 1) in order to complete importation of goods into the United States. Importing firms must record information about the value, quantity, and 10-digit Harmonized Tariff Schedule of the United States (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 MID 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).

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	26. Importer of Record Name and Address
City State Zip	City State Zip

27. Line No.	28. Description of Merchandise: 29. A. HTSUS No. B. ADA/CVD No.	30. A. Grossweight B. Manifest Qty.	31. Net Quantity in HTSUS Units	32. A. Entered Value B. CHGS C. Relationship	33. A. HTSUS Rate B. ADA/CVD Rate C. IRC Rate D. Visa No.	34. 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
			D. Ascertained Other	39. Other
			E. Ascertained Total	40. Total

36. DECLARATION OF IMPORTER OF RECORD (OWNER OR PURCHASER) OR AUTHORIZED AGENT

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)

FIGURE 1 CBP Form 7501

Customs Directive No. 3550-055 lays out the method for deriving the MID 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 International Organization for Standardization (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 15 characters.

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.

The multi-step process for constructing the MID described above may raise concerns about potential for erroneous data entry. There are some mitigating factors, though. First, 96% of all entries are filed electronically through the CBP's Automated Broker Interface, which reduces 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. Customs Broker License Examinations administered by CBP (passage of which is required if transacting customs business on behalf of others) typically include questions about MID construction.[5] Third, 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 owing to manual data entry.

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 collect export filings to the United States. Instead, they substitute U.S. import statistics for Canadian exports to the U.S. in accordance with a 1987 Memorandum of Understanding, signed by the U.S. Census Bureau, U.S. CBP, Canadian Customs, and Statistics Canada.[7] Based on extensive discussions with employees at the U.S. Census Bureau and Statistics Canada, we believe that the data exchange provided the main impetus for the generation of the MID. Filling out the MID was made a requirement for U.S. imports from all countries soon after.

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 measures and best practices.[8] 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 enforces trade-related regulatory requirements that rely on the identity of foreign suppliers to the country. For instance, anti-dumping measures are foreign-firm specific in nature. Furthermore, it is clear from U.S. regulations that the MID is used to track compliance with U.S. restrictions for textile shipments. MID criteria for textiles are the most stringent, 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 textile shipments 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 preceding discussion highlights the regulatory imperatives to provide an accurate MID and the incentives for U.S. importers to accurately identify the foreign manufacturers from whom they are importing.

2.3 | Missing MIDs

The previous sections described the construction of the MID on the part of the U.S. importing firm as mandatory, thus providing a window into the universe of suppliers exporting to the United States. The field is not always populated, however: MIDs are missing in 1.9 percent of the 59 million import transactions in 2011. On a value-weighted basis, 13 percent are missing an MID, indicating that transactions without MIDs tend to be large.

Why might an MID be missing? We report coefficients from regressing a dummy variable equal to one for a missing MID on a host of covariates and report the results in Table 2. The first column—

TABLE 2 Determinants of missing MIDs

	(1)	(2)	(3)	(4)	(5)
Size					
Size Q2	0.017***				0.027***
	(0.000)				(0.000)
Size Q3	0.023***				0.037***
	(0.000)				(0.000)
Size Q4	0.035***				0.045***
	(0.000)				(0.000)
Related party status					
Related party		−0.035***			−0.046***
		(0.000)			(0.000)
Sector					
Mineral products			0.023***		0.039***
			(0.000)		(0.000)
Chemical products			0.013***		0.017***
			(0.000)		(0.000)
Plastics and rubber			0.022***		0.022***
			(0.000)		(0.000)
Hides/skins/leather/fur			0.026***		0.013***
			(0.000)		(0.000)

(Continues)

TABLE 2 (Continued)

	(1)	(2)	(3)	(4)	(5)
Textiles			0.009***		−0.001***
			(0.000)		(0.000)
Footwear/headgear			0.042***		0.023***
			(0.000)		(0.000)
Base metals			0.016***		0.013***
			(0.000)		(0.000)
Machinery/electrical			0.025***		0.022***
			(0.000)		(0.000)
Vehicles			0.022***		0.028***
			(0.000)		(0.000)
Optical and arms			0.017***		0.011***
			(0.000)		(0.000)
Source					
North America				−0.018***	−0.027***
				(0.006)	(0.006)
Central America				−0.014***	−0.008
				(0.006)	(0.006)
South America				−0.013***	−0.004
				(0.006)	(0.006)
Europe				0.022***	0.026**
				(0.000)	(0.000)
Asia				0.004	−0.002
				(0.006)	(0.006)
Australasia and Oceania				−0.016***	−0.012**
				(0.006)	(0.006)
Africa				−0.001	0.005
				(0.006)	(0.006)
Constant	0.001***	0.035***	0.001***	0.019***	−0.002
	(0.000)	(0.000)	(0.000)	(0.006)	(0.006)
R^2	0.01	0.02	0.01	0.01	0.05

Note. The dependent variable is 1 if a transaction's MID is missing. Omitted catgeories are Q1, arm's length, live animals, and Puerto Rico & U.S. posessions. There are 46,000,000 observations. *,**,***Denote significance at 1%, 5%, and 10% levels, respectively.

based on importer size bins—shows that bigger buyers are more likely to be missing MIDs. One possible explanation for this is that 98.8 percent of missing MIDs (and thus, some big importers) are associated with foreign trade zone transactions. A foreign trade zone (FTZ) is 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.[9] Since firms that import in high volumes at a regular frequency are the main participants in foreign trade zones, they tend to be larger firms.

Table 2 also shows that related-party transactions are less likely to be missing an MID, while broad sectors such as "footwear/headgear," "hides/skins/furs/leather," and "machinery/electrical" are more likely to be found without an MID.[10] Transactions with European source countries tend to have a higher likelihood of missing MIDs on average, while the Americas tend to have a lower likelihood. Our takeaway is that transactions without MIDs tend to be conducted by larger importers, but incidence does not vary dramatically across sectors or countries.[11]

2.4 | Checking exporter identification under MID rules

Even if U.S. importers are completing the MID correctly, there remains the possibility that the information collected is too limited to uniquely identify distinct suppliers.[12] Foreign production data can be used to construct MIDs according to the rules laid out above, allowing determination of how often this identifier uniquely identifies the foreign country's suppliers.[13]

We carry out this exercise using Chinese production data, translating the universe of exporter names and addresses from the Chinese Annual Survey of Industrial Firms (ASIF) in 2005 and constructing "MIDs" following the algorithms set forth by CBP. This allows us to assess (indirectly) how common it is for an MID to uniquely identify an exporter, both overall and within a sector. Are different cities combined into a single city when only identified with three letters? The firm-level data is collected by the Chinese National Bureau of Statistics (NBS), and we romanize Chinese characters according to the Hanyu Pinyin system.[14] We use the four-digit China Industrial Classification (CIC) to report sectoral results.

This is not an attempt to link Chinese exporter information in the ASIF to the actual MIDs reported in U.S. import data; rather, the results below are only a general test of MID rules. We highlight three caveats that our findings are premised on. One, observations of Chinese production data are at the firm level, while the MID is meant to capture manufacturers. Second, it is possible that a single supplier appears multiple times by different filers in the U.S. import data, while the information on the manufacturing census is a year-end snapshot, presenting additional opportunity for discrepancies. Finally, we note the possibility that our concordance between Chinese characters and English may differ from what reporting firms use, or that Chinese firms may not use a direct translation of their name on their invoices.

We begin with results on uniqueness. Overall, the ASIF reports approximately 75,000 exporters in 2005. The "MID" is unique for 63.4 percent of the reported exporters. This raises a non-negligible possibility of MID duplication when used in a vacuum without any other identifying information. The first line of Table 3a shows that out of 515 total CIC4 sectors, the average sector has 95.7 percent of its exporters uniquely identified by an "MID", a major improvement. Limiting to sectors with large numbers of exporters still shows greater uniqueness—among sectors with over 1,000 exporters, 84.7 percent of exporters are uniquely identified by an "MID". Even adding two-digit CIC2 sectors significantly improves the identification, as shown in the bottom half of the table. Thus a supplier identifier combined with industry information greatly increases the likelihood of generating unique identification.[15] That said, there is a large discrepancy between the number of Chinese exporters according to the Chinese data and that calculated according to the U.S. data—there were about 173,000 MIDs from China in 2005, over twice as many as in the NBS data.[16] As we will show in Section 4.1, this is large

TABLE 3A Analysis of "MIDs" as constructed from China industrial production data: Uniqueness of the MID, 2005

	% Unique "MIDs", Average	Number of CIC4s
All CIC4s	95.7	515
CIC4s with > 100 Exporters	92.1	185
CIC4s with > 500 Exporters	83.8	24
CIC4s with > 1000 Exporters	84.7	7
	% Unique "MIDs", Average	**Number of CIC2s**
All CIC2s	87.5	39
CIC2s with > 100 Exporters	85.2	32
CIC2s with > 500 Exporters	84.6	27
CIC2s with > 1000 Exporters	84.9	22

Note. This panel uses name, address, and city information from China NBS data to construct an "MID" for each exporter, following rules from U.S. CBP Form 7501. CIC4 is the four-digit China Industry Code, and CIC2 is its two-digit counterpart. For the English name of the firm, the Hanyu Pinyin romanization of Chinese characters, with two to three characters per word, is used. An "MID" is unique if it corresponds to one *faren daima* firm identifier.

compared with other country sources, so where necessary, we present our empirical results excluding China in order to establish robustness of our findings.

Table 3b uses the same augmented Chinese production data, but illustrates how common it is for a sector to have multiple cities with the same three-letter city code. With 145 cities of over 1 million people in 2010, China represents a difficult country for trying to identify cities uniquely with only three letters. The table shows that the more cities that export a particular CIC4 category, the smaller the share of unique city identification within that CIC4 category. Even so, in the average CIC4 sector, cities are uniquely identified by their code 86.8 percent of the time.

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 United States. Nonetheless, we undertake both

TABLE 3B Analysis of "MIDs" as constructed from China industrial production data: Uniqueness of the city code, 2005

	% Unique City Codes, Average	Number of CIC4s
All CIC4s	86.8	515
CIC4s with > 10 Cities	84.0	417
CIC4s with > 50 Cities	72.6	97
CIC4s with >100 Cities	64.4	11

Note. This panel uses city codes from the "MIDs" constructed above. A city code—the first three letters of a city—is unique if there is only one city with that code in a CIC4 industry.

probabilistic matching methods and basic checks in order to increase the reliability of the data. In this section, we describe our methodology for cleaning the MID and offer some summary measures of the resulting supplier data. We use the 2011 Linked Firm Trade Transaction Database (LFTTD) in our main empirical analyses.

The first stage of our cleaning implements a number of common sense adjustments to the MID. We exclude MIDs that do 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 exclude MIDs that have a country code corresponding to no known ISO2 code.

3.1 | Bigram matching

We use a character matching protocol known as bigram matching to combine very similar MIDs into a single MID. A bigram is an approximate string comparator, computed from the ratio of the number of common two-letter combinations within the two different strings and the average string length minus one. We use the STATA-based bigram matching algorithm developed by Wasi and Flaaen (2015). All possible MID pairs within a country are assigned a field similarity score in order to set a standard for determining if any MID is "similar enough" to another MID.[17] 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? We identify a few rules of thumb for field similarity (where 1 means a 100 percent match): a score of 0.98 or higher tends to match MIDs with a few characters being different, while scores between 0.97 and 0.98 tend to match MIDs that are identical in all aspects, other than one has a numeric address field and the other does not. A score of 0.99 or higher typically has only a single character being different. For our main results, we adopted a field similarity score of 0.98, such that we are likely to combine MIDs that differ owing to simple typographical errors (for instance, one character differences or one MID only using the first name of a company), but we will consider 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 combine two distinct suppliers.[18]

The implementation procedure is as follows: within a country, we match each MID to every other MID, and generate 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.[19] Retaining one MID per group leaves a "best MID" (or BMID) variant for each MID in the underlying data, which enables us to generate relationships and other supplier-specific variables (such as size) at the BMID level.[20] All told, these changes together with the above methodology end up reducing the total number of suppliers in 2011 from 1,287,630 to 911,765, a 29 percent decrease.

3.2 | Cross-validation tests of the BMID

In this section, we offer two tests in order to assess how well the bigram matching procedure is capable of grouping together similar MIDs and hence the extent to which the BMID can be viewed as a valid identifier of foreign suppliers to the United States.

First, we examine related-party trade relationships. U.S. firms are required to write down (in column 32C, Form 7501) whether the transaction took place 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

TABLE 4 Validity checks for bigram matching results

	Broad sector	HS2	HS10	Buyer
Changed MIDs that match any (%)	64.0	40.1	26.5	30.4
Random MID pairs that match any (%)	4.1	2.3	0.3	0.6
Changed MIDs that match all (%)	47.4	23.9	9.4	16.4
Random MID pairs that match all (%)	1.3	0.5	0.0	0.2

Note. This table shows how similar information from changed MIDs is to information from the original MID.

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. This implies that when examining the raw MID in relation to the BMID, we would expect a smaller fraction of total relationships that are marked as being related in one transaction while having a missing or unrelated indicator in another. Consistent with our hypothesis, when using the BMID, we find a decrease, from 5.8 to 5.5 percent, in the share of relationships that mix the related and nonrelated indicators across transactions.

Second, we examine if the changes are consistent with other information in the U.S. import data, including sector, product and buyer information.[21] The exercise is as follows. Suppose the bigram matching method designates Supplier A to be the same as Supplier B in 2011, and thus Supplier A's MID is replaced with Supplier B's. How often do Supplier A and Supplier B share all the same sectors, products, or buyers? How often do they share any sector, product, or buyer?[22] Table 4 presents the results. At our preferred score of 0.98, 64 percent of changed MIDs have any "broad sector" in common with the MID they are being changed to, 40.1 percent share the same HS2 code, 26.5 percent share the same HS10 code, and 30.4 percent share the same buyer (U.S. importer). In comparison, if an MID is randomly matched to some other MID within its country, the probability the two MIDs share any broad sector is 4.1 percent, any HS2 category is 2.3 percent, any HS10 category is 0.3 percent, and any buyer is 0.6 percent.[23] This comparison highlights that the bigram matching method groups MIDs that share characteristics other than the identifier itself, even though the routine does not require matches within products or buyers. We thus believe that our probabilistic matching routine, resulting in the BMID, is capable of identifying "similar" MIDs, providing an improved identifier of foreign suppliers.

3.3 | Describing the MID sample

We next illustrate some of the properties of our sample, using the BMID. The minimum length of any BMID in the data is 11 characters, and the maximum is 15 characters. Table 5 shows that BMIDs are almost evenly split between 11, 12, 13, 14, and 15 characters. Nineteen percent of these codes are the maximum length allowable—15 characters. Table 6 Panel (a) shows how often the address component

TABLE 5 Distribution of BMID lengths

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

Note. The maximum MID length is 15 characters. Cleaned sample BMIDs have a minimum of 11 characters.

TABLE 6 BMID 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 0–4 numeric characters in the address field, taken from the supplier's invoice.

of the BMID is populated: the vast majority of BMIDs (89 percent) have at least some address information included.[24]

A potential issue concerning the address component of the MID is the presence of nonnumeric address conventions in Latin America. For example, according to a 2007 *Los Angeles Times* article, "most Costa Rican addresses are expressed in relation to the closest community landmark."[25] Theoretically, this could result in fewer fully-populated address codes.[26] Table 6 Panel (b) shows that BMIDs from South America and "Mexico and Central America" do not actually exhibit lower rates of numeric address components compared with other regions. Europe, Asia, and Africa all have larger fractions with no address information. However, Costa Rica (Table 6 Panel (c)) is an exception, as about 18 percent of Costa Rican MIDs have no address information. Table 6 Panel (b) also shows that North American MIDs (predominantly Canada) have full address information for almost half of all MIDs—not surprising, given that Statistics Canada successfully matches MIDs to their domestic establishments.

An additional concern may be 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 that an MID-identified supplier is shipping. Intermediaries are more likely to export products spanning different industries (Ahn, Khandelwal, & Wei, 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 7 shows that 96 percent of BMIDs export five or fewer HS2 codes, and 97 percent of BMIDs export 10 or fewer HS10 codes. In subsequent analysis, we exclude BMIDs with more than 10 HS2 codes from our sample.

TABLE 7 Distribution of BMIDs, 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 BMIDs to the United States by the number of products or industries exported.

4 | FINDINGS FROM RELATIONSHIP-LEVEL TRADE DATA

In this section, we present a set of empirical regularities, relying on our BMID variable and the 1,579,983 importer–exporter relationships formed by the combination of a BMID and a U.S.firm identifier.

4.1 | Comparisons with foreign export data

We use foreign data on the number of exporters exporting to the U.S. in 2011, and compare the total with the number of exporters calculated using U.S. data, stratifying by both country and sector. We use the World Bank's public-use Exporter Dynamics Database (EDD) that contains destination-specific information on exporting firms for 70 countries between 1997 through 2014 (Cebeci, Fernandes, Freund, & Pierola, 2012). The source of the underlying micro data, which is not publicly available, varies from national government statistics (such as in Peru) to figures collected by private companies (such as in Chile) and are thus wholly different sources than the U.S. customs data. The idea is to compare statistics from the two distinct sources and analyze how closely they align, keeping in mind that the definition of what exactly constitutes a foreign exporter is specific to the U.S., and need not match across different countries.

The country comparison is presented in Table 8. The table contains 41 countries, a number determined by both the availability of destination-specific data and official Census Bureau disclosure rules. The total number of exporters calculated from the EDD is 73% of the total using U.S. data.[27] For some countries, such as Mexico and Spain, the exporter counts match particularly well, while others, such as Germany and Portugal, show much less agreement. Why might using the Manufacturer ID to generate counts of firms exporting to the U.S. result in too many exporters relative to source country data? One answer rests on one of CBPs requirements in constructing the Manufacturer ID: trading companies, sellers other than manufacturers, and similar trading agents cannot be used to create MIDs. Since source countries may count intermediaries (who purchased from multiple manufacturers) as exporters in their customs data, origin-country data compared with the U.S. data is likely to yield lower counts of exporters. Another reason for differences is that MIDs are potentially written down multiple times by different importing firms throughout a year, while firm-level export data is collected only a single time. Thus if firm names change over time, or importing firms construct the MID differently, U.S. data will report a greater number of foreign exporters.

Next, we show a comparison of different exporter counts by "broad sector."[28] The EDD contains the number of exporters to the U.S. in a HS2 category for each country in Table 8. We sum these counts across countries by HS2 sector, aggregate to the "broad sector" level, then compare it with the

TABLE 8 Number of exporters to the United States, 2011

Country	World Bank EDD	BMIDs	Share
Albania	50	63	0.79
Bangladesh	2,051	2,667	0.77
Belgium	4,589	6,823	0.67
Bolivia	268	476	0.56
Brazil	5,772	9,080	0.64
Cameroon	94	107	0.88
Chile	2,072	2,959	0.70
Colombia	2,663	4,045	0.66
Costa Rica	1,599	1,476	1.08
Cote d'Ivoire	139	123	1.13
Croatia	421	295	1.43
Denmark	3,139	4,783	0.66
Dominican Republic	1,613	1,477	1.09
Ecuador	1,466	1,915	0.77
Egypt	851	1,200	0.71
Estonia	302	265	1.14
Ethiopia	240	220	1.09
Georgia	102	63	1.62
Germany	28,229	48,398	0.58
Guatemala	1,370	1,699	0.81
Jordan	574	346	1.66
Kenya	413	377	1.10
Lebanon	318	398	0.80
Madagascar	174	173	1.01
Mauritius	156	184	0.85
Mexico	24,802	27,523	0.90
Morocco	429	863	0.50
Nepal	573	759	0.75
Nicaragua	382	621	0.62
Norway	1,940	2,512	0.77

(Continues)

TABLE 8 (Continued)

Country	World Bank EDD	BMIDs	Share
Paraguay	68	144	0.47
Peru	2,396	3,271	0.73
Portugal	2,413	3,809	0.63
Romania	854	1,420	0.60
South Africa	3,257	3,416	0.95
Spain	13,888	13,115	1.06
Turkey	4,316	8,275	0.52
Uruguay	379	517	0.73
Total	114,362	155,857	0.73

Note. This table compares the number of exporters from two different datasets. The last column is the exporter count from the World Bank EDD as a fraction of the exporter count from the U.S. import data.

same object in the U.S. data.[29] Table 9 shows that counts within sector groupings match fairly well on average, with a few exceptions—the number of "Chemical and allied industries" exporters in the U.S. data exceeds the number in the EDD data by a factor of 4, and "Wood & wood products" by a factor of 3. The EDD/LFTTD count ratio for textiles—which we had described earlier as being a likely candidate for well-constructed MIDs—is extremely close to one, as are sectors such as Plastics, Footwear/Headgear, and others.

4.2 | Dynamic behavior of suppliers

One key feature of the import data is that importer–exporter relationships are extremely short-lived. Monarch and Schmidt-Eisenlohr (2016) find that close to half of all U.S.–foreign supplier relationships in 2011 are newly created. We use BMIDs to demonstrate that exporter exit is, in fact, extremely high. Table 10 shows that of all BMIDs found in 2011, only 54 percent are found in the U.S. data in 2010, while only 56 percent are found in 2012. The similarity to the relationship numbers in previous work shows that exporter exit accounts for a large share of relationship dissolution over time, as fewer and fewer exporters manage to survive into later years. Importantly, as the second row of the table demonstrates, this stylized fact is unchanged when considering simply the raw MID variable in place of our BMID variable, meaning that collapsing similar MIDs does not alter dynamic features of the supplier data. One other salient fact about suppliers over time is they do occasionally change their exported product: about 68 percent of MIDs have the same HS2 category or set of categories in 2011 that they had in 2010.

4.3 | 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 percent of relationships (accounting for 29 percent of total U.S. imports), the exporting

TABLE 9 Number of exporter–sector combinations to the United States, 2011

Broad HS category	EDD/LFTTD share
Plastics & rubber	0.92
Raw hides, skins, leather, & furs	0.94
Footwear & headgear	0.96
Textiles	1.03
Vegetable products	1.16
Transportation	1.23
Prepared foodstuffs	1.24
Machinery & electrical	1.43
Miscellaneous	1.54
Animal & animal products	1.58
Stone & glass	1.58
Mineral products	2.01
Metals	2.57
Wood & wood products	3.27
Chemical & allied industries	4.29

Note. This table compares the number of exporters to the United States in an HS2 sector from two different datasets, for the countries listed in Table 8. The counts are aggregated to the "broad sector" classification. The last column is the count from the World Bank EDD as a fraction of the count from the U.S. import data.

country does not 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 origin country? CBP Instructions state that "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, such a discrepancy likely means that the "exporting country" is reexporting the goods. In other words, if already-produced goods were not substantially transformed, but instead repackaged or re-sold from a second country, then the second country would be listed as the official exporting country.

TABLE 10 Share of 2011 suppliers found in other years

	2010	2012	2013	2014
BMID	54%	56%	45%	40%
Raw MID	54%	56%	46%	40%

Note. This table displays the percent of MIDs in 2011 that were also found in 2010, 2012, 2013, or 2014, as a share of all MIDs in 2011. BMIDs are those that were combined via the bigram matching procedure, while "Raw" refers to the MID as it appears in the trade transactions data.

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 11 presents the top 10 exporters to the U.S. in 2011 by both origin and production countries. Interestingly, although 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 a major reexporter with a comparatively low value-added to export ratio (Johnson & Noguera, 2012). 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.

4.4 | Related-party relationships

According to official Census Bureau trade statistics, trade within related parties typically accounts for about 40 percent of all U.S. annual merchandise imports. Since we can use the BMID to identify related-party relationships in the data, we can contrast them to arm's-length relationships. We find that related-party relationships occupy a very small share of total relationships, only 6.6 percent. 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 are associated with high-value transactions. Indeed, a simple regression with product and source country fixed effects shows that related-party relationships—at the buyer–

TABLE 11 Top export sources to the United States, 2011

Country	Share of total value (%)
(a) By "exporting country"	
China	18
Canada	14
Mexico	12
Japan	6
Germany	5
South Korea	3
United Kingdom	2
Saudi Arabia	2
Venezuela	2
Taiwan	2
(b) By BMID "country of origin"	
China	15
Mexico	13
Canada	12
Japan	9
Germany	5
Taiwan	4
South Korea	3
United Kingdom	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 12 Related party relationships

	Log Trade	Log Price
(a) All relationships		
Related	0.149***	0.107***
	(0.002)	(0.003)
Country FE	Yes	Yes
Product FE	Yes	Yes
Adj. R^2	0.22	0.66
N	4,440,000	3,110,000
(b) Excluding China		
Related	0.262***	0.044***
	(0.002)	(0.003)
Country FE	Yes	Yes
Product FE	Yes	Yes
Adj. R^2	0.25	0.66
N	3,394,000	2,341,000

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, 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. ***Denotes significance at the 1% level.

supplier-product level—trade more than nonrelated parties (Table 12, column 1). We also find that related-party relationships tend to have higher unit values (Table 12, column 2). This effect is precisely estimated, as we use trade and unit values at the relationship level, rather than at the firm level only.

TABLE 13 Share of total relationships, by export source to the United States

Country	(a) 2011 cleaned sample Share (%)	(b) Sample in 2010 & 2011 Share (%)	(c) 2010 raw sample Share (%)
China	27	20	27
Canada	7	6	9
Hong Kong	6	4	6
Italy	6	4	5
Taiwan	5	4	5
Germany	5	4	5
United Kingdom	4	3	4
India	4	2	3
Japan	3	2	3
South Korea	3	2	3

Note. This table ranks U.S. export partners by the total number of importer–exporter relationships. Column (a) uses BMIDs from our cleaned data for 2011. Column (b) uses only MIDs that were found in both 2010 and 2011 U.S. import data. Column (c) uses only raw 2010 U.S. import data.

As we documented in Section 2.4, there is a substantial difference in the number of Chinese exporters to the U.S. across datasets. Therefore, as a robustness check, we run the same specification as above, excluding China. The results remain qualitatively similar—related parties have higher trade values and higher prices compared with nonrelated-party trade relationships.

4.5 | Relationships and country characteristics

Using our BMID measure, Table 13 shows which countries have the most supplier relationships with the U.S. in 2011. Over a quarter of all importer–exporter relationships in 2011 were between U.S. buyers and mainland Chinese suppliers, a share that bumps up to one-third of all relationships if we include Hong Kong. The rankings are not significantly altered by restricting only to MIDs found in both 2010 and 2011, or by using 2010 data alone, as shown in columns (b) and (c).

We next examine why some source countries have more U.S. relationships than others. We estimate a gravity-like specification, regressing a host of country attributes on the (log) number of supplier relationships and the average trade value per relationship. Table 14 shows that larger countries (measured by log GDP) tend to have more relationships, and also have more value per relationship. Farther away locations and source countries without a common language with the U.S. have fewer

TABLE 14 Country characteristics and relationships

	All sectors		Textiles only	
	No. of relationships	Value/relationship	No. of relationships	Value/relationship
Log GDP	0.97***	0.27***	0.82***	0.14
	(0.05)	(0.05)	(0.05)	(0.09)
Distance	−0.00***	−0.00	0.00	0.00*
	(0.00)	(0.00)	(0.00)	(0.00)
Contiguity	−0.34	−0.49	0.25	0.42*
	(0.34)	(0.65)	(0.89)	(0.64)
Common language	0.94***	0.32	0.163	−0.26
	(0.24)	(0.24)	(0.34)	(0.43)
Former colony	0.11	−0.64	0.36	0.05
	(0.88)	(0.41)	(0.88)	(0.53)
RTA	1.20***	0.66***	1.42***	2.58***
	(0.22)	(0.27)	(0.44)	(0.61)
Common currency	0.44	1.30*	0.21	−0.90
	(0.38)	(0.70)	(0.31)	(0.67)
Observations	177	177	153	153
Adj. R^2	0.72	0.20	0.52	0.17

Note. This is a regression of the number of relationships and value per relationship on source country characteristics. Covariates come from the CEPII gravity database. *,**,***Denote significance at 1%, 5%, and 10% levels, respectively.

relationships, while being in a regional trade agreement with the U.S. strongly predicts both more relationships and higher trade per relationship.

As described above, the textile industry is one where we believe MIDs are particularly likely to represent the manufacturer, given the importance in U.S. law of establishing a proper country of origin for textile products. For this reason, we also conduct our gravity specification using only products that are classified as textiles in the HS system—HS2 50 through 63. As can be seen in the right side of Table 14, the key results of a positive effect of GDP on the number of supplier relationships with the U.S. holds, as does the effect of being in a regional trade agreement.

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 study U.S. importer–foreign exporter relationships. We document the rules and laws that govern the generation of the MID, then propose a set of cleaning algorithms and procedures meant to augment the reliability of the MID as a measure of unique foreign suppliers. This includes collapsing very similar MIDs into one, as well as commonsense checks for erroneous entries. Finally, we illustrate new findings about foreign buyers and their relationships with U.S. buyers.

In any national dataset attempting to measure information on foreign firms, there are bound 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 that 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 computationally intensive 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.

ACKNOWLEDGMENT

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, Tim Schmidt-Eisenlohr, Christian Volpe, and two anonymous referees 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.

NOTES

[1] We use the Linked Firm Trade Transaction Database (LFTTD) as maintained by the U.S. Census Bureau. See http://www.census.gov/ces/dataproducts/datasets/imp.html for further description.

[2] Because of 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 at the same link for more detailed instructions on constructing MIDs.

[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.

[7] This is a reciprocal data exchange, designed to reduce respondent burden, where Canada provides U.S. Canadian merchandise import shipments from the U.S. that the U.S. substitutes for exports to Canada.

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

[9] There are about 250 foreign-trade zones in the United States. See http://enforcement.trade.gov/ftzpage/info/ftzstart.html.

[10] "Related parties" refer to transactions with shared ownership or interest (see Section 3.2). The "broad sector" classification contains 15 sectoral groupings derived from grouping similar HS2 categories (see Online Appendix B). For access to the Online Appendix, see the Supporting Information at the end of this paper.

[11] These patterns suggest an area for further research on using probabilistic matching to assign an MID to such transactions. Large importers have more transactions. If an importer in a FTZ imports a similar product from the same country from a non-FTZ location and reports the MID in the non-FTZ transaction, it may be possible to assign a likely MID to the missing transactions.

[12] Using the MID to study supplier switching across or within cities may also pose challenges as countries may have multiple cities that begin with the same three letters. City codes have been used to study buyer–supplier relationships within and across countries in conjunction with country-specific knowledge (Kamal & Sundaram, 2016, 2017; Monarch, 2014).

[13] Note this exercise simply checks how well the MID coding procedure is capable of uniquely identifying suppliers and does not require linking directly to U.S. import data.

[14] In general, we take two to three Chinese characters to be one word of the company name. More detail is provided in Online Appendix C in the Supporting Information.

[15] Although the CIC classification system also comes from non-U.S. data, it is similar to the U.S. Standard Industrial Classification (SIC) (Brandt, Van Biesebroeck, & Zhang, 2012). This is why SIC/North American Industrial Classification System (NAICS) industry classifications (or the even more detailed HS classification) are likely to strengthen identification of foreign suppliers in U.S. data.

[16] We discuss reasons why counts from U.S. data often exceed those of source country data in Section 4.1.

[17] Other papers that use bigram matching include Anderson, Davies, Signoret, and Smith (2016), Ernstberger and Grüning (2013), Flaaen (2014), Green and Jame (2013), Chodorow-Reich (2014), and Braun and Raddatz (2010).

[18] We also implement a match score of 0.99 for some of the analysis in the Online Appendix (for access, see Supporting Information at the end of this paper).

[19] 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.

[20] The related-party status of a BMID relationship with both related and nonrelated-party transactions will be random.

[21] We thank two anonymous referees for this suggestion.

[22] We examine these categories in 2010 for matches made in 2011.

[23] These probabilities are calculated by repeating the following procedure 50,000 times: select a random MID in 2011, assign it another random MID in the same country, and calculate the frequency that they share the same broad sector, HS2, HS10 or buyer.

[24] Leading zeros are not allowed in the address component—since the preceding and following fields can only be based on letters, we can isolate how many numbers (if any) the address field of the MID contains.

[25] "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.

[26] 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.

[27] Section 2.4 showed that exporter counts from Chinese data (not in the EDD) are also substantially smaller than the same count from U.S. data.

[28] We thank two anonymous referees for suggesting this analysis.

[29] This exercise uses the number of exporter–HS2 combinations (not the number of unique exporters) within each "broad sector", since we cannot eliminate exporters who export more than one HS2 sector within a "broad sector" in the EDD. Thus the results of this exercise are not directly comparable with Table 8.

ORCID

Ryan Monarch http://orcid.org/0000-0002-2853-0893

REFERENCES

Ahn, JaeBin, Khandelwal, A. K. & Wei, S.-J. (2011). The role of intermediaries in facilitating trade. *Journal of International Economics*, *84*(1), 73–85.

Anderson, M. A., Davies, M. H., Signoret, J. E., & Smith, S. L. S. (2016). *Firm heterogeneity and export pricing in India* (Working Paper No. 2016-09-B). Washington DC: U.S. International Trade Commission.

Benguria, F. (2014). Production and distribution in international trade: Evidence from matched exporter–importer data (mimeo). Lexington, KY: University of Kentucky.

Bernard, A. B., Moxnes, A., & Ulltveit-Moe, K. H. (2014). *Two-sided heterogeneity and trade* (NBER Working Paper No. 20136). Cambridge, MA: National Bureau of Economic Research.

Blum, B. S., Claro, S., & Horstmann, I. J. (2013). Occasional and perennial exporters. *Journal of International Economics*, *90*(1), 65–74.

Brandt, L., Van Biesebroeck, J., & Zhang, Y. (2012). Creative accounting or creative destruction? Firm-level productivity growth in Chinese manufacturing. *Journal of Development Economics*, *97*(2), 339–351.

Braun, M., & Raddatz, C. (2010). Banking on politics: When former high-ranking politicians become bank directors. *The World Bank Economic Review*, *24*(2), 234–279.

Carballo, J., Ottaviano, G. I. P., & Martincus, C. V. (2013). *The buyer margins of firms' exports* (Discussion Paper No. 9584). London: Centre for Economic and Policy Research.

Cebeci, Tolga, Fernandes, A., Freund, C., & Pierola, M. (2012). *Exporter dynamics database* (Policy Research Working Paper No. 6229). Washington, DC: World Bank.

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

Dragusanu, R. (2014). Firm-to-firm matching along the supply chain (mimeo). Cambridge, MA: Harvard University.

Eaton, J., Eslava, M., Krizan, C. J., Kugler, M., & Tybout, J. (2014). *A search and learning model of export dynamics* (mimeo). State College, PA: Pennsylvania State University.

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

Flaaen, A. (2014). *Multinational firms in context* (Working Paper). Ann Arbor, MI: University of Michigan.

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

Heise, S. (2016). *Firm-to-firm relationships and price rigidity: theory and evidence* (CESifo Working Paper Series No. 6226). Munich, Germany: CESifo.

Johnson, R., & Noguera, G. (2012). Accounting for intermediates: Production sharing and trade in value added. *Journal of International Economics*, *86*(2), 224–236.

Kamal, F., & Sundaram, A. (2016). Buyer–seller relationships in international trade: Do your neighbors matter? *Journal of International Economics*, *102*, 128–140.

Kamal, F., & Sundaram, A. (2017). Spatial concentration of sourcing in international trade: The role of institutions (mimeo). Washington, DC: Center for Economic Studies, U.S. Census Bureau.

Melitz, M. (2003). The impact of trade on intra-industry reallocations and aggregate industry productivity. *Econometrica*, *71*(6), 1695–1725.

Monarch, R. (2014). *It's not you, it's me: Breakups in U.S.–China trade relationships* (Working Paper No. 14–08). Washington, DC: Center for Economic Studies, U.S. Census Bureau.

Monarch, R., & Schmidt-Eisenlohr, T. (2016). *Learning and the value of trade relationships* (CESifo Working Paper Series No. 5724). Munich, Germany: CESifo.

Pierce, J. R., & Schott, P. K. (2012). The surprisingly swift decline of U.S. manufacturing employment (NBER Working Paper No. 18655). Cambridge, MA: National Bureau of Economic Research.

U.S. Department of Homeland Security. (2012). CBP Form 7501 instructions. Washington, DC: Author.

Wasi, N., & Flaaen, A. (2015). Record linkage using STATA: Preprocessing, linking and reviewing utilities. *The Stata Journal*, *15*(3), 672–697.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article. [Correction added on 6 September 2017, after first online publication: The correct Supporting information file is now available.]

S1: roie_12306_sup_online_appendix

How to cite this article: Kamal F, Monarch R. Identifying foreign suppliers in U.S. import data. *Rev Int Econ*. 2018;26:117–139. https://doi.org/10.1111/roie.12306

APPENDIX A: EXAMPLES OF THE BIGRAM MATCHING PROGRAM

In Section 3.1, we describe the procedure whereby we collapse "similar" Manufacturer IDs into a single MID, where "similar" is defined as a score, calculated according to the number of matching bigrams within the MID. The procedure follows Wasi and Flaaen (2015) 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 MID for this firm would be: CNQUAKAO1234BEI. Table A1 presents seven permutations of this MID, along with their accompanying bigram matching score.

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

As can be seen from Table A1, 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.