# Are Customs Records Consistent Across Countries?

NASEM Workshop on "Innovation, Global Value Chains, and Globalization Measurement

C.J. Krizan[a], J. Tybout[b], Z. Wang[c], and Y. Zhao

[a]Dept. of Labor, [b]Penn State U. and NBER, [c]Shanghai U. Finance and Economics, [d]George Washington U.

May 6, 2021

# Acknowledgement and Disclaimer

*This research was supported by the National Science Foundation (Grant No. SES-1426645). All results have been reviewed by the U.S. Census Bureau to ensure that no confidential information is disclosed. Any opinions, findings, conclusions and recommendations expressed herein are those of the authors and do not necessarily represent the views of the NSF or the U.S. Census Bureau.*

# Motivation

- In recent decades, many trade researchers have exploited customs records to study:
  - Firm-level trade dynamics
    - market entry and exit
    - market penetration, customer accumulation
  - Patterns of technology diffusion
  - Study international business networks, GVCs
- This work relies heavily on micro patterns in the data
  - Firm-to-firm connections and durations of relationships in international markets
  - Firm-to-firm shipment frequencies, product clasifications, and values.
- But how accurately do the data describe these phenomena?

# This paper

- Our objective: contribute to a small literature assessing the reliability of these micro features of the data
- Compare export records generated by shipments leaving Colombia to the import records generated by the same shipments when they enter the U.S. as imports.
  - Begin with aggregates
  - Go down to HS2 industry
  - Go down to firm to firm
  - Then down to transaction by transaction

# This paper

- Our objective: contribute to a small literature assessing the reliability of these micro features of the data
- Compare export records generated by shipments leaving Colombia to the import records generated by the same shipments when they enter the U.S. as imports.
  - Begin with aggregates
  - Go down to HS2 industry
  - Go down to firm to firm
  - Then down to transaction by transaction
- Consider alternative explanations for discrepancies and implications for research.

# This paper

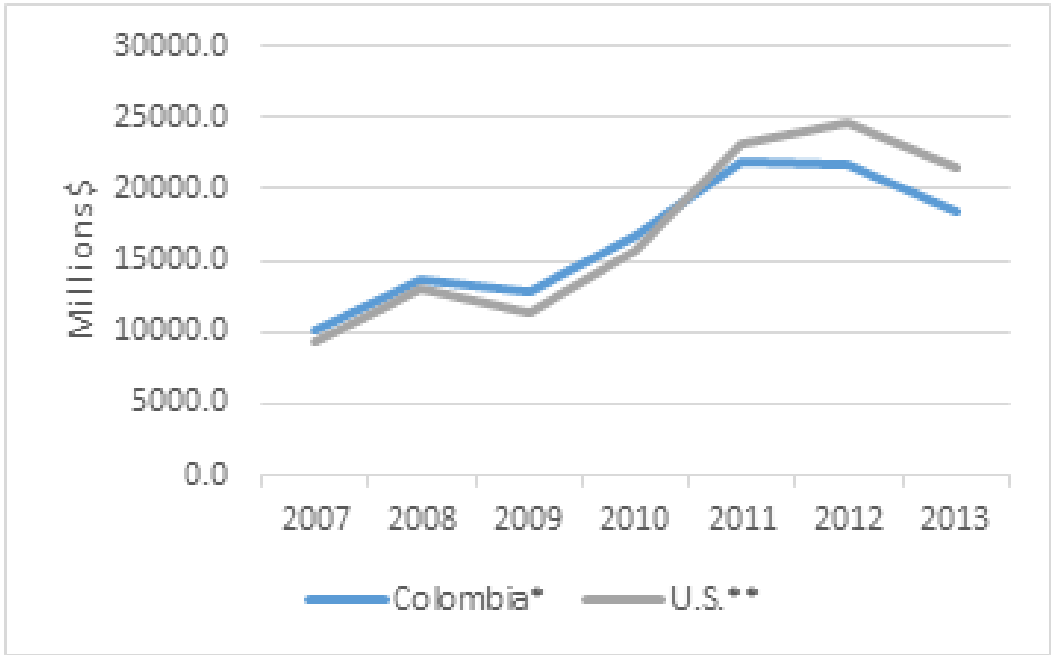
- Our objective: contribute to a small literature assessing the reliability of these micro features of the data
- Compare export records generated by shipments leaving Colombia to the import records generated by the same shipments when they enter the U.S. as imports.
  - Begin with aggregates
  - Go down to HS2 industry
  - Go down to firm to firm
  - Then down to transaction by transaction
- Consider alternative explanations for discrepancies and implications for research.
- Finish with brief discussion of scope for improvements in tracking records.

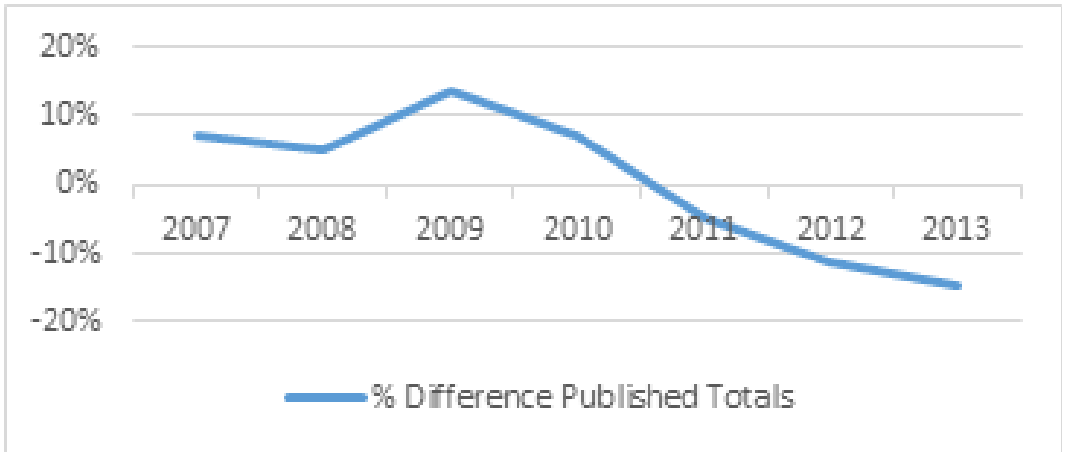
# Related literature

- **Record matching**: massive literature; Christen's (2012) book provides nice overview
- **LFTTD**: Bernard et al. (2009), Barresse et al. (2017), Kamal and Ouyang (2020), Kamal and Monarch (2017)
- **Trade reconciliation studies**: U.S. Census (1996); Orsini and dos Santos (2015); U.S. Census (2012); Fisman and Wei (2004); Mishra et al. (2008); Stoyanov (2012); Ferrantino et al. (2012); 2008 Javorcik and Narciso (2017); Kellenberg and Levinson (2019)
- **Studies using matched customs records** (very incomplete list): Eaton et al. (2008,2014); Blum et al. (2010, 2018); Bernard and Dhingra (2015); Dragusanu (2014); Kamal and Sundaram (2014); Sugita et al. (2019); Dragusanu (2014); Eaton et al (2017); Monarch and Schmidt-Eisenlohr (2017); Bernard et al. (2018a, 2018b); Carballo et al (2018); and Monarch (2019); Helper and Munasib (2021)

# Official aggregates

## Colombian Exports (FOB) to the U.S



# Are discrepancies concentrated in a few industries?

Consider 3 largest HS2 categories, by year


GRAY: ceramic products; ORANGE: apparel not knitted; BLUE: knitted apparel

# A mismatch in transactions or sales per transaction?

- Not uncommon for U.S. importers to split apart transactions for administrative purposes (U.S. Census Bureau experts)
- LFTTD-CO records imply about 8 percent more transactions than the DIAN data.
- But LFTTD-CO also reports 12 percent higher value, so more transactions *and* larger shipment sizes recorded by U.S.
- Not just a matter of shipments generating multiple records in one country but not the other.

# Interpreting the patterns

- Excess of transactions in LFTTD over transactions in DIAN can't just be a record matching problem.
- Is the discrepancy caused by entrepot trade?
  - Country of origin (CO) should differ from shipping country (CS) when entrepot trade occurs.
  - Both databases *should* be based on CO concept, but there's some room for slippage.
    - exporter reports "last known destination"; may not know true destination.
    - importer may not know country of origin.
    - lots of Colombian trade goes through Panama
    - "[T]he majority of trade–80 percent–is shipped indirectly. . ." Ganipati et al. (2021)

# A reflection of Entrepot Trade?

**Figure 4: Monthly Shipments over 2007-2013: LFTTD-CS vs. LFTTD-CO**

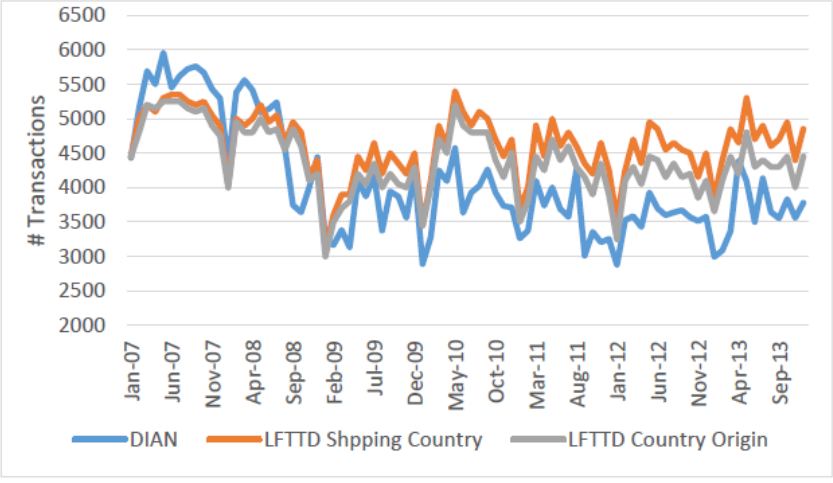
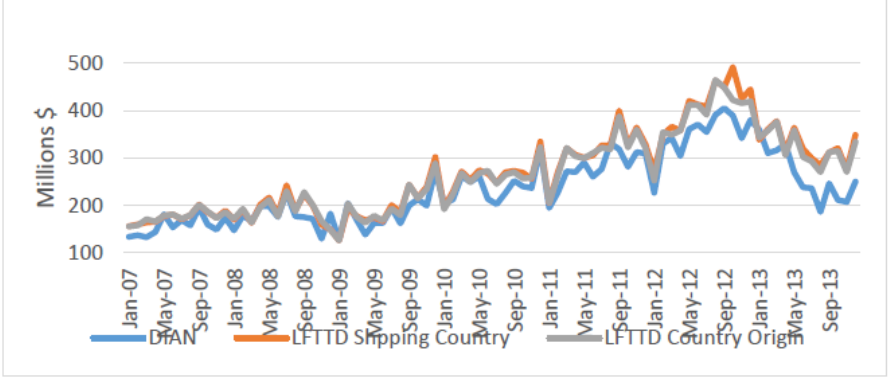



**Figure 5: Monthly F.O.B. value over 2007-2013: LFTTD-CS vs. LFTTD-CO**

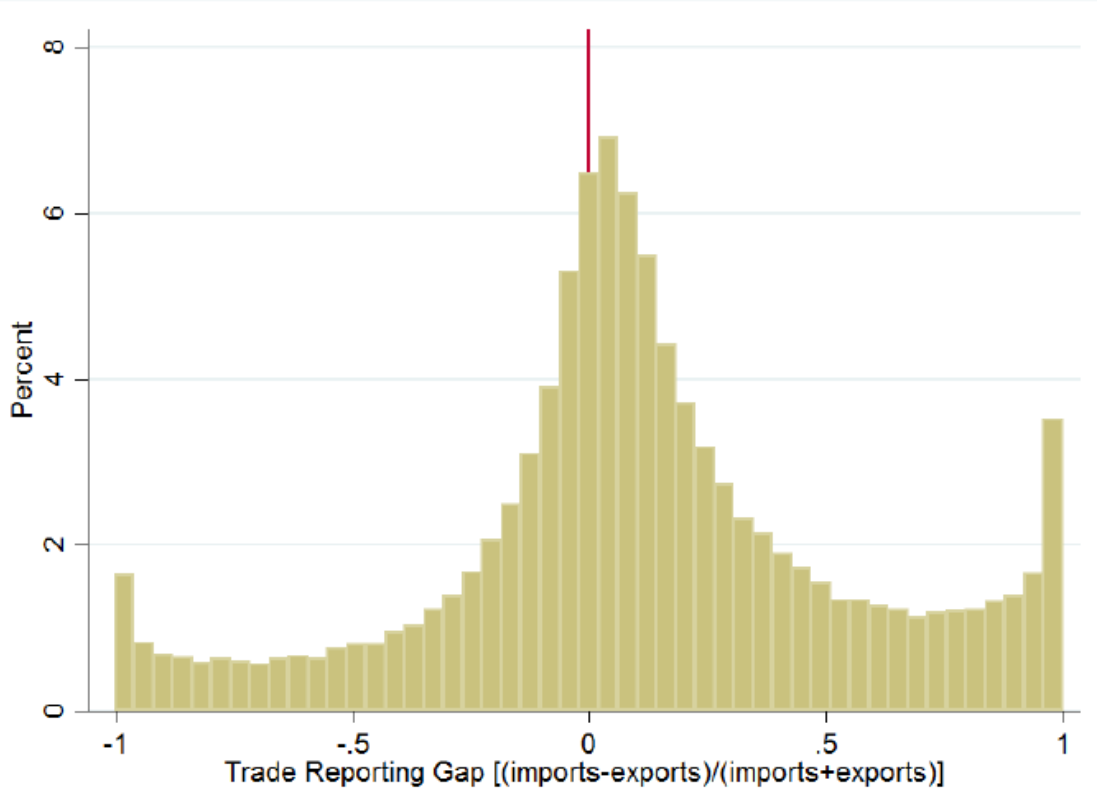



- Entrepot trade *may* be getting more important, still, it doesn't explain growing gap in U.S./Colombia aggregate trade series

details

# Are the gaps especially big?

- 4-7 percent for Australia (U.S. Census, 1996)
- 11-17 percent for Brazil (Orsini and dos Santos, 2015)
- 22-48 percent for China (U.S. Census, 2012).
- wide range in 11 year Comtrade panel,126 countries:




source: Kellenberg and Levinson, 2019

# Correlates of discrepancies

- **Tariffs** (Fisman and Wei, 2004; Mishra et al., 2008; Stoyanov, 2012; Ferrantino et al., 2012; Kellenberg and Levinson, 2019)
  - Colombia-U.S. FTA in 2012 may have reduced incentives in U.S. to understate imports.
  - But the gap didn't close after 2012, it grew.
- **Domestic profit taxes** (which create incentives to overstate the value of intermediate inputs)
- **Preferential trade agreements** (which reduce incentives for tariff avoidance when importing from partner countries)
- **Inflation** (which proxies for incentives to avoid capital controls)
- **Corruption and auditing standards** (Javorcik and Narciso, 2017; Kellenberg and Levinson, 2019).

# Now to our focus: micro examination of misreporting

- Do customs records consistently characterize firm-to-firm trading patterns?
  - Who trades with whom?
  - Within particular buyer-seller relationships, how well to shipment records match up?
- Are reporting issues concentrated among particular types of firms?
- Implications for studies that use these data.

# Matching importers and exporters

| | LFFTD (U.S.) | DIAN (Colombia) |
|---|---|---|
| **Importer identifier** | **EIN** → Name, Address (business registry) | Name, Address |
| **Exporter identifier** | **MID** (string based on Name, Address) | **NIT,** Name, Address |

- Longitudinal Firm Trade Transactions Database (**LFTTD**)
  - Employer Identification Number (**EIN**)
  - EIN can be used to retrieve importing firms' and address from Business Register
- Colombia's National Directorate of Customs and Taxes (**DIAN**)
  - includes tax identification number (**NIT**) of exporter
  - includes name and address of importer

# Identifying trading partners

| | LFFTD (U.S.) | DIAN (Colombia) |
|---|---|---|
| **Importer identifier** | **EIN** → Name, Address (business registry) | Name, Address |
| **Exporter identifier** | **MID** (string based on Name, Address) | **NIT,** Name, Address |

- Trading partner matching strategy.
  - we'll use name and address matching to mimimize loss of information.
  - will also briefly examine properties of a psuedo MID constructed from Colombian data
- Once this is done, we'll try to link the individual transactions reported by each pair of trading partners

# Two-stage linking strategy

- **Preliminary data cleaning**
  - Using code that created the LFTTD, clean records to standardize and reduce noise due to recording errors
- **First stage: firm matching** details
  - Block on zip code or state to reduce dimensionality of the pairwise comparison problem.
  - 3 rounds of exact greedy matching on names and addresses;
  - 3 rounds of fuzzy greedy matching on names and addresses
  - outcome: list of importer-exporter pairs, each agent lists the other as a trading partner within a common time window, up to noise
- **Second stage: transaction matching:** details
  - Block on exporter-importer pairs
  - Use transaction info. (HS2, shipment value, date) to link transactions across datasets.
  - outcome: for each pair of trading partners, a list of transactions they have recorded in a roughly consistent way

# Trading partner matching results

| | **LFTTD** | **DIAN** |
|---|---|---|
| No. Colombian exporters identified | 13,500 | 7,281 |
| No. U.S. importers identified | 9,400 | 18,194* |
| Matched no. importing firms | 2,500 | 2,500 |
| Matched firms, total no. transactions | 259,000 (72.9% LFTTD) | 324,707 (95% DIAN) |
| Matched firms, FOB exports ($USM) | 18,430 (84.8% LFTTD) | 19,268 (97.3% DIAN) |

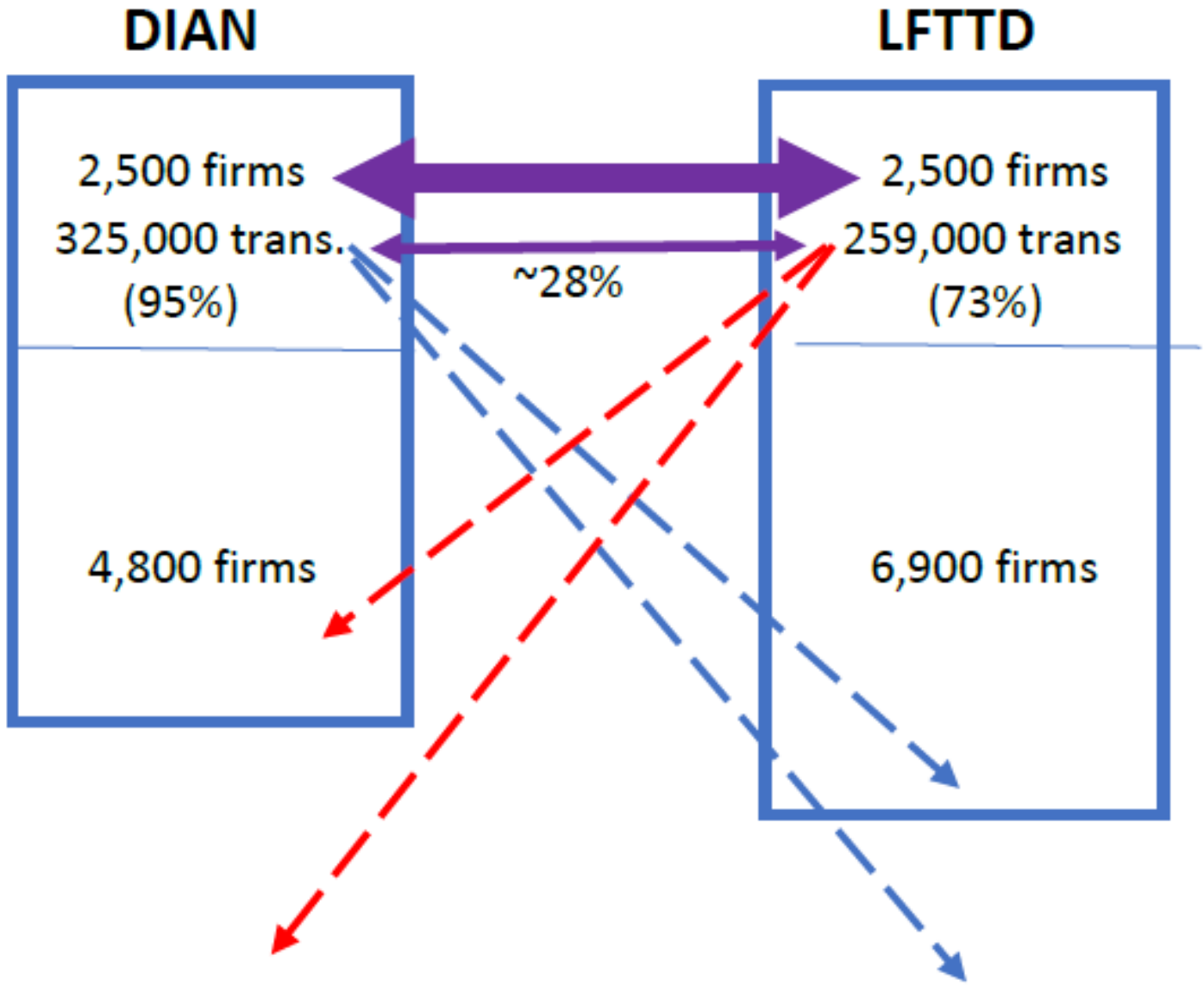
*Based on psuedo MID constructed from importer info. in DIAN.

- Matched *firms* account for most trade, but most firms not matched
- Almost twice as many exporters identified by MID than actually appear in DIAN; similar pattern using psuedo-MID on the DIAN data. (See also Kamal and Monarch, 2018)

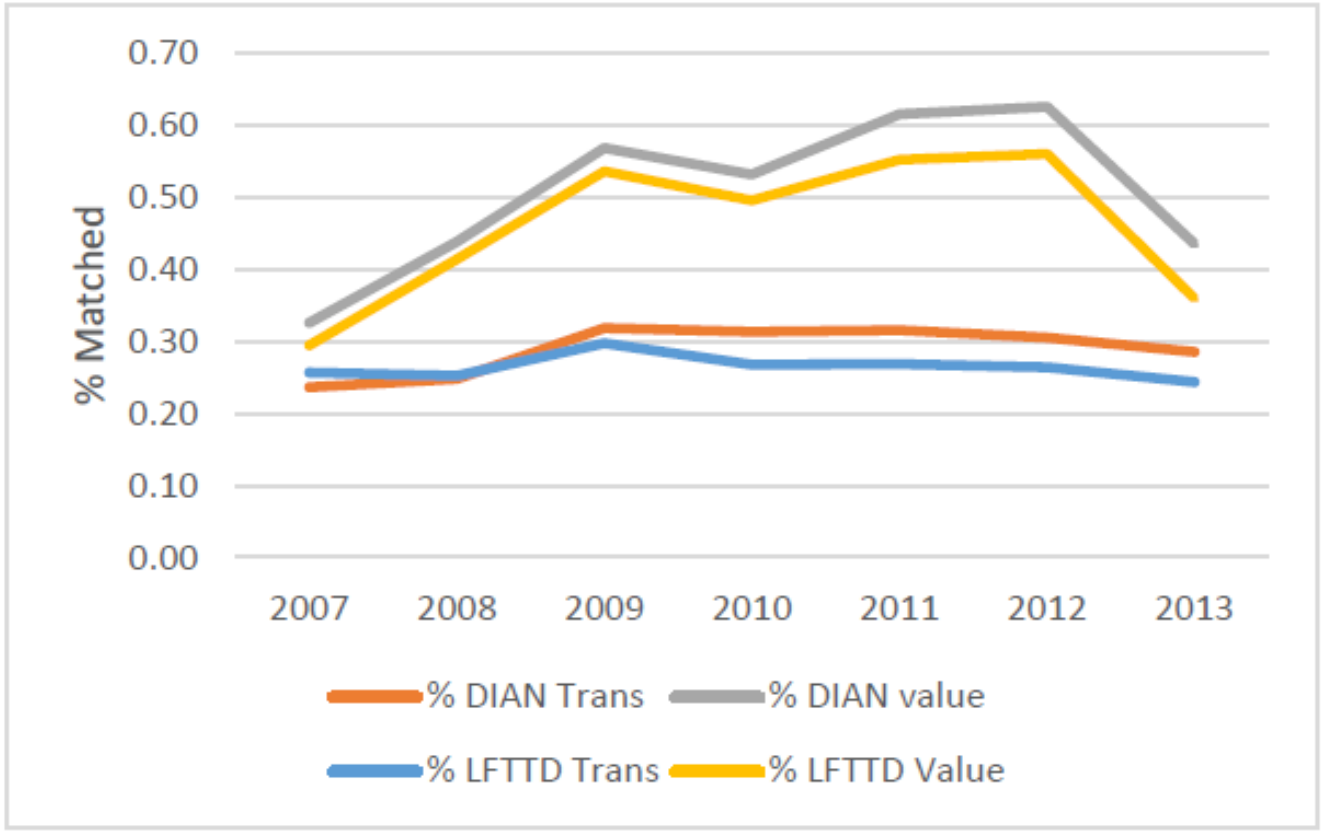
# Transaction matching results

| | **LFTTD** | | **DIAN** | |
|---|---|---|---|---|
| | level | %LFFTD | level | %DIAN |
| # trans. matched | 97,000 | 27.3% | 97,000 | 28.4% |
| FOB value, matched trans. | 10,400 | 48.0% | 10,383 | 52.4% |

- Matched transactions of matched firms account for
  - <30% total transactions,
  - ~50% total value Colombian exports to U.S..


DIAN
LFTTD
2,500 firms
325,000 trans.
(95%)
~28%
2,500 firms
259,000 trans
(73%)
4,800 firms
6,900 firms

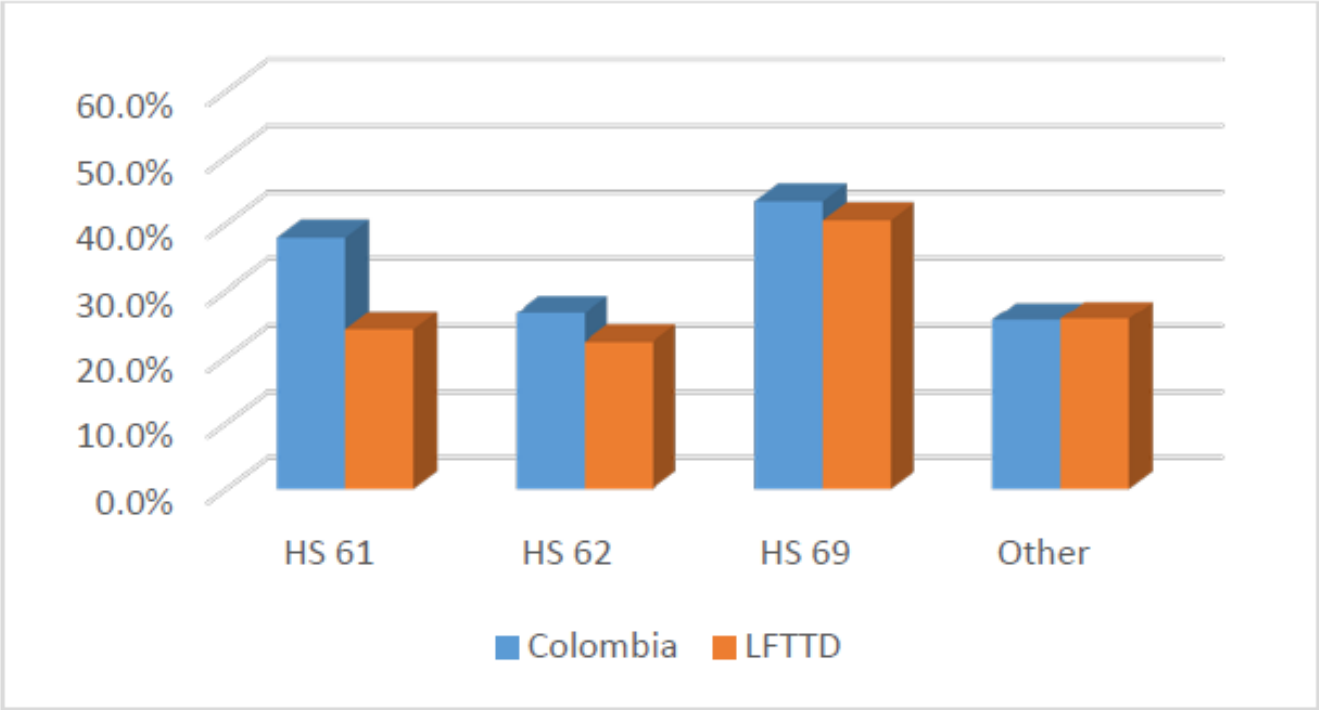
# % Transactions, % FOB value matched over time




The Great Trade Collapse appears to knock out smaller exporters, which are less likely to match.

# % Transactions matched, by industry

3 leading sectors

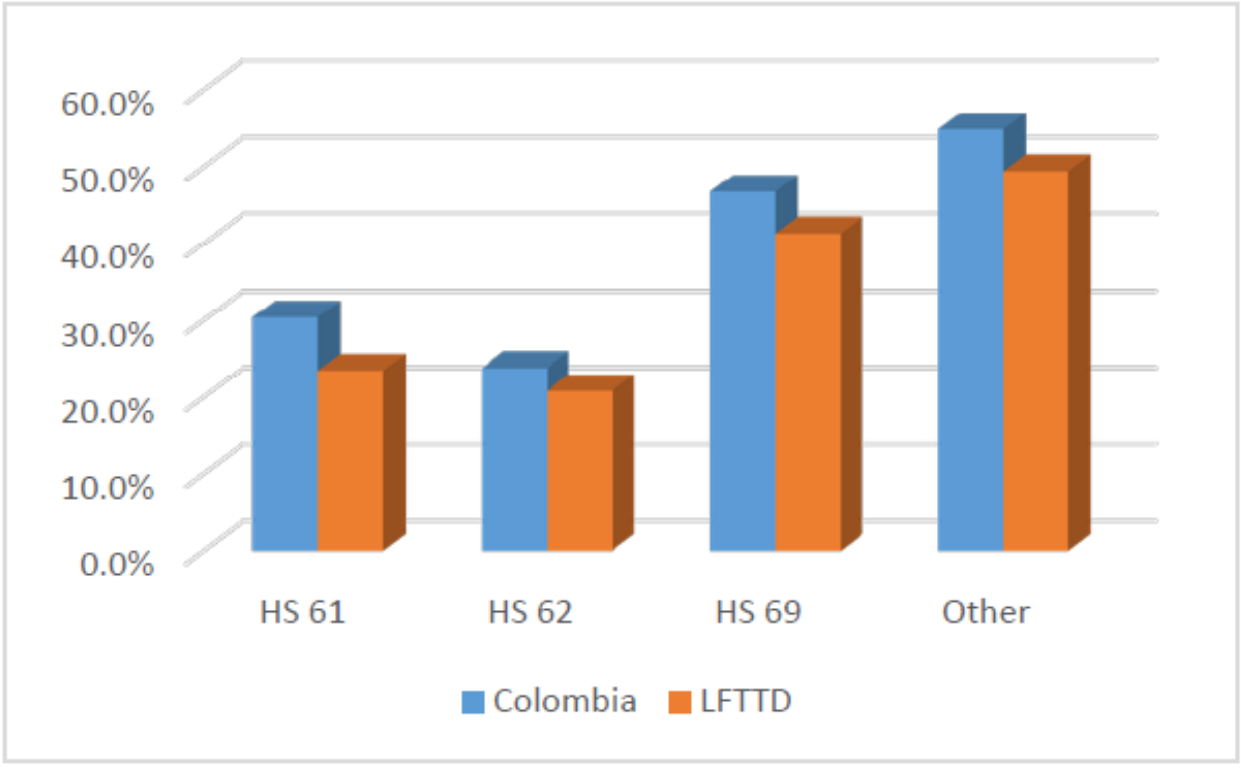



HS 61: Articles of apparel and clothing accessories, knitted or crocheted.
HS 62: Articles of apparel and clothing accessories, not knitted or crocheted.
HS 69: Ceramic products.

- Industries with relatively high turnover, small firms, appear to match less frequently

# % FOB value matched at transaction level

3 leading sectors



# Related party versus Arm's Length (LFTTD only)

| | **Related party** | | **Arm's Length** | |
|---|---|---|---|---|
| | level | %LFFTD | level | %LFTTD |
| # Trans. bet. matched firms | 81,000 | 82.2% | 178,000 | 66.9% |
| # Trans. matched | 30,500 | 30.2% | 67,700 | 25.2% |
| FOB value, matched firms | 2,930 | 80.9% | 15,500 | 83.8% |
| FOB value, matched trans. | 1,257 | 34.7% | 9,165 | 49.5% |

- Related party firms are more likely to match, but their transactions are less likely to match.
- Shades of Bernard et al. (2006) on transfer pricing?

# Match rates by shipment value

| | LFFTD | | DIAN | |
|---|---|---|---|---|
| quintile | trans. | value | trans. | value |
| 1 | 24% | 25% | 25% | 25% |
| 2 | 28% | 28% | 29% | 29% |
| 3 | 26% | 26% | 27% | 27% |
| 4 | 26% | 26% | 30% | 30% |
| 5 | 29% | 51% | 32% | 56% |

- Better match rates in largest quintile, especially in value terms.

# Wholesale/Retail vs Other Importers (LFTTD only)

| | **Non W/R trade** | | **W/R trade** | |
|---|---|---|---|---|
| | level | %Non W/R | level | %W/R |
| # trans. bet. matched firms | 86,000 | 55.1% | 173,000 | 82.0% |
| FOB value, matched firms | 13,770 | 83.4% | 4,660 | 83.2% |
| # FOB. matched trans. | 32,500 | 20.8% | 65,000 | 30.8% |

- Poor match rate for non-W/R transactions, but similar share of FOB value accounted for.

# Summary

- Discrepancies in aggregate flows trace to more Colombian shipments in LFTTD than in DIAN, and larger value per shipment.
  - possibly entrepot trade via Panama, not recorded by Colombians as destined for U.S.
  - not just a few problematic sectors.
- Aggregates hide more dramatic inconsistencies in customs records at the transactions level.
- Transactions that can be matched are accounted for by 2,500 firms–about 35% of the exporting firms in Colombia and 27% of the importing firms in the United states.
  - The firms with at least one matched record account for 97% of Colombian exports to the U.S. and 85% of U.S. imports from Colombia, by value
  - But even for these firms, most transactions cannot be matched.

# Summary, continued

- Match rates for shipments within matched buyer-seller pairs are low, especially for:
  - non-affiliated trade
  - small shipments
  - non-wholesale/retail trade

# Poor linkages mess things up

- **Reseach**
  - characterization of international B2B networks, GVC's:
    - missed links and imagined links affect network statistics, undermining inference
    - mismeasured longevity of relationships
  - analysis of exporter dynamics:
    - mismeasured entry costs, search costs, learning
  - analyses of technology transfers
    - mismeasured effect of interactions
- **Government**
  - construction of trade aggregates
  - enforcement of commercial policy
  - identification of bad actors abroad

$\Rightarrow$ Well-recognized payoff to doing better

# Better identifiers?

- **Shipment Invoice numbers**
  - Issued by the seller to the buyer, lists good, price, etc.; carries numeric code.
  - Main documentation of the sale between the two parties
  - U.S. import declaration form requires the importer to attest to its accuracy and use it to fill-out several vital fields
  - **but** required information on foreign export customs records?
    - In the U.S.: invoice # can be used to identify the "Importer of Record".
- **Bills of Lading (BOL) numbers**
  - BOL issued by the shipment carrier, carries a numeric code
  - establishes receipt of the goods, provide evidence of title to the goods' ownership.
  - **but** they can refer to multiple invoices and they may identify a container rather than a shipment
  - required information on foreign export customs records?
    - In the U.S.: BOL # can be used to identify the "Importer of Record".

# Better identifiers?

- **Private sector firm identifiers**
  - Dunne and Bradstreet
  - others?
- Global coordination on new recording norms may be infeasible.

But . . .

- Collecting data on a subsample could still be valuable
  - provides a sample of known true and false matches
  - opens the possibility of "supervised" matching algorithms that are based on "training" samples

# Entrepot Trade as an obscuring factor

- "[T]he majority of trade–80 percent–is shipped indirectly. The average shipment stops at two additional countries before its destination . . ." Ganipati et al. (2021)
- DIAN records exports according to their "last known destination."
  - If exporters don't know final destination, could bias trade statistics in either direction.
- LFTTD records list Colombia as country of origin (**CO**), shipping country (**CS**), or both. Suppose
  - **CO** = **Colombia, CS** $\neq$ **Colombia:** not a problem if properly recorded in U.S.
  - **CS** = **Colombia, CO** $\neq$ **Colombia:** *could* be a problem if such goods pass through Colombian customs and lose their country of origin identity, while importer knows CO $\neq$ Colombia.

# First stage: exact greedy matching rounds

1. Using standardized names, match on the first five words in each string.

   **example from DIAN:**

   "CEDAR BRIDGE NURSERIES DBA WORLD CLASS FLOWERS" becomes
   "CEDAR BRIDGE NURSERIES DBA WORLD"

2. If no exact match is found for a record, or if less than five words are available, it is passed to a second round based on the first four words.
3. Analogously, records that could not be exactly matched in the second round advance to a third round, which matches on only the first three words

# First stage: fuzzy matching criteria

- Definitions:
  - Let $\gamma$ be a string based on record characteristics
  - let $\mathbb{A}$ and $\mathbb{B}$ be the set of all upstream and downstream records, respectively.
  - let metric $s(\gamma_a, \gamma_b)$ measure similarity (inverse Generalized Edit Distance) between $a \in \mathbb{A}$ and $b \in \mathbb{B}$.
  - let $B \subseteq \mathbb{B}$ be the set of downstream records for which $a$ is the best match: $B = \{b' | a = \max_{a' \in \mathbb{A}} s(\gamma_{a'}, \gamma_{b'})\}$
  - let $I_a$ and $I_b$ be the domestic-record-based firm identifiers associated with records $a$ and $b$, respectively. (For us, $I_a$ is the exporting firm's NIT, and $I_b$ is the importing firm's EIN.)
- Trading partner identification
  - Then we say firms $I_a$ and $I_b$ are trading partners if (1) $b \in B$, (2) $b$ is the best match for $a$:

$$s(\gamma_a, \gamma_b) > \max_{b' \neq b} s(\gamma_a, \gamma_{b'}),$$

and (3) the similarity between $a$ and $b$, $s(\gamma_a, \gamma_b)$, exceeds a threshold value.

# First stage: fuzzy greedy matching rounds, continued

- the rounds:
  - Block on zip code, fuzzy match on name and street address
  - Block on state, fuzzy match on name and street address
  - Block on 2-digit zip code, fuzzy match on name
  - Still no match? **Give up**.

# Second stage: fuzzy matching on transactions

- Blocking on exporter-importer pairs, use transaction information to look for consistency in the transactions reported.
- Match on the similarity between the target transaction in DIAN and each of the potential counterpart transactions in LFTTD.
- Analogously with name/address matching we say records $a$ and $b$ are a match if
  - $a$ is the best match for $b$,
  - $b$ is the best match for $a$ and
  - the similarity between $a$ and $b$, $s(\gamma_a, \gamma_b)$, exceeds a threshold value.
- Now, howerer, $\gamma = [HS2, \textit{shipment date}, \textit{value}]$
- Similarly function somewhat ad hoc, but robustness tested.

**The index**: $$s' = \left(s_{HS2} + s_{date} + s_{value}\right)/3$$

where:

- 2-digit HS codes:

$$s_{HS2} = \begin{cases} |HS2_{DIAN} - HS2_{LFTTD}| = 0 & 1 \\ |HS2_{DIAN} - HS2_{LFTTD}| = 1 & 0.7 \\ |HS2_{DIAN} - HS2_{LFTTD}| = 2 & 0.5 \\ |HS2_{DIAN} - HS2_{LFTTD}| = 3 & 0.3 \\ |HS2_{DIAN} - HS2_{LFTTD}| > 3 & 0 \end{cases}$$

- export month: $s_{date} = 1 - |Month_{DIAN} - Month_{LFTTD}|/12$
- and F.O.B. values: $s_{value} = \frac{|Value_{DIAN} - Value_{LFTTD}|}{Value_{DIAN} + Value_{LFTTD}}$