Machine Learning for Missing Data on Illicit Trade

Alice Lépissier, PhD
Brown University

7th January 2024
Machine Learning in International Trade and Finance
ASSA 2024

WATSON INSTITUTE
INTERNATIONAL & PUBLIC AFFAIRS
BROWN UNIVERSITY

THE WILLIAM R.
RHODES CENTER
International Economics & Finance

Agenda

- A primer on illicit trade
- Existing estimates of illicit trade
- The problem of missing data
- Using machine learning for data imputation

Illicit financial flows (IFFs)

- Illicit financial flows refers to the movement of money across borders that is illegally earned, transferred, or utilized (Baker, 2005).
- Trade misinvoicing is a subset of IFFs which involves misreporting trade invoices to clandestinely shift money abroad or repatriate money domestically.
- Diverse motivations:
  - laundering proceeds of crime
  - financing of terrorism
  - corruption
  - tax evasion by shifting profits to lower-tax jurisdictions (multinationals) or hiding transfers of wealth (individuals)
  - market and regulatory abuse

Trade misinvoicing 101

1. Manipulate shipment invoices according to motivations below.
2. Present doctored invoices to customs authorities.
3. Optional: bribe customs officials, collude with a foreign partner.

	Imports	Exports
Outflow	**Over-invoicing** Disguise illicit capital flight Evade taxes Avoid capital controls	**Under-invoicing** Disguise illicit capital flight Evade taxes Avoid capital controls
Inflow	**Under-invoicing** Repatriate undeclared capital Launder money Evade tariffs	**Over-invoicing** Repatriate undeclared capital Launder money Exploit export subsidies

Consequences of trade misinvoicing

- Depletes state coffers
- Erodes tax base
- Undermines state institutions
- Weakens governance

Inflows are just as corrosive as outflows

- Used to finance illicit sectors of the economy
- Untaxed and invisible to governments

Existing estimates of illicit trade

Estimating illicit trade from bilateral trade statistics

- Illicit financial flows are unobservable: they are deliberately hidden, so they will not be systematically recorded.
- What can we learn about illicit trade from observed bilateral trade flows?
- Leverage macroeconomic identity that imports = mirror exports for given country dyad, year, and commodity.
- Exploit the "double entry" accounting system of international trade statistics, where a transaction is recorded twice.

The atlas of illicit financial flows from trade misinvoicing

Paper: https://dx.doi.org/10.2139/ssrn.3984323

Data: https://doi.org/10.5281/zenodo.3610557

Code: https://github.com/walice/Trade-IFF

The "atlas" of trade mis-invoicing

Scrape entire UN Comtrade database of bilateral trade flows for all 99 commodity sectors, $n \approx 23$ million.

Locates trade misinvoicing in the discrepancies in mirror trade statistics, e.g., difference between reporter imports and mirror partner exports (and vice versa).

… A bunch of statistical adjustments later …

The "atlas" database

- ✓ Bilateral estimates of misinvoiced trade
- ✓ 167 countries during 2000-2018
- ✓ Disaggregated by commodity and year
- ✓ Publicly available

Machine learning for missing data on illicit trade

Paper: https://alicelepissier.com/docs/illicitAI.pdf

Code: https://github.com/walice/illicitAI

The problem

- The “atlas” database is missing data on 10 African countries that do not provide customs declarations to UN Comtrade.
- Weak administrative systems for statistical reporting in developing countries leads to issues with data quality and availability.
- Data will not be missing at random.

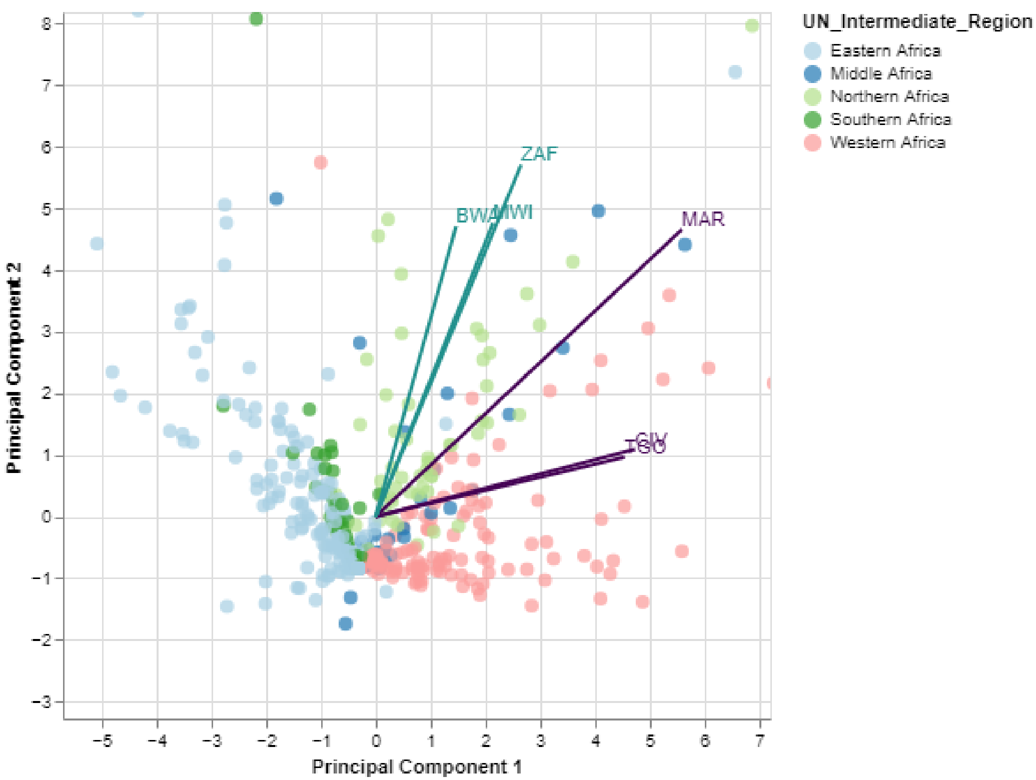
We need a strategy to impute missing data on illicit trade without relying on the underlying bilateral trade transaction.

Research question

- Recent innovations use machine learning on satellite imagery to predict measures of economic well-being in developing countries.
- Machine learning on transaction-level data collected by financial institutions can identify risky financial transactions.
- But these approaches rely on high-resolution data that is passively collected (e.g., nightlights), or that has clearly labeled outcomes (e.g., “fraud”, “not fraud”).
- Will machine learning work to detect illicit financial flows in aggregate economic and political data?

How well do machine learning models trained on information about country-level characteristics predict bilateral flows of misinvoiced trade for African countries?

Unsupervised machine learning recovers regional clusters of illicit trade



The Principal Components Algorithm is blind to what the regions were.

Training the models

Dimensions of analysis

❶ **Gravity models of international trade** ➞ Push-pull gravitational factors of regular trade flows

❷ **Trade-based money laundering** ➞ The "illicit premium" refers to the attractiveness of destination countries for illicit business, which requires some corruption (but not too much), and political and macroeconomic stability

❸ **Policy incentives for trade misinvoicing** ➞ Economic policies such as tariffs or capital controls create incentives for market and regulatory abuse

Predicting illicit trade from country pairs

- What happens if official trade statistics are missing? Machine learning models trained without data on observed trade flows are able to recover 70% of variation in illicit trade outcomes in Africa.
- Variables used to train the models are either unilateral (e.g., perceptions of corruption) or bilateral characteristics (e.g., distance between countries).
  - ✓ Gravity variables: GDP, geographical & cultural distance, etc.
  - ✓ Financial integrity: cooperation on anti-money laundering, etc.
  - ✓ Governance: corruption, rule of law, etc.
  - ✓ Regulatory environment: tariffs, capital controls, etc.

The specific combinations of country characteristics capture some underlying structure that is highly predictive of illicit trade.

Features used to train the models

Gravity variables
GDP, population, geographical distance, cultural distance, barriers to trade

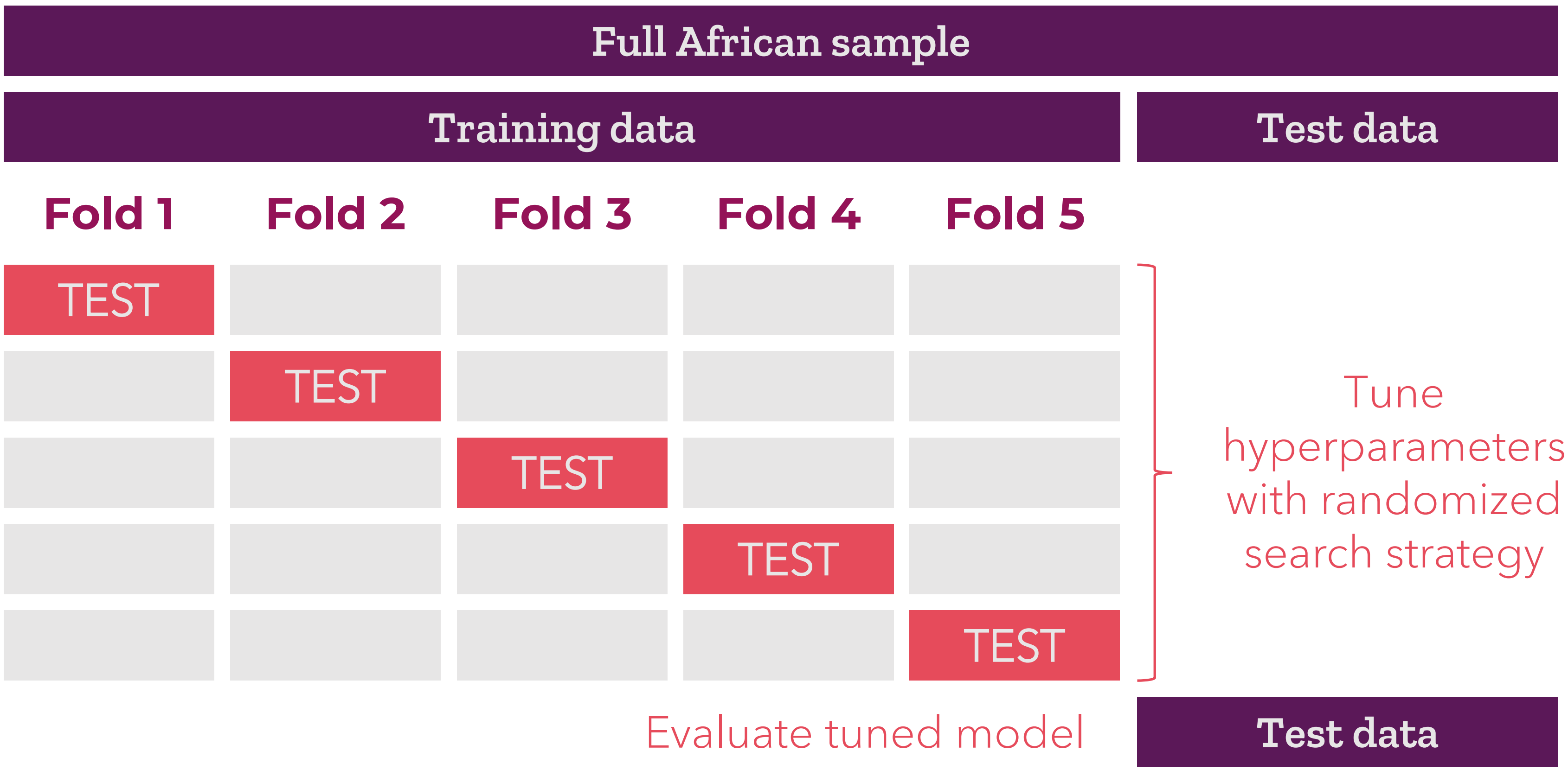
Financial integrity variables
Secrecy score & rank on Financial Secrecy Index, promotion of tax evasion, AML laws, cooperation on AML judicial matters

Governance variables
Corruption, quality of private sector regulations, rule of law

Regulatory environment
Tariffs, capital controls, controls on commercial trade and direct investment

Total of 42 predictors from publicly available databases

Tuning & training the models



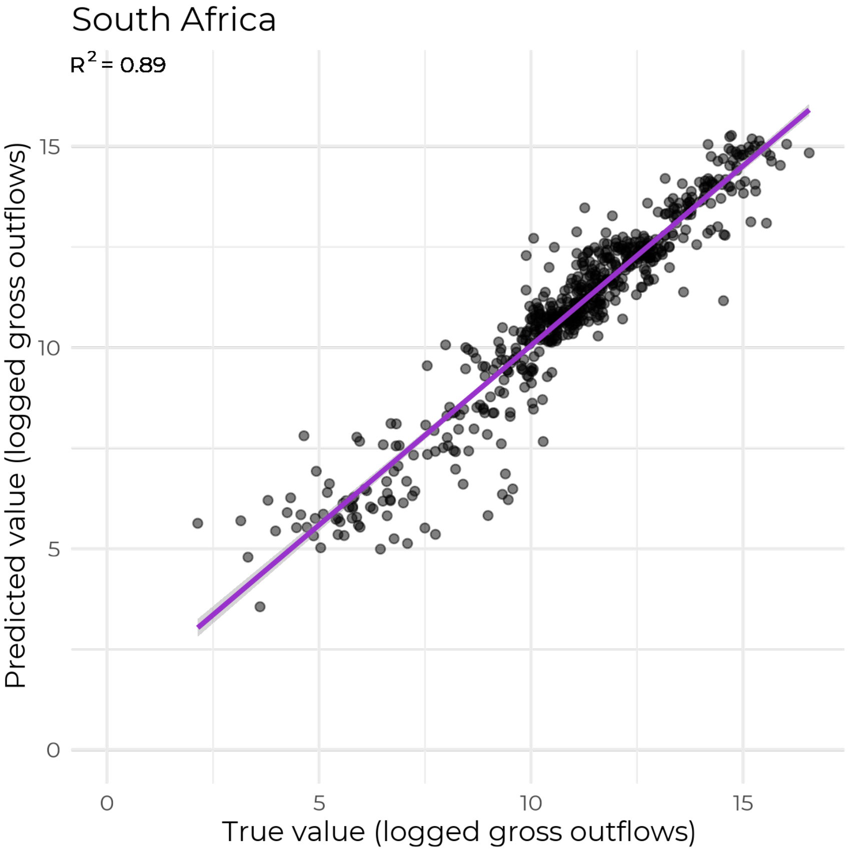
Performance of the models

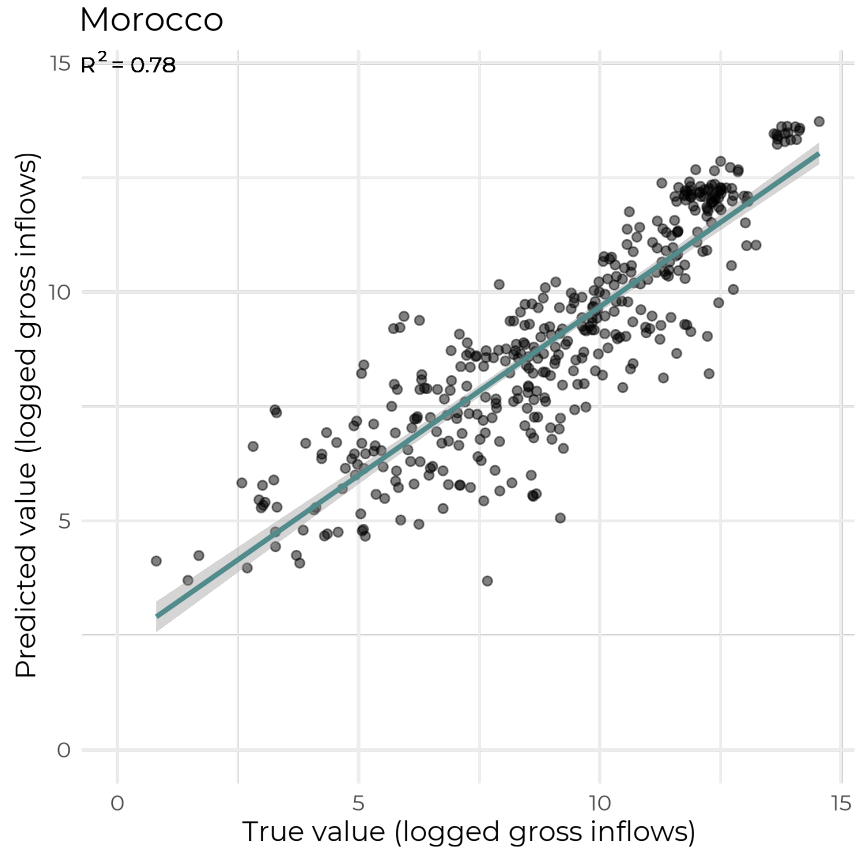
Predictive performance

- The machine learning models can reliably recover the variation in illicit trade outcomes.
- Cross-validated scores are estimates of the generalization performance of the tuned models in the population.
- Scores on an unseen test set which has not been used for model selection are used to evaluate the final performance.

	R-squared		Mean Square Error (MSE)	
	Outflows	Inflows	Outflows	Inflows
Cross-validated	68%	70%	3.23	3.04
On unseen test set	71%	73%	3.00	2.87

Cross-validated predictions

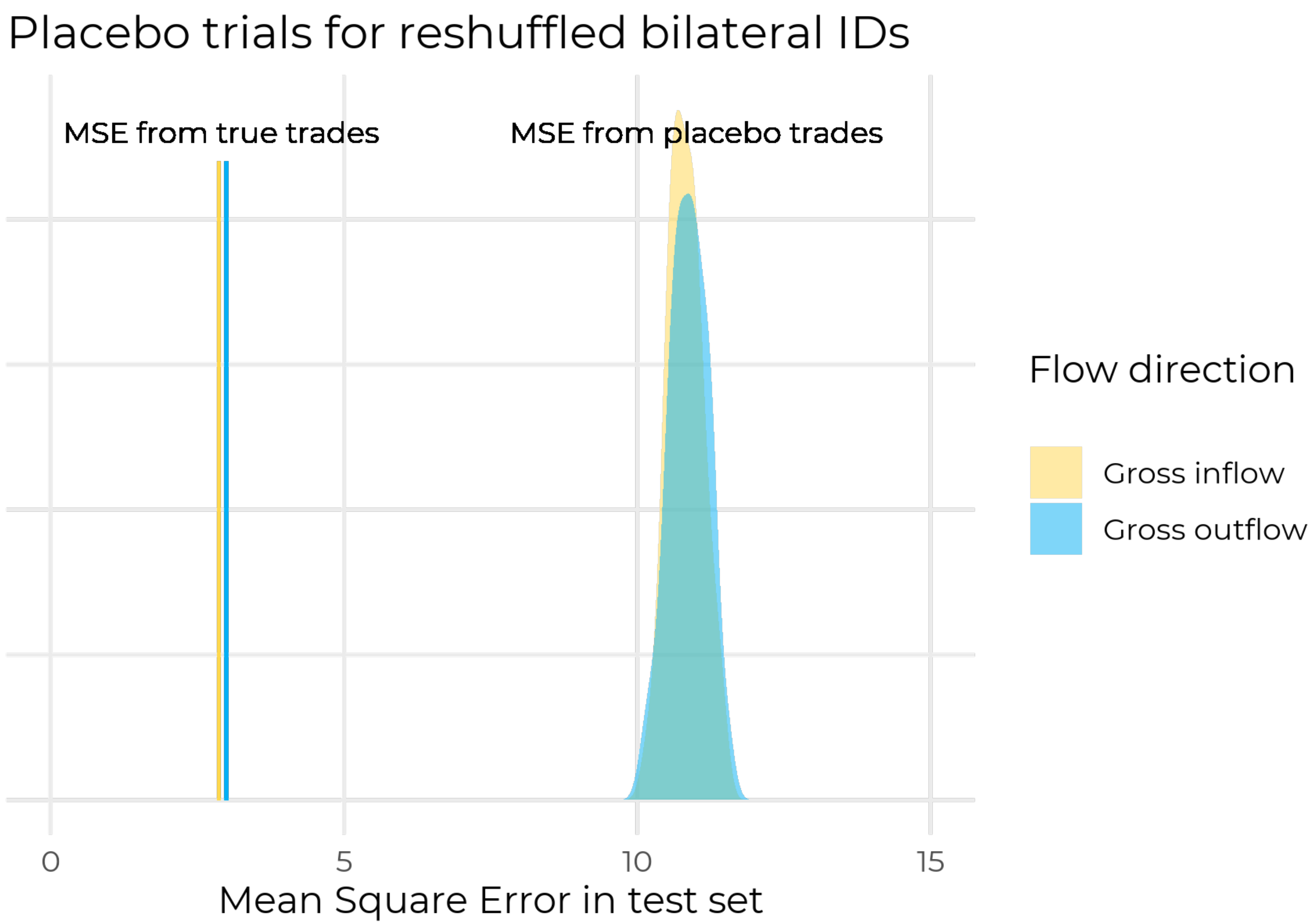




Randomization inference

- Conduct an experiment to test whether the results are the product of chance.
- Randomly reshuffle the identities of the bilateral trades and re-train the machine learning models on the fake illicit trades.
- Repeat this experiment for 100 trials and compare the accuracy of the models trained with the correct transactions to the models trained with the reshuffled transactions.
- If the results are due to chance, should expect to see the Mean Squared Error (MSE) of the models trained with the real illicit trades appear within the distribution of the MSE of the placebo models.

Inference with placebo experiment



Generalization across borders

- Does the model "travel" across country borders?
- Group countries by income level and use these samples as new test sets to evaluate the models.
- Tests of increasing difficulty because LMIC sample includes some African countries, whereas evaluating models on HIC set will indicate the extent to which models trained on different countries can be expected to generalize to new countries.

	Low & lower-middle income countries		High income countries	
	Outflows	**Inflows**	**Outflows**	**Inflows**
Cross-validated R^2	38%	38%	61%	59%
R^2 on country group sample	60%	56%	54%	42%

Robustness check – linear regression

- Could a simpler linear regression model have performed better?
- No. The superior performance of the Random Forest model (a more flexible predictor) suggests that the covariates interact in highly complex and non-linear ways to predict illicit trade.

	R-squared on test set	
	Outflows	Inflows
Linear model (reduced)	44%	39%
Linear model (full sample)	58%	57%
Random forest model	71%	73%

Contributions & applications

- Contributes to broader literature that uses creative quantitative approaches to estimate economic outcomes.
- Demonstrates that machine learning models can also reliably be trained using country-level data.
- Uses publicly available data and off-the-shelf machine learning algorithms.

Application for Democratic Republic of Congo

✓ Since DRC does not report trade data, it also won't report the identity of its trading partners.

✓ Use Comtrade to find the mirror declarations of trades with DRC: this yields the dyads that DRC is a part of.

✓ Collect the unilateral and bilateral features of those dyads and use them as out-of-sample set to generate predictions for missing data using the tuned models.

Limitations

- The "atlas" measure is taken as ground truth, so any conclusions about the accuracy of the machine learning algorithms will be conclusions about the atlas model and not about the unobservable illicit trade. If the atlas model is a poor emulation of nature, then the conclusions may be wrong (Breiman, 2001).
- The features used to train the model still need to be compiled, and some variables like Gross Domestic Product may also suffer (to a lesser extent) from data scarcity in poor countries.
- Exercise caution when using this technique for unit-level imputation. A more prudent strategy is to use the method to fill in the bilateral gaps, and then to aggregate the predictions over partners or years.

Thank you! Questions?

Get in touch

- https://alicelepissier.com
- alice_lepissier@brown.edu
- https://www.linkedin.com/in/alicelepissier
- https://github.com/walice

Glad you asked...

Inferential framework for the "atlas"

The paper constructs a measure of the dollar value of illicit trade that is embedded in each bilateral trade transaction.

Estimand

Population-level amount of trade misinvoicing conditional on observed bilateral trade flows

Estimator

The "atlas" method

Estimates

The "atlas" database provides bilateral estimates for 167 countries during 2000-2018, disaggregated by year and sector.

Inferential framework for missing data

The paper accomplishes a predictive task to augment the “atlas” database of illicit trade when trade data is missing.

Estimand

Taking the “atlas” measure as ground truth, the population-level quantity of illicit trade conditional on country-level features

Estimator

Random forest algorithm

Preview of estimates

The tuned machine learning models recover up to 70% of the variation in illicit trade outcomes in an unseen test set.

Key features of "atlas" methodology

imports = mirror exports + latent factors + noise

- Distinguishes between non-illicit and illicit reasons for trade gaps.
- Econometrically estimates transport costs and adjusts for other benign reasons for discrepancies:
  - Delays in the arrival of shipments
  - Asymmetric recording of re-exports
- Harmonization approach: accounts for the quality of country declarations to generate a reconciled value, a weighted average of reporter and partner declarations.
- Residual approach: strips reported imports of non-illicit predictors of gaps and takes difference with reconciled value.
- The remainder is the misinvoiced part of the trade.

Correlation matrix of continuous features