Machine learning for missing data on illicit trade

Paper: Please email me for the latest draft

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?

Inferential framework

Chapter 3 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.

Theory-guided variable selection

- Extract theoretical insights from three relevant literatures:
  - Gravity models of international trade
  - Trade-based money laundering
  - Policy incentives for trade misinvoicing
- Select theoretically important variables along different dimensions:
  - Push-pull gravitational factors of regular trade flows
  - 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
  - Economic policies such as tariffs or capital controls create incentives for market and regulatory abuse
- All variables are either unilateral (e.g., perceptions of corruption) or bilateral characteristics (e.g., distance between countries).

Features to train the models

Gravity variables

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

Governance variables

Corruption, quality of private sector regulations, rule of law

Financial integrity variables

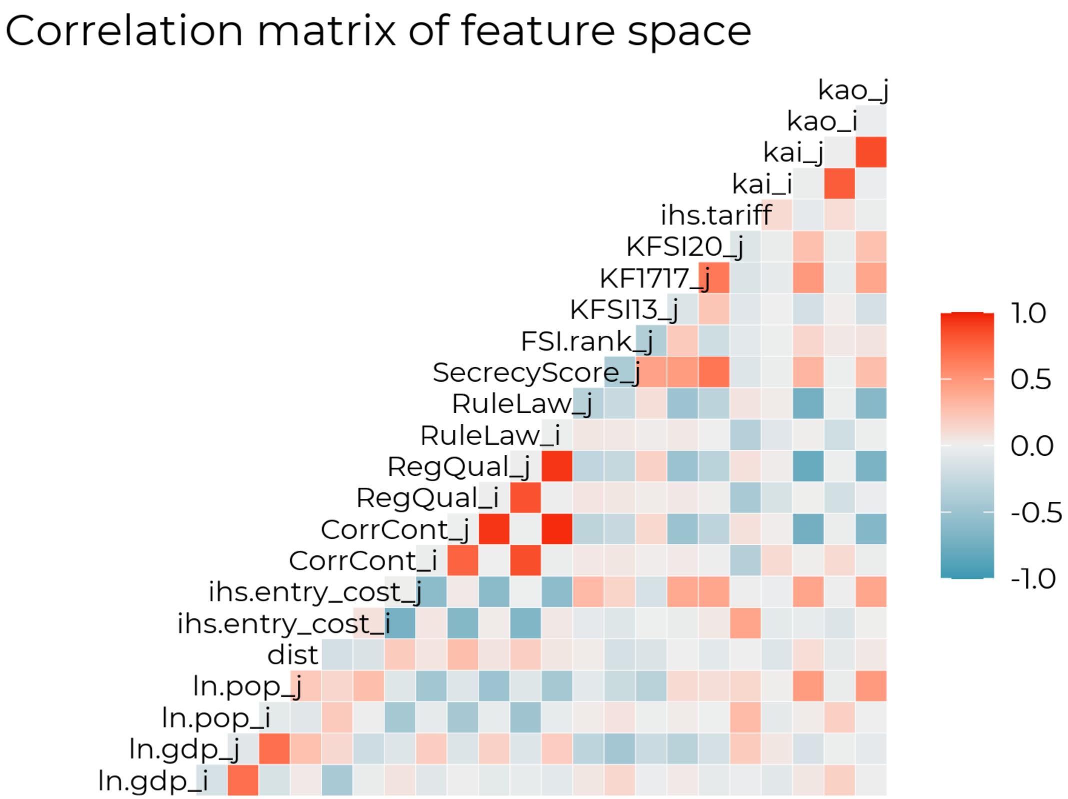
Secrecy score & rank on Financial Secrecy Index, promotion of tax evasion, AML laws, cooperation on AML judicial matters

Regulatory environment

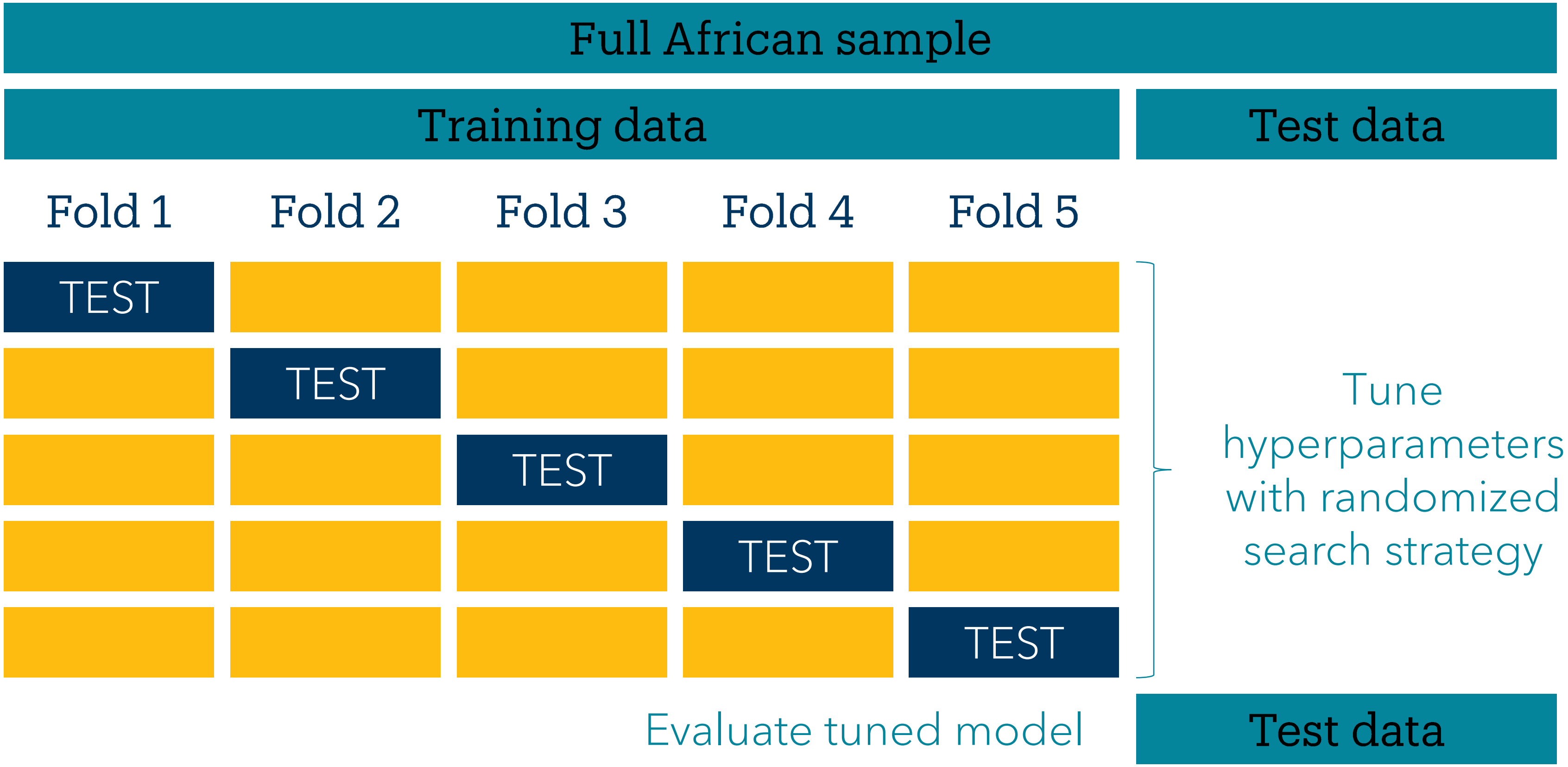
Tariffs, capital controls, controls on commercial trade and direct investment

Total of 42 predictors from publicly available databases

Correlation matrix of continuous features



Tuning & training the models



Full African sample

Training data

Test data

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
TEST				
	TEST			
		TEST		
			TEST	
				TEST

Tune hyperparameters with randomized search strategy

Evaluate tuned model

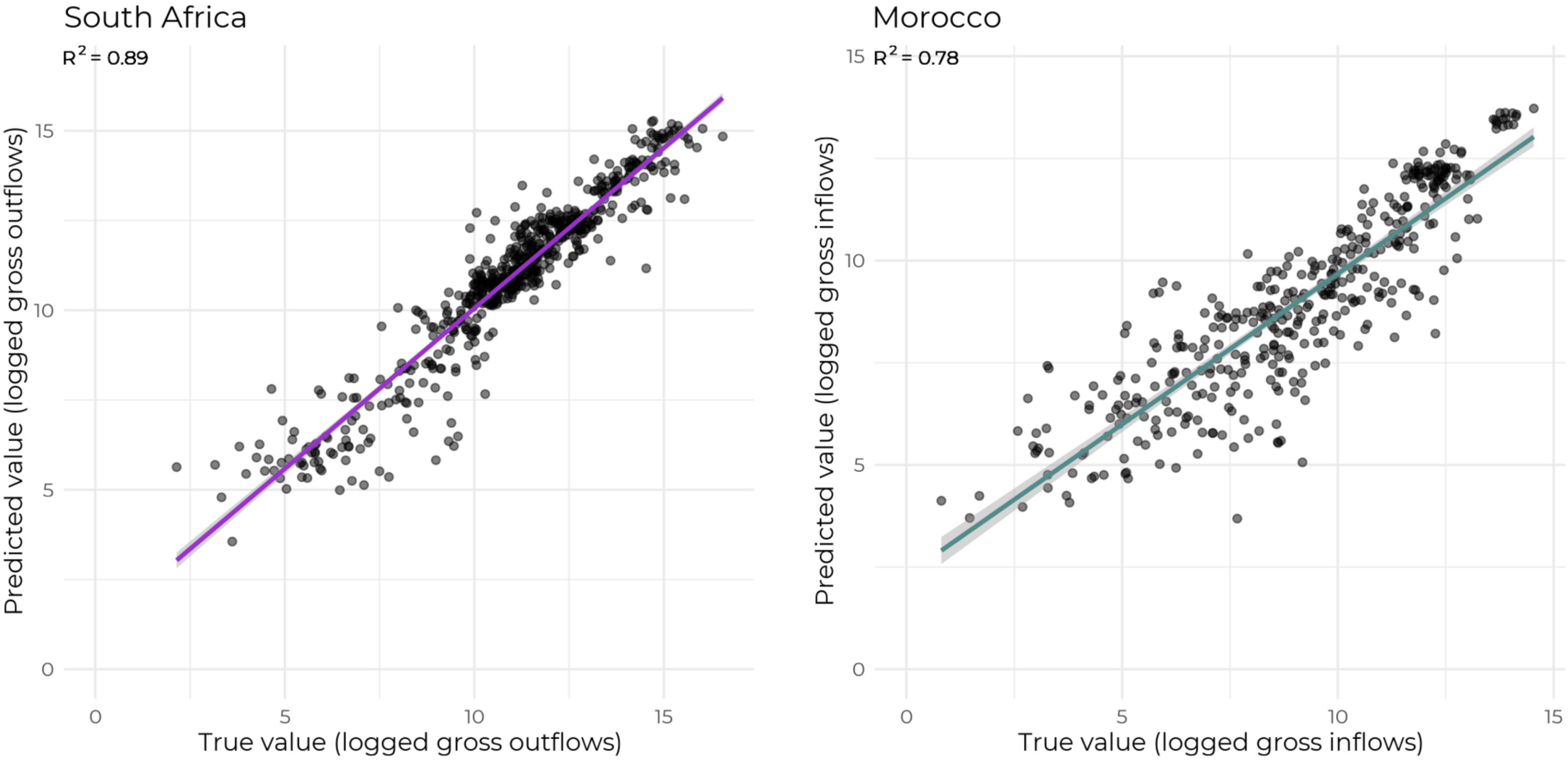
Test data

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

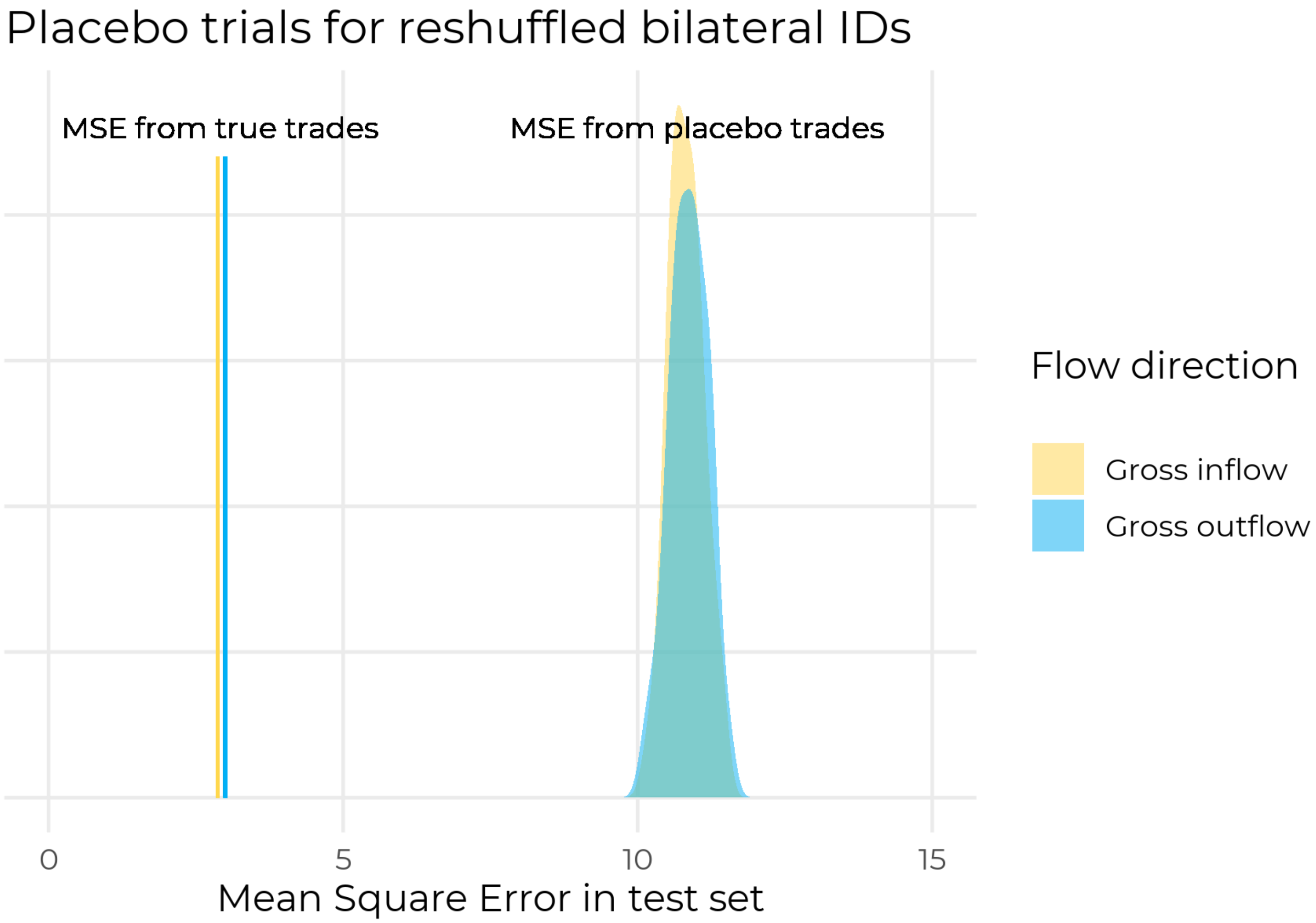
Cross-validated predictions



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 and & 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)	58%	57%
Random forest model	71%	73%

Conclusions & policy 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 to impute data 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.