Machine Learning for Missing Data on Illicit Trade

Alice Lépissier*

Brown University, Providence RI, USA

March 25, 2023
Latest version

Abstract

Trade misinvoicing is an illicit practice designed to move money in and out of a country through the deliberate falsification of customs invoices at import and export. Illicit trade hampers development by diluting public revenues, undermining tax authorities, weakening governance, and eroding state institutions. Yet, progress towards combating these illicit financial flows (IFF) is hampered by a lack of credible measurement given that IFFs concern behavior which, by definition, seeks to remain hidden. Extant dollar estimates of illicit trade are derived from bilateral trade statistics reported by countries. However, some low-income countries do not report customs data, and it is implausible to assume that these data are missing completely at random (MCAR). In this article, I present a machine learning approach to ameliorate the problem of missing data from developing countries, where administrative systems for data collection tend to be weaker, and which do not collect data on customs declarations. I predict illicit trade using machine learning models that are trained on readily available data without relying on official trade statistics. Findings show that the models are able to recover 70% of the variation in illicit trade outcomes. I show how machine learning approaches can be used for data imputation in order to yield robust inferences in real-world situations of data scarcity. This demonstrates the promise of predictive approaches to augment existing measures of illicit finance in data-constrained settings.

*Postdoctoral Fellow, Brown University. Correspondence to: alice_lepissier@brown.edu.

Contents

1 Introduction

The deployment of machine learning techniques in the field of illicit finance has enabled significant progress in the detection of fraudulent transactions using financial data, but it remains to be seen whether machine learning can produce similar advances when analyzing other types of illicit financial flows problems. This paper investigates whether machine learning approaches can also be fruitfully applied to the analysis of illicit financial flows that usually requires government data that is hard to collect and not always available. The success of machine learning in fraud applications stems from the fact that financial data is typically abundant, high-resolution, transaction-level data that is passively collected by financial institutions in the course of their usual operations. However, other important dimensions of illicit finance are captured by aggregate economic data such as bilateral trade statistics used to measure trade misinvoicing – the illicit practice used to shift money in and out of a country by deliberately manipulating the trade invoices presented to customs authorities. This type of data needs to be actively compiled and reported by government authorities and is often scarce in data-poor African countries (Devarajan, 2013; Jerven, 2009; Sandefur & Glassman, 2015). How well do machine learning models trained on more readily available information about country-level characteristics predict bilateral flows of misinvoiced trade for African countries?

Reliable methods to address missing data issues are needed in the study of illicit financial flows (IFFs); a field characterized by the difficulty of measuring the problem given that these flows are, by definition, deliberately hidden. Efforts to quantify trade misinvoicing have thus relied on official trade statistics to detect instances of illicit activity, but this data is not always systematically recorded by national customs authorities. The prejudice of missing data is compounded for developing countries, where data on economic activity is more scarce (Beegle et al., 2016; Devarajan, 2013; Jerven, 2013; Paige et al., 2020), and who are particularly afflicted by the harmful consequences of illicit financial flows (Reuter, 2012; UNCTAD, 2020; UNECA, 2017). Here I demonstrate a machine learning approach for predicting outcomes on IFFs that is reliably accurate and does not rely on data that has to be compiled by governmental agencies, and consequently mitigates the adverse effects of missing data from developing countries, with an application to Africa. The method predicts bilateral illicit trade flows from African countries without relying on trade data from customs declarations, and instead leverages more readily available data such as information on distance between countries. This paper contributes to the field of illicit finance by adding a reliable tool in the technical repertoire of IFF analyses. In addition, the paper advances scholarship on the use of machine learning in the social sciences by demonstrating a novel application of machine learning to illicit finance using aggregate country-level economic data.

This paper uses the Random Forest algorithm to predict the "atlas" measure of trade misinvoicing; the dollar value of misinvoiced trade that is embedded in a bilateral transaction for any given country pair in any given year. The predictor variables are country-level features that denote either bilateral (e.g., the existence of a trade agreement between the partners) or unilateral (e.g., population size) characteristics. These variables are either directly observable, such as whether countries share a common language, or they are proxy measures of an underlying political or economic phenomenon, such as perceptions of corruption, that originate from publicly available databases that have wide country coverage. The models that are trained in this paper are based on the Random Forests algorithm. The model hyperparameters are tuned using a randomized search strategy using 5-fold cross-validation. The predictive accuracy of the models is evaluated with R^2 and the Mean Square Error (MSE). Predictions are generated using cross-validation to guard against an overly optimistic assessment of the models' ability to predict the outcome in new, unseen data.

Results show that machine learning models trained on readily available country-level characteristics explain up to 73% of the variance in the dollar amount of misreported trade in Africa. Variables related to gravitational push-pull factors, the quality of governance in a country, the integrity of its financial system, and macroeconomic regulations have high predictive power for illicit trade outcomes. These models were trained using publicy available data and off-the-shelf machine learning algorithms that do not require significant modifications to their computational architecture. The results imply that the method presented here can be used to supplement existing measures of trade misinvoicing and can add to the evidence base on illicit financial flows. The method proposed by this paper does not rely on government-compiled data; it is economical and it is straightforward to extend. Therefore, machine learning approaches show considerable promise to mitigate missing data problems in the study of illicit financial flows, suggesting broader applications across complex policy problems that are resistant to quantification.

The rest of this paper proceeds as follows. Section 2 motivates the contributions of this paper to scholarship on IFFs and its practical policy implications by outlining the major areas of difficulties in the field and showing how the method presented here overcomes some common problems. Section 2.1 presents the outcome measure of illicit trade that is the object of the paper, discusses the nature of the missingness of the data, and the reasons why the proposed method can ameliorate some aspects of this problem. Section 2.2 locates the paper within the literature on prediction policy problems to argue that illicit finance as a field requires solving many tasks that are predictive in nature, and as such is poised to benefit from the inferential framework of machine learning. Then, section 2.3 reviews the existing uses of machine learning methods to the study of illicit finance and shows that this paper provides a novel application of machine learning to economic rather than financial data. After presenting the goals and intended contributions of this paper, section 3 follows best practices in the literature and uses theoretical insights from the literature on trade misinvoicing to identify a set of predictor variables that will be used in the analysis. Three relevant literatures are identified and critically reviewed: the economic literature on gravity models of international trade (section 3.1), studies of trade-based money laundering (section 3.2), and analyses of the determinants of trade misinvoicing (section 3.3). The rest of the paper presents the data, methodology, and findings. Section 4 describes the outcome and predictor variables in more detail. Section 5 introduces the Random Forest algorithm employed in the paper and presents the approach used to tune, train, and validate the models. Findings on the performance of these models and robustness checks are reported in section 6. Section 7 discusses potential applications of the method and its limitations, and section 8 concludes.

2 The application of machine learning methods to illicit finance

2.1 Predicting the "atlas" measure of trade misinvoicing

2.1.1 Definining trade misinvoicing

A common working definition of illicit financial flows is that they are cross-border flows that are deliberately hidden in order to obscure the illicit nature of their origin (e.g., proceeds from criminal activities, theft of state assets, etc.) or the illicit nature of the transaction (e.g., abusive transfer pricing by multinational companies, hiding wealth in offshore tax havens, etc.) (Baker, 2005; Cobham & Janský, 2020; High Level Panel on Illicit Financial Flows from Africa, 2015; Reuter, 2012). Trade misinvoicing is the faking or manipulation of invoices presented to customs for the purpose of illicitly moving money. Trade-based money laundering (TBML) is a subset of trade misinvoicing and is used to "wash" dirty money by co-mingling it with legitimate trade flows so that it can be used in the legal marketplace. TBML can be used to launder the proceeds of criminal activities such as drug or human trafficking (UNODC, 2011), illegal logging or fishing (Nelleman & INTERPOL Environmental Crime Programme, 2012; G. Rose, 2014), and grand corruption (Findley et al., 2020; van der Does de Willebois et al., 2011). Combating TBML is also an integral part of the post-9/11 international security architecture put in place to track and dismantle the financing of terrorism (FATF, 2019; Morse, 2019).

Another subset of trade misinvoicing is related to transactions that originate from legal commercial practices but are then purposely distorted or hidden in order to evade taxes on those capital flows. The most common practice in that category is abusive transfer pricing by multinational companies (MNC) to shift corporate profits to lower tax jurisdictions in order to abate their tax bill (Clausing, 2003; Davies et al., 2018). Much of international trade today is carried out by multinational corporations that have subsidiaries in several countries. According to OECD rules, subsidiaries of the same MNC should buy and sell goods to each other at the prevailing market price as if they were unrelated parties (according to the "arm's length principle").[1] In practice, there is little oversight and guidance on how to proceed when benchmark prices are not readily available,[2] and so transfer mispricing is one of the main mechanisms through which MNCs evade taxes. In recent years, tax evasion by multinational corporations has been the object of sharp criticism from civil society campaigns and state institutions alike (see, e.g., Christian Aid (2009), Cobham et al. (2020), and UNECA (2019)). By "booking" profits in tax havens instead of the countries where the economic activity originated, multinational corporations abscond from their responsibilities to appropriately compensate the jurisdictions of origin for the factors of production that they provided and that were necessary for the realization of MNCs' profits (e.g., infrastructure, an educated labor force, etc.); in doing so, they do not behave as "good corporate citizens".

Finally, another dimension of trade misinvoicing is hiding wealth and capital offshore, away from the purview of regulators and tax collectors. By creating shell companies in offshore financial centers, and in collusion with a trade partner in the country of origin, funds can be transferred offshore by manipulating trade invoices. Zucman (2013) estimates that up to 8% of global wealth is held offshore in tax havens. Shifting money offshore can enable the wealthiest individuals to avoid paying their fair share of taxes, which deepens inequality and robs governments of revenues to finance the needs of the state. Offshore financial centers that specialize in providing a combination of legalized opacity and lax regulation can also threaten democratic outcomes by harboring money that entrenches the power of unaccountable political leaders and corrupt elites (Christensen, 2012; Shaxson, 2011; Shaxson & Christensen, 2013). Andersen et al. (2017) found that exogenous increases in petroleum rents were associated with an increase in hidden wealth in

[1]The authoritative statement of the arm's length principle is found in paragraph 1 of Article 9 of the OECD Model Tax Convention (see OECD (2017)).

[2]Notably, in the context of intangibles such as intellectual property and brand names; observe for example the case of market-leading technology companies, such as Apple, who manage to pay Lilliputian corporate taxes on their profits through strategic arrangements on the sale of their IP (Cobham & Janský, 2020; Tørsløv et al., 2018; UNECA, 2018b).

autocratic countries, and that around 15% of windfall profits were diverted to secret accounts.

Therefore, trade misinvoicing is a phenomenon that has wide-ranging societal ramifications, from international security, to tax justice and the perpetuation of inequality. Combating IFFs from trade misinvoicing has been recognized as an urgent policy priority at the highest political levels and has propelled international cooperation between countries. The fight against IFFs has been enshrined as a United Nations Sustainable Development Goal (SDG 16.4), endorsed by the African Union,[3] and is the object of ongoing intergovernmental policy efforts (see FACTI (2021) and UNODC and UNCTAD (2020)). Thus, trade misinvoicing is squarely acknowledged as a pressing concern for developing countries: combating trade misinvoicing is crucial to domestic resource mobilization for low income countries to be able to finance their own sustainable development (High Level Panel on Illicit Financial Flows from Africa, 2015; O'Hare et al., 2014; UNCTAD, 2020; UNECA, 2017).

2.1.2 Measuring trade misinvoicing

Illicit trade flows are, by definition, hidden, and so challenges to quantification remain a significant impediment to studying and tracking the phenomenon of trade misinvoicing. In the previous chapter of this dissertation, I presented an original database of illicit trade flows, developed using a methodology that delivers improvements on long-standing concerns in the literature regarding the credibility of "trade gaps" approaches. This database – the "atlas of misinvoicing" – provides a measure of the dollar amount of illicit activity that is embedded in each bilateral trade transaction. The methodology looks for gaps in bilateral trade statistics that are reported to the United Nations Commodities Trade (UN Comtrade) database. The customs authorities in each country participating in Comtrade regularly report the dollar value of commodities that were

[3] The High Level Panel on Illicit Financial Flows from Africa (HLP), chaired by former South African president Thabo Mbeki, was established with a mandate to assess the extent and causes of IFFs from Africa. The HLP established that IFFs were a significant drain on the resources of the continent and that combating IFFs was imperative in order to empower African countries to rely on their own resources to finance development. The policy recommendations of the concomitant report (High Level Panel on Illicit Financial Flows from Africa, 2015) were subsequently endorsed at the Twenty-Fourth assembly of the African Union in January 2015 in Addis Ababa, Ethiopia (see African Union (2015)).

traded internationally, either through imports or through exports. UN Comtrade contains detailed disaggregated data on commodities using the Harmonized System (HS), the international nomenclature for trade classification, where commodities belong to a certain product category that can be hierarchically mapped to a less detailed product category, and so on. The "atlas" measure uses Comtrade data at the highest level of aggregation (the 2-digit HS code) and thus provides estimates that can be broken down in 99 HS chapters. The "atlas" measure can also be aggregated up to the reporter-partner-year level; a feature that is exploited in this paper.

The "atlas" database provides the widest coverage of any existing estimates of illicit trade; with comprehensive bilateral estimates for 167 countries and their trading partners for 99 sectors in each year during 2000-2018. To generate this database, the entire Comtrade database was scraped for a period of 20 years. However, some low income countries do not report to Comtrade at all, and some countries' reports are patchy, because they do not report every year or for every commodity. There are 44 African countries in the "atlas" database, which means that 10 African countries are missing from the database. Yet, the non-reporting countries still export and import goods and participate in the global market for commodities; there are few truly autarkic nations (e.g., North Korea).

Therefore, this paper demonstrates how the problem of missing data in African countries can be mitigated using machine learning approaches. The paper shows that machine learning algorithms can reliably be trained to recover bilateral estimates of trade misinvoicing without requiring trade statistics to train the model. Using publicly available data that is more readily observed (e.g., distance) or collected (e.g., Gross Domestic Product), the Random Forest algorithm is able to explain around 70% of the variance in illicit trade. The predictor variables used in this paper are bilateral or unilateral country characteristics. Features such as the distance between countries and whether a given country pair share a colonial past are directly observed and have been compiled in CEPII's *Gravity* database. Other predictors are measures that are constructed by researchers to proxy some underlying political phenomenon, such as the perceptions of corruption and the rule of law. Those "construct" variables are obtained from the World Bank's *Worldwide Governance Indicators* and the Tax Justice Network's *Financial Secrecy Index*. Finally, some of the independent variables used in the paper describe macroeconomic policies, e.g., the presence of capital controls, and are compiled by the IMF in its *Capital Control Measures* dataset. All the databases used in this paper endeavor to be global databases with comprehensive country coverage.[4] While compiling data on observed economic policies is not trivial and requires work on the part of researchers, any missingness in the data is not a direct result of poor data collection practices in developing countries. In other words, these data are not directly afflicted by weak statistical capacity in developing countries, insofar as they do not rely on active data collection by customs authorities. Of course, the reasons for why a country might not report trade data to Comtrade will be correlated with the reasons for why a researcher would find it onerous to compile data on its economic policies. Both situations will partly have to do with the fragility of a country's statistical institutions that leads to the problem of poor or missing data in developing countries (Devarajan, 2013; Jerven, 2013; Jerven & Johnston, 2015; Sandefur & Glassman, 2015).

The approach presented in this paper is not meant to replace the "atlas" measure presented in the previous chapter. Rather, it should be seen as a supplement that can augment the evidence base on illicit financial flows from trade misinvoicing. Indeed, the methods described in this chapter and the preceding one have fundamentally different ojectives and are tools designed to ameliorate a specific challenge in the study of illicit finance. The "atlas" measure presented in chapter 3 is a measure that is designed to estimate trade misinvoicing with improvements over existing techniques. The methodology of the "atlas" measure includes econometric adjustments so that observed trade irregularities are not uncritically equated with illicit financial flows. A sophisticated estimation strategy is developed to estimate the dollar value of illicit trade with

[4]The *Gravity* database contains data for 252 countries (including national designations that do not exist anymore) for 1948-2019, the *Worldwide Governance Indicators* database covers over 200 countries and territories since 1996, the *Financial Secrecy Index* provides scores for 112 jurisdictions in 2018, and the *Capital Control Measures* dataset presents data for 100 countries for 1995-2017.

some precision and with methodological rigor. The "atlas" approach seeks to create an outcome measure that can be scaled across countries in order to have wide country coverage. Despite the fact that the measure is explicitly designed so as to be generalizable across all countries, since it relies on bilateral trade statistics as a starting point, the coverage of the "atlas" database is necessarily limited by the data coverage of UN Comtrade, and the "atlas" does not contain data for 10 out of the 54 African countries. Therefore, the objective of this chapter is to evaluate the potential of predictive methods to accurately fill missing data gaps. Predicting IFF outcomes in order to supplement existing estimates is a task that solves a certain type of "prediction policy" problem (Kleinberg et al., 2015). This paper provides suggestions for how predictive tasks can help researchers and practitioners get a better handle on the problem of illicit finance. The next section places this paper's contribution in the context of wider prediction problems that exist in the field of illicit finance, and suggests that machine learning techniques can confer specific advantages to address these types of problems.

2.2 Prediction policy problems in illicit finance

Causal policy problems are distinct from prediction policy problems (Kleinberg et al., 2015). The first class of problems asks questions of the type "should I invest in this policy intervention to tackle the social problem?" while the other tends to ask "is this going to be a problem?".[5] The first requires answering the causal question of "what is the effect of the policy intervention on the social problem?". By contrast, pure prediction problems only require information about the predicted outcome in order to answer the question "what is the likelihood that this problem will occur?".[6] Kleinberg et al. (2015) make the case that prediction problems are more common than

[5]Kleinberg et al. (2015) call them "rain dance" (if there is a drought, should a policy-maker invest in rain dances to increase the chance of rain?) versus "umbrella" (if she sees clouds, should a policy-maker grab an umbrella to avoid getting wet?) problems. The rain dance problem requires causality (do rain dances cause rain?) while the umbrella problem is a prediction problem (is the predicted chance of rain high enough to warrant an umbrella?). In fact, the umbrella problem is a pure prediction problem because solving it requires only knowing the predicted level of rain; the umbrella has no incidence on the level of rain.

[6]Different machine learning methods are used to answer variations of this question. The aforementioned question is answered by soft classifiers that produce predicted probabilities (e.g., logistic regression, discriminant analysis). Hard classifiers (e.g., K-nearest neighbors, decision trees) will answer the question "will this problem

is usually understood in policy domains, and that improving our ability to solve prediction policy problems can not only lead to large welfare gains but also generate useful theoretical insights.

Decision-makers and academics working on illicit finance could potentially reap substantial benefits from tackling prediction policy problems, because illicit finance is a domain where identifying risk and predicting unit-level responses is valuable. The use of machine learning (ML) to accomplish predictive tasks is ubiquitous in fields such as criminal justice, social policy, finance, and healthcare (see, e.g., Chandler et al. (2011), Ge et al. (2020), Kleinberg et al. (2017), and West and Bhattacharya (2016)); applications which, similarly to illicit finance, also involve assessing risk and predicting heterogeneity in outcomes. Predicting whether an accused person is a flight risk helps judges decide whether to grant bail or not, and predicting at-risk youth aids in targeting social policy interventions.[7] Financial actors routinely use machine learning to assess the likelihood of a transaction being fraudulent or the risk that a potential borrower will default; while individualized diagnoses are improving patient care in medicine.

The types of problems that decision-makers working on the fight against IFFs have historically tackled are: (1) tracking and recovering, to the extent that it is possible, illicit funds; (2) assessing whether a particular financial transaction is at risk of being an illicit activity given the features of the transaction; and (3) understanding the underlying determinants of IFFs to devise policy interventions that address the root causes of illicit finance. In this paper, a new type of prediction policy problem in illicit finance is identified: (4) augmenting existing estimates of IFFs when constraints imposed by the data-poor environment of developing countries preclude data

occur, yes or no?". Outside of the classification realm, regression questions will answer "how much of a problem will there be?". All of these are examples of supervised machine learning.

[7] It is vital to be cognizant of the hazards of using artificial intelligence (AI) in fields where fairness and equity should be first-order concerns, such as law enforcement. The well-known ML adage "garbage in, garbage out" is particularly premonitory here: models trained on data collected from racist, unequal, or biased interactions will reflect those systematic biases in the predictions they make. For example, candidate-screening programs used to screen resumes and sort between job applicants can reflect existing discriminatory hiring practices. Predictive policing programs are dangerously prone to bias against poor and minority communities (O'Neil, 2016). Movements towards ethical AI recognize that the development of AI models with a singular focus on predictive accuracy should be jettisoned in favor of an approach that makes space for concerns about equity, privacy, and explainability (the so-called "right to an explanation"). One distinction to note is that this paper is not concerned with making predictions for individual persons, but rather generates predictions about country-level patterns.

collection and measurement. Problems (1), (2), and (4) are prediction policy problems that will benefit from the use of machine learning methods, while (3) requires some causal knowledge.

Given the damage that they occasion, stopping IFFs is an urgent priority for policy-makers, and many multilateral cooperation initiatives have concentrated on "following the money", tracking it, and recovering it. International organizations such as the United Nations Office on Drugs and Crime (UNODC), Interpol, and a multitude of counter-terrorism agencies coordinate efforts to tackle the activities that generate criminal profits. Stopping financial crime is the dominion of Financial Intelligence Units (FIU), law enforcement units in different jurisdictions that cooperate with each other in order to fight money laundering across borders. Further programs are aimed at repatriating money to governments if the IFFs involve stolen state assets, such as the Stolen Asset Recovery Initiative (StAR), a partnership between the World Bank and UNODC. In the case of IFFs, even if the root causes of the problem are ill-identified, treating the symptoms is important. Predictive tasks can help sharpen forensic analysis to detect instances of IFFs.

Machine learning approaches can help policy-makers know where to invest resources for monitoring and detection. The questions of which shipment the customs official should inspect or of which tax return the assessor should audit are at their core resource allocation problems, where the government must decide how to spend limited resources in order to have the best chance of catching and stopping the IFF. Solving this resource allocation problem requires answering predictive questions about which transaction is the riskiest. To address this need, the TRACE program was launched in July 2021 by a consortium of European law enforcement agencies, NGOs, and universities to detect and disrupt illicit flows in real-time using AI technology.[8]

Recognizing the potential of establishing risk profiles for generating policy-relevant intelligence (FATF & Edgmont Group, 2020), Lépissier and Cobham (2019) develop a dataset and methodology for an index of countries' vulnerability and exposure to IFFs, that is subsequently used in the Tax Justice Network's data tool IFF Vulnerability Tracker.[9] Given that secrecy is required

[8]See https://www.vicesse.eu/trace.

[9]See https://iff.taxjustice.net/. The dataset presents a global atlas of countries' vulnerability and exposure

to obscure an illicit financial flow, this approach measures the vulnerability of a country to IFFs based on its partners' financial secrecy. If a jurisdiction transacts primarily with highly secretive countries, the country will score as highly vulnerable to IFFs. Risk-based approaches cohere with the conceptualization of IFFs as a wicked problem and reflect the view that IFFs are not just about criminality where the control of money laundering can be enforced through market discipline. Rather, illicit finance is seen as reflecting broader problems of inequality and barriers to achieving tax justice globally.

Risk-based approaches consider that features of the transaction can be indicative of vulnerability to IFFs, and as such can assist the development of more targeted policy interventions. Learning what type of country characteristics are predictive of the risk of IFFs is still a useful heuristic when deciding what policy priorities should be, even if the causal mechanism is unknown. Governments work bilaterally with their foreign counterparts and negotiate to exchange information that might be useful to detect IFFs. Initiatives such as the OECD's Base Erosion and Profit Shifting (BEPS) action plan have highlighted the value of sharing information on individual taxpayers, company registries of beneficial ownership, and activity reports of multinational companies.[10] For government officials, knowing which countries among their economic and financial partners have a high propensity for IFFs is a useful guide when entering bilateral negotiations on a variety of issues such as trade agreements, tax treaties, and conventions on information-sharing. Therefore, approaching illicit finance as a predictive problem has helped guide efforts to combat IFFs.

This chapter employs machine learning models to tackle prediction policy problem (4), a type of problem that is distinct from, but related to, problem (2) of assessing country-level risk.

scores to IFFs based on cross-border transactions in 8 different channels: banking positions (claims and liabilities), foreign direct investment (outward and inward), foreign portfolio investment (outward and inward), and trade (exports and imports).

[10] Various policy proposals are aimed to reduce the informational asymmetry between countries that hinders the identification of IFFs. Automatic Exchange of Tax Information (AEOI) between fiscal authorities is useful to prevent tax evasion. Registries of the ultimate beneficial ownership of companies can help identify if an individual or corporation is using shell companies to disguise illicit activities through a complex web of corporate ownership. Finally, country-by-country reports (CBCR) of the activities and profits of a parent multinational company broken down by the country of their subsidiaries can point to instances of profit shifting by multinational corporations to avoid taxes.

Specifically, when existing measures of country-level IFFs rely on official statistics where data availability is limited, the challenge is recast as a missing data problem. The method presented here predicts the amount of illicit trade between two countries, without requiring international trade statistics as part of the training process. Most methods to estimate trade misinvoicing, including the "atlas" measure presented in this dissertation, require mirror trade statistics as an input. Consequently, when countries lack the reporting capacities to provide those statistics, the misinvoicing measure will be limited. This paper shows that the issue of missing data in developing countries can be viewed as a prediction policy problem that involves harnessing predictive tasks to augment existing data. The chapter contributes a method can be viewed as a type of unit-level imputation procedure. The phenomenon of IFFs is more generally a missing data problem too, since it concerns flows that are not systematically recorded and are deliberately hidden. Moreover, given that the flows result from activities that are in contravention of laws and norms, they are concealed by design, and thus the data cannot be assumed missing at random (Molenberghs et al., 2014). Here, this paper aims to attenuate one specific aspect of the missing data problem of illicit finance, so that a measure of misinvoiced trade can still be provided even if the countries do not report to Comtrade.

However, machine learning is no panacea, and ML methods will struggle with problems of type (3) that relate to the design of policy interventions that require an understanding of the determinants of IFFs. Finding that a risk factor is highly predictive of IFFs does not imply that enacting policies which affect the risk factor will lead to changes in the amount of IFFs. While prediction problems can help answer the question of targeting IFF interventions on the sectors most afflicted by IFFs, the problem of how to direct IFF interventions to the sectors that would benefit the most from the intervention is much harder because it requires knowing the causal effect of the intervention, and this necessitates some counterfactual statement about what would have happened under an alternative policy scenario (Athey et al., 2017). The problem can also be difficult because of heterogeneity across units. It may be the case that some countries with high levels of corruption will be less responsive to the policy treatment than countries with lower levels of corruption, and so the policy will work to reduce more IFFs in low-corruption countries. If countries with low corruption experience less IFFs to start with (which is less than certain), this further underscores the difficulty of optimally targeting political action against IFFs.

An important caveat is that, in practice, many policy problems require a combination of causal and predictive inference to be solved. Although exclusively focusing on prediction will not help us address problems that contain underlying causal questions, Kleinberg et al. (2015) contend that elucidating those problems can still yield substantive and theoretical insights. For example, they suggest that understanding how agents change their behavior as a response to the way that law enforcement change their monitoring strategies can shed light on the game theory of enforcement. Gonzalez-Lira and Mobarak (2019) show how regulated agents engage in subversive adaptation to circumvent monitoring attempts. They provide empirical evidence of how Chilean fish vendors adapted to changing monitoring schedules during a fish ban in order to keep selling illegal fish. Predicting IFFs once new rules are in place, e.g., more stringent AML provisions, can help identify the new types of stratagems and previously untapped loopholes that agents exploit as a response. Given the "whack-a-mole" nature of the wicked IFF problem, an iterative process of identifying IFFs will be crucial to make progress.

2.3 Types of data used in machine learning studies of illicit finance

The previous section has argued that the study of illicit finance is well-suited to a predictive inferential framework, given the prevalence of policy prediction problems in the field. Machine learning methods have a comparative advantage in tackling tasks that require making predictive inferences in order to successfully accomplish them. In particular, the applications of ML in the financial sector have flourished. Next, this section reviews specific applications of machine learning to illicit finance and contrasts the use of high-resolution transaction-level data that is passively collected by financial institutions with the aggregate economic statistics that must be actively collected or compiled by governmental authorities, in order to underscore the difficulty of measuring illicit trade in data-constrained environments, and thus the value of using predictions from machine learning as a mitigation strategy to fill the gaps.

The application of machine learning techniques to the study of IFFs has progressed faster in the finance literature than in social sciences for several reasons. Financial data, such as banking transactions, exist at a higher cross-sectional resolution than data on macroeconomic variables such as international trade, which greatly increases the number of observations N that are available for model training. Likewise, financial data have a higher temporal resolution, often offering daily if not hourly records of transactions, which can be fruitfully exploited by ML algorithms. Finally, financial data tend to have clear and well-documented class labels, e.g., "fraud" or "not fraud", that are amenable to classification tasks at which ML techniques excel. For example, credit card companies collect masses of data on documented cases of fraud and on benign transactions in the course of their usual business operations. Financial institutions possess troves of data: FATF recommendations state that banks should be legally required to file so-called "suspicious activity reports" with the Financial Intelligence Unit (FIU) of their country in cases where they have grounds to believe that the funds are the proceeds of crime or are related to terrorist financing.[11] By contrast, the type of macroeconomic and "macropolitical" data pertaining to IFFs that are common in social sciences, e.g., country-level trade, governance variables, policy variables describing the regulatory environment, etc., exist in lower volumes than microdata, and measure concepts that are harder to pin down than a binary classification of "good" or "bad".

One limitation of financial data is that they are often confidential and not publicy available to researchers, who must instead expend considerable effort in negotiating a memorandum of understanding with the financial institution that collects the data, and then must conduct further pre-processing operations in order to properly anonymize the data. This limits the usefulness of these datasets for work that is in service of public policy. The use of Generative Adversarial Networks (GAN) to create synthetic datasets that have the same statistical properties as the original financial datasets is a promising development (Efimov et al., 2020). GANs can be used to

[11] See FATF recommendation 20: https://www.cfatf-gafic.org/index.php/documents/fatf-40r/386-fatf-recommendation-20-reporting-of-suspicious-transactions. The specific grounds for reporting are detailed in each jurisdiction by the government agency in charge of financial crimes in that respective country.

create artificial datasets that replicate with high fidelity existing datasets (Goodfellow et al., 2014) that contain sensitive or confidential information, and thus can allow researchers and practitioners to learn things about the original dataset while sidestepping many legal and regulatory difficulties. By contrast, while aggregate country-level statistics pertaining to economic or political outcomes will face issues of data scarcity, the constraints associated with proprietary data will be less binding.

Practitioners in the financial industry – including regulators, FinTech companies, and management consultants – have recognized the business value of Artificial Intelligence (AI) systems that deploy data mining in real-time (Deloitte, 2018; SAS, 2019), and the innovative potential of AI for financial regulation (Brainard, 2021; FATF, 2021).[12] Opportunities to leverage learning from data in the fight against money laundering and global terrorism have long been recognized (Senator et al., 1995), though are still largely untapped, and form the basis of an active area of research (Canhoto, 2020; Labib et al., 2020; Tiwari et al., 2020). For example, Natural Language Processing can be used to screen customer names against global lists of known criminals, and black-listed or sanctioned organisations and individuals (Deloitte, 2018).

Advances in machine learning in the financial realm to detect illicit transactions can broadly be classified into supervised learning methods on the one hand, and unsurpervised and self-supervised learning approaches on the other (West & Bhattacharya, 2016). Supervised learning methods are used on labelled data, that is, when data has been collected on an observed outcome variable, e.g., a transaction is labelled "fraud" or "not fraud". Support vector machines, boosting algorithms, and logistic models have variously been used to classify transactions into either of the "good" or "bad" categories (Jullum et al., 2020). A well-known problem when using classifiers to identify suspicious transactions is that of class imbalance: for every transaction that is labelled as "fraud", there are orders of magnitude more transactions that are not fraudulent. When the distribution of observations across the known classes is not equal, the classifier will be biased towards the

[12] Following the meteoric rise of "FinTech" as a concept, regulators are now calling the application of technologies to financial regulation "RegTech", see https://www.fatf-gafi.org/publications/fatfgeneral/documents/fintech-regtech-mar-2016.html.

majority class ("not fraud") and will struggle to predict cases in the minority class ("fraud"). Thus, a naive binary classifier for anomaly detection by Financial Intelligence Units would be useless because it would miss most of the true cases of money laundering (i.e., there would be too many false negatives). Approaches to solve the class imbalance problem include over-sampling the minority class, under-sampling the majority class, and synthetic sampling (Sudjianto et al., 2010).

Sometimes, the analyst does not have access to labelled data, i.e., there is no response Y, and they must do with the characteristics of the transactions themselves, X. The goal of unsupervised learning is thus to gain insights from the distribution of X, to identify hidden patterns and expose previously ignored similarities and differences between groups of observations. Clustering can be used to derive client profiles (which can then be used as inputs to supervised learning models) (Alexandre & Balsa, 2016). Graph-based approaches exploit the network structure of financial transactions and have been used to perform community detection (Fortunato, 2010), to analyze group suspicious behavior collectively (Savage et al., 2016), or to infer the suspiciousness of entities from the directionality of the transaction (Joaristi et al., 2019). Some approaches use the temporal nature of the data to look for sequential irregularities (Gupta et al., 2014; Li et al., 2009). Other approaches use statistical methods that search for deviations from a benchmark in order to identify anomalies (Badal-Valero et al., 2018; Raza & Haider, 2011). Finally, Paula et al. (2017) use an unsupervised deep learning AutoEncoder network to identify anomalous patterns in the exports of Brazilian corporations that might be indicative of export fraud or money laundering.

Recently, the applicability of AI techniques to the international trade setting has been explored. Machine learning algorithms been used to predict bilateral trade flows (with no attempt to distinguish between their licit or illicit nature). Authors have highlighted the timeliness of using contextual AI to predict international trade in the face of outlier events such as global pandemics and trade shocks (Batarseh et al., 2019, 2020; Gopinath et al., 2020). Forecasting future trade patterns is a prime concern of policy-makers given the impact of trade on employment and wages. Likewise, trade predictions are valuable because they assist with GDP forecasting and macroeconomic planning (Batarseh et al., 2020).

Bilateral trade data can be disaggregated to the commodity level, where the unit of observation is a reporter-partner-commodity-year tuple, or it can be aggregated over commodities to obtain aggregate trade patterns between countries. Quimba and Barral (2018) and Wohl and Kennedy (2018) use neural networks to predict aggregate trade patterns for the US and APEC countries, respectively, and find that they have a substantially lower out-of-sample prediction error compared to linear regression models. Neural networks have a superior predictive performance possibly because they are able to combine features in complex non-linear ways compared to parametric models. However, parametric models have the advantage that coefficients can often be interpreted as elasticities, which is useful for economic analysis, whereas neural networks are often criticized for being a black box (Wohl & Kennedy, 2018).

Other authors have used tree-based methods such as random forests or boosting to predict imports or exports of specific agricultural commodities and find once again that they are able to generate reliably accurate predictions (Batarseh et al., 2019, 2020; Gopinath et al., 2020). When forecasts of international trade are made, they traditionally rely on a combination of expert case studies, simple forecasting models, and large Computable General Equilibrium models (Batarseh et al., 2020). As a result, forecasts can be *ad hoc* and are not able scale to easily. The application of ML techniques to predicting international trade is a recent development which has shown promise, though results are still limited to a specific subset of countries or commodities.

This paper demonstrates an alternative use case for the application of ML to the trade setting, by showing that predictions of *illicit* trade can be generated even if the underlying data on the *observed* trade flow is missing. While the extant literature on machine learning and trade seeks to solve a forecasting problem, this paper shows that machine learning can also be used to address a missing data problem by generating reliable predictions of illicit trade in cases where the underlying trade data is missing or is patchy.

Next, the paper follows best practice in the application of machine learning to social scientific analysis by leveraging theory-guided domain knowledge to inform the selection of features that will be used as predictors in the machine learning algorithms (Mullainathan & Spiess, 2017; Storm et al., 2020). Section 3 below critically reviews the literature in order to identify a set of variables that are likely to realize high predictive returns for misinvoiced trade.

3 Theory-guided variable selection

There are three major literatures that are pertinent to trade misinvoicing and which can inform the selection of predictor variables. Using insights from these literatures, I identify a set of predictors that are positioned along the dimensions of gravitational push-pull factors, the "illicit premium", and market and regulatory abuse. First, since the "atlas" measure is constructed using bilateral trade statistics, the international trade literature in economics is germane to trade misinvoicing. Thus, section 3.1 discusses the most commonly used model in international trade analysis – the gravity model – and the debates on how the variables that represent the economic forces of attraction between countries should be specified in the model. Second, I trace the intellectual history of the extant literature on trade-based money laundering to the literature on gravity models. Section 3.2 discusses the Walker-type models of trade-based money laundering: modified gravity models that have been augmented with variables that proxy the desirability of a country for illicit business. Finally, section 3.3 presents a typology of the various stratagems that are used to manipulate trade invoices, and categorizes the literature on the determinants of trade misinvoicing according to these manipulations. This literature emphasizes the type of regulatory environment that generates differential incentives to misinvoice in order to abuse or evade market rules. Together, these literatures provide valuable insights on how to approach variable section.

3.1 The gravity dimension: gravity models of international trade

The "atlas" trade misinvoicing measure estimates the portion of any given bilateral trade transaction that is illicit. The database provides estimates of the dollar value of illicit trade for any given country pair in any given year. Thus, the "atlas" measure is in the same range as bilateral trade statistics. In other words, since the amount of illicit trade is in part a function of reported trade, we can use theories of international trade as a starting point to identify relevant variables. The gravity model of international trade has long been the workhorse of international economics. First developed by Tinbergen (1962) and refined by Anderson (1979), the gravity model provides an explanation grounded in Newtonian mechanics for bilateral trade flows. The gravity model holds that, like celestial objects, countries attract international trade flows as a function of their size and distance from others. The modern form of the model also includes a measure of relative trade costs (Anderson & Wincoop, 2003). The basic model is of the form $V_{ij}^X = \beta_0 Y_i^{\beta_1} N_i^{\beta_2} Y_j^{\beta_3} N_j^{\beta_4} D_{ij}^{\beta_5} \exp(\beta_6 P_{ij})$ where V_{ij}^X represents the value of exports from country i to its partner j, Y is the Gross Domestic Product (GDP) of the countries, N is their population size, D_{ij} is the geographical distance between them, and P_{ij} is a measure of bilateral trade facilitation (such as the existence of a Preferential Trading Agreement, or alternatively of barriers to trade) (Anderson, 1979; Anderson & Wincoop, 2003; Disdier & Head, 2008; Ferwerda et al., 2013; Tinbergen, 1962).

The model is often presented in its logarithmic form to linearize the parameters in order to allow for estimation with linear regression models: $\ln V_{ij}^X = \ln \beta_0 + \beta_1 \ln Y_i + \beta_2 \ln N_i + \beta_3 \ln Y_j + \beta_4 \ln N_j + \beta_5 \ln D_{ij} + \beta_6 P_{ij}$ (McCallum, 1995; Tinbergen, 1962). Variations include assuming that $\beta_2 = \beta_4 = 0$ to remove population size from the equation (Bergstrand, 1985; Ferwerda et al., 2013; Tinbergen, 1962), using GDP per capita as a measure of economic size instead of GDP, or entering distance as a negative term to reinforce the connection with Newton's equation for universal gravitation (Disdier & Head, 2008; Ferwerda et al., 2013).[13]

[13] To see this, subsume trade costs in the constant and write $\ln V_{ij}^X = \ln \beta_0 + \beta_1 \ln Y_i + \beta_2 \ln Y_j - \beta_3 \ln D_{ij}$ and take exponentials on both sides, so that the equation more closely resembles Newton's formulation: $F = G\frac{m_1 m_2}{r^2}$, where F is the gravitational force between the objects, m_1 and m_2 are their masses, r is the distance between

Despite the intuitive appeal of the model, disagreements are rife on how to specify the model in ways that are theoretically consistent. Anderson (1979) provides the first attempt to square the gravity model with economic theory. Others seek to prove that the gravity model coheres with the Heckscher-Ohlin theory of international trade (Deardorff, 1998; Ferwerda et al., 2013; Yotov et al., 2016).[14] Yet, despite the model's consistently high explanatory power, the model has been criticized for a lack of strong theoretical foundations (Anderson & Wincoop, 2003; Bergstrand, 1985).

One area of difficulty is the famous "distance puzzle" (McCallum, 1995). A persistent empirical result in international economics is that bilateral trade decreases with distance (Disdier & Head, 2008), despite falling trade costs and the notion that the world has become flatter/smaller with globalization (Yotov, 2012). As a response, Anderson and Wincoop (2003) argue that the gravity model is not correctly specified because it does not take into account what they call "multilateral resistance", which is that the more resistant a country is to trade with all others, the more it will be pushed to trade with a given partner, and so controlling for multilateral resistance terms theoretically solves McCallum (1995)'s border puzzle. One of the proposed empirical solutions is to augment the traditional gravity equation with importer and exporter fixed effects (Anderson & Wincoop, 2003; Feenstra, 2004; Piermartini & Yotov, 2016; Santos Silva & Tenreyro, 2006; Yotov et al., 2016) in order to estimate multilateral resistance with cross-sectional data.

However, it is subsequently argued that the correct way to account for multilateral resistance is to incorporate exporter-time and importer-time fixed effects in panel data estimations (Piermartini & Yotov, 2016; Yotov, 2012; Yotov et al., 2016). Note that including country-year fixed effects comes with its own challenges. In a predictive setting, including country-year fixed effects might overfit the training data and as a result the model would generalize less well for out-of-sample predictions (Wohl & Kennedy, 2018). In a parameter estimation setting, the inclusion of country-year fixed effects would preclude the inclusion of any explanatory variables that vary across

them, and G is the gravitational constant (Disdier & Head, 2008).

[14] Disconcertingly, the HO model is silent on the role of distance in international trade, and considers that countries are disembodied entities that trade with each other in a frictionless world (Deardorff, 1998).

countries and time (such as GDP or trade costs), and the effect of potentially relevant explanatory variables for dirty money flows would not be identified (e.g., whether a country has a particular AML provision in place in a given year) (Kellenberg & Levinson, 2019).

The appropriate estimation technique and the relative merits of various estimators have also been debated. Santos Silva and Tenreyro (2006) argue that the gravity model should be estimated in its multiplicative form rather than in its additive form, because Jensen's inequality stating $\mathbb{E}[\ln(X)] \neq \ln(\mathbb{E}[X])$ means that estimating the parameters of the log-linearized model with OLS can lead to biased estimates and incorrect interpretations of the elasticities. Instead, the authors argue for the use of the Poisson Pseudo Maximum Likelihood (PML) estimator to estimate the model in its multiplicative form, in order to address problems of heteroskedasticity and the existence of zeroes in trade data (Piermartini & Yotov, 2016). Others suggest that the Gamma PML estimator is preferable under certain conditions given by the data (Head & Mayer, 2014). Clearly, developing theory-consistent estimation methods remains an open problem in the economics of international trade (Head & Mayer, 2014). Moreover, difficulties with interpretability do not solely afflict machine learning approaches. The various refinements that have been proposed to deal with the problem of zero trade flows that occur when taking logarithms of V_{ij}^X, such as taking the logarithm of $(V_{ij}^X + 1)$ (Eichengreen & Irwin, 1995) or estimating multiplicative models with PPML and GPML (e.g., Disdier and Head (2008) and Santos Silva and Tenreyro (2006)), also require transformations of that data that complicate the interpretation of coefficients.

Despite struggling with economic and econometric theory, the alluring resemblance of the gravity model to one of the most universal representations of the natural world explains its continued presence in the pantheon of international trade theories. The role that distance (both geographical and cultural) and barriers to trade play in predicting observed international trade flows suggests that these variables will also be useful to predict illicit trade flows.

3.2 The "illicit premium" dimension: models of trade-based money laundering

Gravity models have been used in work related to illicit finance to analyze patterns of Financial Direct Investment (FDI) in tax havens and Offshore Financial Centers (OFC). A. K. Rose and Spiegel (2007) use a gravity model to analyze the determinants of cross-border portfolio holdings and show that proximity to an OFC has the surprising consequence of increasing the competitivity of the domestic banking sector. Haberly and Wójcik (2014) use departures from a partial gravity equation[15] to detect anomalies in patterns of global FDI, and find that between 30-50% of global FDI is intermediated through networks of offshore shell companies. Haberly and Wójcik (2015) use gravity equations to investigate the determinants of real and offshore FDI and find that colonial relationships play a role in explaining patterns of offshore FDI.

Next, I discuss how gravity models have been applied to the study of trade-based money laundering, and present the main categories of variables relating to illicit activity that have been used to supplement gravity models. Walker (1999) provides the first prototype of a gravity-based model of money laundering (Ferwerda et al., 2013, 2020). Flows of dirty money are geographically allocated between country i and j according to the characteristics of both source (i.e., where the proceeds of crime are generated) and destination countries (i.e., where the criminal proceeds are laundered). Walker (1999) postulates that the amount of money laundering sent abroad depends on the attractiveness of a destination country for the concealment of ill-gotten gains, and seeks to estimate the "illicit premium" of destination/host countries (Collin, 2019) as a function of the presence or absence of banking secrecy provisions, government attitudes to money laundering, corruption and conflict, and geographical and trading proximities. He posits that criminals and money launderers would favor countries with more stable banking regimes (Walker & Unger, 2009). Likewise, corruption is expected to have ambiguous effects on whether a host country

[15] Since FDI is meant to reflect long-term investments in the real economy, the authors use a gravity equation that predicts what meaningful bilateral investment should be between countries on the basis of product of origin and the host country's nominal GDP. Departures from predicted FDI flows are used to calculate anomalies.

attracts dirty money flows: presumably, prospective money launderers appreciate authorities that can be persuaded to look the other way, but do not want so much corruption that their money is put at risk.

Implicit in this view is the notion that money, whether dirty or clean, seeks out safe havens. Thus, even if some IFFs are generated in a criminogenic environment (recalling that not all IFFs are criminal or corrupt proceeds), some degree of political and financial stability is desirable to prevent expropriation by the state or erosion of value due to macroeconomic instability. Therefore, illicit finance shares many of the characteristics of non-illicit finance since criminals have similar motivations to a traditional investor when it comes to deciding where to place money. The same way that increasing financial integration between countries has led to the explosion of opportunities for capital to thrive,[16] financial globalization has also been a boon for illicit sectors of the economy. In this sense, the Walker model is an important innovation because it is explicitly global and reflects the transnational nature of many forms of crime (Yikona et al., 2011).

The model of TBML is then revised by Walker and Unger (2009) who now explicitly interpret it as a gravity model and refine the concept of distance. Geographical distance might be less relevant to flows of money than flows of merchandise, given that transportation costs are negligible since money can be transferred by the click of a mouse (Walker & Unger, 2009), but cultural distance might still matter because it gives rise to communication and transaction costs. Explanatory variables such as whether countries share a common language or colonial background are thus added to the model (Walker & Unger, 2009).

The somewhat *ad hoc* inclusion of the variables on the right-hand side has led to criticisms that the model is atheoretical, and the fact that the left-land side variable (money laundering flows) is unobserved has meant that the model is impossible to empirically validate (despite attempts by Walker and Unger (2009) to triangulate the results) (Collin, 2019; Ferwerda et al., 2013,

[16] And perhaps more so, to exceptional societal challenges such as the deepening and perpetuation of global inequalities (Piketty & Goldhammer, 2014), and the explosive potential for system-wide failures brought about by the destabilizing combination of increased financial interdependence and deregulation, cf. the global financial crisis of 2008.

2020). Indeed, the outcome variable in the Walker and Walker-Unger models is the percentage of proceeds from foreign crime that flow into a host country, which they operationalize by making unverifiable assumptions about the amount of domestic crime that is generated in the source country and about its relative economic profitability (Walker, 1999; Walker & Unger, 2009). Since we do not have data on flows of money laundering, the predictive power of the Walker model has not been empirically tested, and the question of whether a gravity-type equation can properly explain flows of trade-based money laundering has not been settled (Ferwerda et al., 2013). In this paper, the concern is not the underlying causal mechanisms that drive TBML; rather, the goal is to generate reliably accurate predictions of misinvoicing. While this paper does not isolate which variables can explain trade misinvoicing the best, theory does suggest that Walker-type variables may collectively have high explanatory power in predicting a measure of misreported trade.

3.3 The market and regulatory abuse dimension

3.3.1 Types of manipulations and their motivations

The challenge with the Walker model is that trade misinvoicing flows represent an amalgamation of money laundering, tax evasion, market abuses, and corporate profit-shifting. Baker (2005) launched the research agenda on illicit financial flows with the book *Capitalism's Achilles heel: Dirty money and How to Renew the Free-Market System*, where he distinguished between three broad categories: grand corruption by state officials, laundering of criminal proceeds, and commercial tax evasion by multinational companies through the manipulation of intra-firm subsidiary prices (Baker, 2005). Cobham and Janský (2020) and Cobham et al. (2014) extend Baker's classification and distinguish between IFFs stemming from tax abuse (by individuals and companies) and between IFFs which are the result of market or regulatory abuse, such as the circumvention of capital controls or taking advantage of export credits. The next section examines the literature on the determinants of trade misinvoicing to identify an additional set of variables that are likely to explain some of the variation in illicit trade outcomes. As a preliminary step to understanding how various determinants of misinvoicing might be relevant, the four ways in which trade can be misinvoiced are discussed next, since those stratagems are associated with different illicit motivations.

Trade invoices can be faked by either the importer, the exporter, or both, which gives rise to four different types of manipulations that are executed for varied reasons. Table 1 provides a typology of the types of motivations for trade misinvoicing that result in either illicit financial outflows or inflows into the importing or exporting country. In a legal trade transaction, the invoice value declared by an importer (exporter) should match with the payment (receipt) of funds recorded by financial institutions, which should accord with the (unobserved) true value of the goods (World Customs Organization, 2018). Since the true value of the goods is an unknown quantity, one strategy is to exploit the bilateral nature of international trade statistics to infer trade misinvoicing from gaps between the importer's record of the trade and the mirror record from the exporter's perspective, an approach known as a "trade gaps" analysis.

The type of manipulation depends on the aims of the misinvoicer. Shifting or retaining money abroad can be accomplished by import over-invoicing or export under-invoicing, which result in an illicit outflow where either excessive funds or merchandise leaves the country. This is a type of "technical smuggling" as opposed to the "pure smuggling" that occurs when illegal goods such as drugs are clandestinely traded (Schuster & Davis, 2020). When the value of imports is overstated, excess funds leave the country disguised as a form of trade payment (Schuster & Davis, 2020; World Customs Organization, 2018). When the value of exports is understated, this results in an outflow of merchandise in excess of the foreign exchange that is received in return. Export under-invoicing can be used to conceal profits abroad, since commodities leave the country but the corresponding financial flows stay partly in foreign accounts (Schuster & Davis, 2020), which deprives countries of precious foreign exchange and erodes their tax base. It has been argued that export under-invoicing is a more likely vehicle for illicit capital flight than import over-invoicing because customs officials tend to pay more attention to imports in order to monitor potential tariff evasion, and as a result controls on exports tend to be less restrictive (Schuster & Davis, 2020). However, the empirical record is mixed: Gara et al. (2019) provide evidence from Italian trade to suggest that export under-reporting is preferred over import over-reporting as a way to shift money abroad.

Import under-invoicing and export over-invoicing, on the other hand, will result in an inflow (or a negative trade gap). The potential to evade tariffs by understating the value of imports has been pointed out since Bhagwati (1964). Export over-invoicing, on the hand, is used to take advantage of incentives that the government puts in place to encourage exports, such as subsidies or tax credits. This paper treats trade manipulations that give rise to negative gaps or inflows as IFFs, contrary to other studies (Schuster & Davis, 2020; World Customs Organization, 2018). Although tariff evasion via import under-invoicing will look like an inflow into the importing country, it actually robs governments of tax revenues, and taking advantage of export subsidy regimes is a form of market abuse that can make it more difficult for the state to finance other socially beneficial activities. This line of reasoning is adopted by those who favor a broad rather than legalistic conceptualization of IFFs and fits within an analytical framework that determines "illicitness" following a criterion of harm, that is, an illicit flow is one that has the potential to damage economic development, and whose removal would improve social outcomes (Blankenburg & Khan, 2012; Cobham & Janský, 2020; Cobham et al., 2014; Kar & Cartwright-Smith, 2008). Therefore, this study considers trade manipulations that result in an illicit inflow as an integral part of the problem on the basis that inflows can be just as corrosive to good governance and state institutions as illicit outflows (Blankenburg & Khan, 2012; Salomon, 2019; Spanjers & Salomon, 2017). Moreover, illicit inflows may themselves be used to fund illicit sectors in the economy through the repatriation of profits by transnational crime organizations, or may be used to finance terror (Cobham & Janský, 2020).

		Imports		Exports
Out	*Manipulation*	Over-invoicing	*Manipulation*	Under-invoicing
	Motivation	Disguising illicit capital flight: money laundering of criminal or corrupt proceeds	*Motivation*	Disguising illicit capital flight: money laundering of criminal or corrupt proceeds
		Retaining money abroad		Retaining money abroad
		Tax evasion: shifting corporate profits abroad to reduce domestic tax burden (corporate), shifting undeclared income (individuals)		Tax evasion: concealing corporate profits abroad to reduce domestic tax burden (corporate), shifting undeclared income (individuals), avoiding export taxes
		Market abuse: avoiding capital controls by obtaining excess foreign exchange		Market abuse: avoiding capital controls (on profit repatriation, on foreign currency denomination)
In	*Manipulation*	Under-invoicing	*Manipulation*	Over-invoicing
	Motivation	Repatriating undeclared capital	*Motivation*	Repatriating undeclared capital
		Money laundering: incorporating proceeds into the domestic legal financial system		Money laundering: incorporating proceeds into the domestic legal financial system
		Tax evasion: evading tariffs		Market abuse: exploiting export subsidy regime: obtaining duty drawbacks or concessional rates on export finance awarded to top exporters

Table 1: Typology of trade misinvoicing manipulations by trade flow and direction. Source: author's typology adapted from Cobham et al. (2014), Gara et al. (2019), Schuster and Davis (2020), and World Customs Organization (2018).

3.3.2 Determinants of trade misinvoicing

As illustrated by Table 1, the direction of misinvoicing and the type of trade flow that is misreported depends on the underlying illicit motivation. Another strand of research is concerned with the determinants of trade misinvoicing and the observable type of trade gaps that are generated, and seeks to analyze the incentives that traders have to misreport and in what direction.

One reason to misreport or fake trade invoices is to evade capital controls or tariffs. A country may place restrictions on imports or exports in an effort to stabilize its currency and the capital account (Patnaik et al., 2012). Likewise, a country might enact tariffs to shore up a domestic infant industry or to level the playing field in terms of its exporters' competitivity when other countries do not meet the same regulatory standards (e.g., labor protections or environmental regulations on carbon emissions). The popularity of using capital controls as prudential tools to manage capital mobility and of using tariffs as protective measures to shield domestic industry has waxed and waned during the various periods of financial globalization, from the post-World War II Bretton Woods era to the Washington Consensus and beyond (Ghosh & Qureshi, 2016), depending on the degree of financial and economic liberalization that economists deemed desirable at the time. Without entering into those macroeconomic debates, the fact remains that capital controls and tariffs are policy instruments that remain the purview of a sovereign state, and that evading them is a form of market abuse that directly threatens the ability of a state to manage its own affairs.

Several authors examine the impact that these kind of policy measures have on the amount of misinvoiced trade. Vézina (2015) finds that statistical irregularities in a country's export statistics of natural resources are more likely when a country has export controls or prohibitions in place. Fisman and Wei (2009) show that corruption in the exporting country is associated with greater under-reporting of exports of cultural objects and antiques, particularly when these are export-restricted cultural objects (e.g., archeological artifacts that were illegally excavated from the country). Fisman and Wei (2004) establish that higher tariffs in China on imports from Hong Kong are associated with greater under-reporting of imports, which they attribute to tariff evasion. Javorcik and Narciso (2008) provide evidence to suggest that tariffs are correlated with trade gaps, and that this effect is stronger for differentiated products due to the difficulties that this presents to customs officials who need to gauge the quality of the products in order to ascertain their prices.

The literature also seeks to identify the relationship between misreported trade and incentives for trade misinvoicing. Carrère and Grigoriou (2015) use a gravity model to analyze the role of incentives to misreport trade and find that tariff rates and Foreign Direct Investment (a proxy for profit-shifting in order to evade taxes) partly explain import gaps. Buehn and Eichler (2011) develop a theoretical model that combines a microeconomic framework on the expected cost and benefits for firms to misreport with macroeconomic incentives such as taxes on trade and income. The authors find robust evidence that the black market premium and high export taxes are associated with export under-invoicing, and thus argue that a major incentive to misinvoice is to evade taxes on trade, but find weaker evidence on the impact of income tax differentials on trade misinvoicing. Gara et al. (2019) determine that trade gaps are correlated with differential tax rates on income and trade, tariff rates, in addition to a country's openness to trade and traditional gravity variables such as whether trading partners are part of a Regional Trade Agreement (RTA).

The quality of governance will also have an impact on the amount of trade misinvoicing but it is hard to specify *a priori* in what ways and in what direction. As Walker (1999) first pointed out, trade misinvoicers and money launderers require some degree of institutional quality and political stability to ensure that their money in safe, but too much regulatory oversight will make it difficult to shift the money in the first place. This suggests that there are non-linearities at play in the relationship between governance and illicit flows. These non-linearities are evident with market abuse too. Kellenberg and Levinson (2019) find a non-monotonic relationship between the trade gap and the tariff level. A startling result is that the gap between imports and exports (where a higher trade gap would indicate import over-invoicing) grows with the first 4 deciles of the tariff

(Kellenberg & Levinson, 2019). However, at higher tariff levels, the authors find that the trade gap shrinks as the tariff increases (which is suggestive of tariff evasion) but also find that it is much lower in absolute terms than the trade gap at lower tariff levels. To explain this hump-shaped pattern, Kellenberg and Levinson (2019) argue that there are two countervailing forces at play: the higher the tariff, the more diligent customs officials will be in monitoring accurate reporting (which would explain why the absolute trade gap is lower at higher tariff levels), but also the more incentive the importer (who pays the tariff) will have to under-invoice (which would explain the negative correlation between tariff levels and the trade gap that is observed at the higher end of the tariff distribution). In addition to this non-linearity, tariff rates are also expected to interact with corruption (S. Jean & Mitaritonna, 2010; Worku et al., 2016), since customs officials can be bribed to doctor invoices in order to evade tariffs through import under-invoicing (which shows up as an illicit inflow in trade gap analyses) or to evade taxes by shifting profits abroad through export under-invoicing (which results in an illicit outflow).

Unsurprisingly, the literature reports mixed results on the impact of corruption on money laundering. Kellenberg and Levinson (2019) find that controlling corruption and stricter auditing standards are associated with reduced export under-invoicing, consistent with the notion that greater oversight (or less corruption) makes it harder for traders to misinvoice. On the other hand, as the Walker model predicts, low corruption might also increase the illicit premium since it reduces the transaction costs of laundering related to paying bribes (Ferwerda et al., 2013). Reflecting this causal ambiguity, Ferwerda et al. (2013) find that the effect of corruption on TBML is statistically insignificant. While the connection between corruption and trade-based money laundering is unclear, the evidence record is stronger for the role of corruption in predicting trade gaps associated with tariff evasion (Carrère & Grigoriou, 2015; Rijkers et al., 2017; Worku et al., 2016).

Consistent with the expectations of the Walker model that the marketability of countries as destinations for laundering depend on their suitability for stashing ill-gotten gains, Gara et al.

(2019) find that anomalies in Italian trade records increase with the degree of financial secrecy of the counterpart, but also with a measure of its financial attractiveness and institutional stability as proxied by an index on business protection from crime and violence. However, Ferwerda et al. (2013) find that, contradicting the notion of the illicit premium, countries with anti-money laundering provisions and more hostile government attitudes to money laundering experience more TBML. The authors' proposed explanation for this finding is that existing AML regimes are almost completely focused on combating money laundering in the financial/banking system, which makes it harder to launder money the "traditional" way. Thus, the stricter a country is with respect to AML rules in the financial system, the more criminals will turn to forging trade invoices in order to launder money (Ferwerda et al., 2013). However, the adoption of AML provisions is endogenous, and the causal arrow might point in the opposite direction: countries that suffer from more TBML are also more likely to put in place AML policies.

The inscrutability of causal explanations is a persistent problem in the study of trade misinvoicing, and few papers have had convincing causal identification strategies. This paper leaves this issue to further research and does not attempt to elucidate the causal mechanisms that lead to misinvoiced trade. Although the specific ways in which these variables enter the data-generating model that gives rise to trade misinvoicing must remain a black box; I consider that they can still provide valuable information. The goal of statistical analysis is to obtain valuable information about the link between a response and predictor variables; being able to interpret the links between those factors is one way of obtaining information, and having causal clarity on these links is not a pre-requisite to generate reliable information about the link between the outcome and the predictors (Breiman, 2001). Therefore, from the point of view of this paper, reviewing extant literature illuminates the types of variables that should enter the black box in the first place. As I have shown, independent variables that can explain variation in the IFF outcome operate along three dimensions: the gravity dimension that contains the push and pull factors of international trade (reviewed in section 3.1); the "illicit premium" dimension introduced by the Walker model (see section 3.2) that augments gravity variables with factors that relate to the attractiveness (or not) of a destination country to carry out illicit activity; and finally, the dimension of market and regulatory abuse (section 3.3). The data used to represent these variables is presented in the next section.

4 Data

4.1 Outcome variable

This paper leverages the "atlas of misinvoicing", an original dataset of trade misinvoicing estimates for 167 countries during the period 2000-2018.[17] This dataset provides estimates of gross and net ouflows for imports, exports, and total trade. At its lowest level of aggregation, the database provides estimates for a reporter-partner-commodity-year tuple, where commodities belong to one of the 99 sectors of the Harmonized System for classifying international trade. This paper follows the notation in chapter 3 and indexes reporters with i, partners with j, and years with t. The database provides the amount of illicit flows embedded in each transaction reported to the United Nations commodities trade (UN Comtrade) database (United Nations, 2020), both in absolute terms (in nominal US dollars) and in relative terms (as a percentage of a country's trade and GDP that year).

The "atlas" database also presents estimates at a higher level of aggregation, by aggregating yearly bilateral flows over commodities; these are the estimates that are used in this chapter. Thus, the observational unit of the outcome variable here is a reporter-partner-year triple. The outcome variables of interest are gross outflows, `GER_Tot_IFF`, from country i to partner j (which can be the result of import over-invoicing or export under-invoicing) and gross inflows, `In_GER_Tot_IFF`, from country j into country i (which can be the result of export over-invoicing or import under-invoicing). In this database, illicit outflows from i to j are represented by positive numbers while illicit inflows from j to i will be negative values.

[17] This dataset is publicly available on a Zenodo repository (https://doi.org/10.5281/zenodo.3610557) and the code to replicate the results is available on a GitHub repository (https://github.com/walice/Trade-IFF). The methodology used to generate these trade misinvoicing estimates is described in chapter 3 of this dissertation.

4.2 Predictor variables

Section 3 has sketched the outlines of theoretical approaches to analyze bilateral trade and trade-based money laundering that are used to delineate the feature space by identifying a plausible set of predictors for trade misinvoicing. The analysis identified three dimensions along which to parse the covariate space. First, there are the canonical gravity variables that have a long history of being used to predict international trade flows and which capture: geographical distance, cultural distance, relative trade costs, and trade facilitation. Second, the "illicit premium" dimension is suggested by Walker-type models of TBML. This will include variables that proxy how attractive countries are as destinations to conceal illicit funds, which is a function of how well they can protect capital assets (through political stability and regulatory quality) and how scrupulously they scrutinize the provenance of the funds. The third dimension relates to trade misinvoicing due to market and regulatory abuse, and includes variables that proxy the incentives to misinvoice trade in order to evade tariffs or circumvent capital controls.

In the paragraphs below, the set of macro-level variables that are used in this paper and their observational level are presented. Micro-level variables are not considered since the most disaggregated level of the outcome measure is reporter-partner-commodity-year. Full details on the data sources, the coverage of the measures, and the unit of analysis is provided in the codebook included in section A of the appendix.

Together, these variables yield a set of $K = 42$ predictors (since some of these variables are recorded for both the reporter and the partner, e.g., GDP). Some of these variables are correlated with each other, and some will be more informative than others to predict illicit trade patterns.

4.2.1 Gravity variables

Gravity variables are provided by CEPII's *Gravity* database Conte et al. (2021) and include:

- country-year variables denoting mass/size characteristics such as Gross Domestic Product (`GDP`) and population (`pop`)

- bilateral variables measuring the distance between a given pair of countries i and j, both geographical (distance in km, `dist`, and a dummy for contiguity, `contig`) and cultural distance (dummies for whether countries share a common official language, `comlang`; a common colonizer, `comcol`; and whether they were in a colonial relationship post-1945, `col45`)

- variables that capture barriers to trade including a measure of trade costs (the costs of entry for doing business in each country, `entry_cost`), and a trade facilitation dummy for whether any given pair of countries have a Regional Trade Agreement (`RTA`) in any given year

4.2.2 Governance variables

Governance variables come from the *Worldwide Governance Indicators* database collected by Kaufmann et al. (2010). All of these variables are measured by percentile rank, where higher values denote better outcomes, and are at the country-year level. They include:

- how well a country controls corruption, `CorrCont`, capturing perceptions of the extent to which public power is exercized for private gain (including petty and grand forms of corruption, as well as "capture" of the state by elites and private interests) (Kaufmann et al., 2010)

- a measure of regulatory quality, `RegQual`, "capturing perceptions of the ability of the government to formulate and implement sound policies and regulations that permit and promote private sector development" (Kaufmann et al., 2010)

- a variable on the rule of law, `RuleLaw`, capturing people's confidence in the rules of society and the extent to which they abide by them, including confidence in contract enforcement, property rights, the criminal and justice system, and the likelihood of crime and violence (Kaufmann et al., 2010)

4.2.3 Financial integrity variables

Financial integrity variables are proxies for the degree of financial opacity in a country, as well as government attitudes to preventing money laundering and tax evasion. All of the variables except for FATF come from the Tax Justice Network (2020)'s *Financial Secrecy Index* (FSI). The index provides assessments of a jurisdiction's secrecy and the scale of its offshore financial activities. I extract the key financial secrecy indicators from this index that measure how tough a jurisdiction is with regard to money laundering and tax evasion. One limitation of these variables is that they are cross-sectional and there is no data on how these vary over time. Moreover, the data from the FSI were missing for many of the African countries considered here. Hence, the FSI variables are only used as unilateral characteristics for partner j. The financial integrity variables are:

- a country's aggregate secrecy score on the FSI, SecrecyScore, where a score of 100 means a country is fully secretive and a score of 0 means a country is fully transparent
- a country's rank on the FSI, FSIRank, where the top ranking corresponds to the top secrecy-weighted jurisdiction according to that country's share of the global market of offshore financial services
- a score between 0 and 100 indicating how much a jurisdiction promotes tax evasion, KFSI13
- a score between 0 and 100 of how poorly a country meets the anti-money laundering recommendations of the FATF, KFSI17
- a score between 0 and 100 of how uncooperative a jurisdiction is with other countries on judicial matters regarding money laundering, KFSI20
- a dummy for whether a country is a member of the Financial Action Task Force, FATF

4.2.4 Regulatory environment variables

Macroeconomic variables are used as proxies for the incentives to misinvoice that are generated by the regulatory environment, such as tariff evasion and the circumvention of capital controls.

Apart from the variable `tariff`, all the variables come from the IMF's *Capital Control Measures* dataset (see Fernández et al. (2015)), and are at the country-year level. From this dataset, I extract the measures of capital controls on inflows and outflows on the asset classes where controls might be plausibly evaded by misreporting trade. The regulatory environment variables are:

- to capture tariff evasion, the average tariff line applied by country i on imports from j in a given year (`tariff`). This variable is used both at a disaggregated level (commodity-level tariff) and aggregated at a country-level (from UNCTAD (2018)).
- an index of average capital controls on inflows (`kai`) and outflows (`kao`)
- restrictions on commercial credits for operations that are directly linked with international trade transactions, including an aggregate measure of controls (`cc`), and controls on inflows (`cci`) and outflows (`cco`)
- restrictions on direct investment accounts for the purpose of establishing lasting economic relations between residents and nonresidents, including an aggregate measure of controls (`di`), and controls on inflows (`dii`) and outflows (`dio`)

5 Methods

5.1 Pre-processing the data

5.1.1 Summary statistics

A brief presentation of the data is provided next. The "atlas" database presented in this dissertation has provided empirical confirmation that African countries are severely afflicted by illicit financial flows (IFFs) from trade misinvoicing, lending support to the political mandate for combating IFFs across the continent that was created by the High Level Panel on Illicit Financial Flows from Africa (High Level Panel on Illicit Financial Flows from Africa, 2015). Cumulatively, the continent experienced \$1.2 trillion of gross outflows from trade misinvoicing during the period 2000-2018 – an amount corresponding to approximately 5% of the continent's GDP and 12% of its total trade. During that period, African countries experienced a loss of \$86 billion a year on average, a figure that dwarfs the amount of aid that the continent received at the same time (UNCTAD, 2020; UNECA, 2018a). Policy action centers around preventing illicit financial flows from trade misinvoicing to constitute a new source of development finance for African countries, in order to reduce their dependence on foreign assistance or to forego it entirely (High Level Panel on Illicit Financial Flows from Africa, 2015; UNECA, 2018b).

Therefore, trade misinvoicing creates significant barriers for the prospects of sustainable development, to which the continent is particularly vulnerable (Abugre et al., 2020). Aggravating the prejudice of IFFs, the lack of data and the poor quality of official government statistics is a well-documented phenomenon in African countries (Devarajan, 2013; Jerven, 2009, 2016; Jerven & Johnston, 2015; Sandefur & Glassman, 2015). Moreover, the quality of intra-African trade statistics is poorer than records on extra-continental trade, since keeping track of customs invoices at porous land borders is more difficult than recording trade that departs from ports (UNCTAD, 2020). Thus, the estimates of intra-African trade misinvoicing in the "atlas" are likely to be underestimated.

It is important to notice that the continent also experiences a large amount of illicit trade that results in inflows, including inflows that originate from other African countries. Figure 1 displays the aggregate amount of gross inflows and gross outflows in the continent during the period of the study, which are calculated using the Gross Excluding Reversals (GER) aggregation strategy detailed in the previous chapter.[18] Gross inflows are displayed as a negative value on the figure. Illicit inflows are included in the study since they represent inflows of revenues that are not recorded and not taxed by African governments, and as such can entrench the power of autocratic leaders

[18] The GER aggregation strategy consists of adding up the illicit flows in each direction separately, ignoring the opposite flow. Thus, country-level gross outflows are the sum over partners of strictly positive IFF values, while country-level gross inflows are calculated by adding up the strictly negative IFF values over partners (since inflows are presented as negative values in the database).

in Africa (Andersen et al., 2017). If the gross outflows are netted of the inflows, the findings from the "atlas" database indicate that Africa experienced mostly net outflows from 2009 onwards.

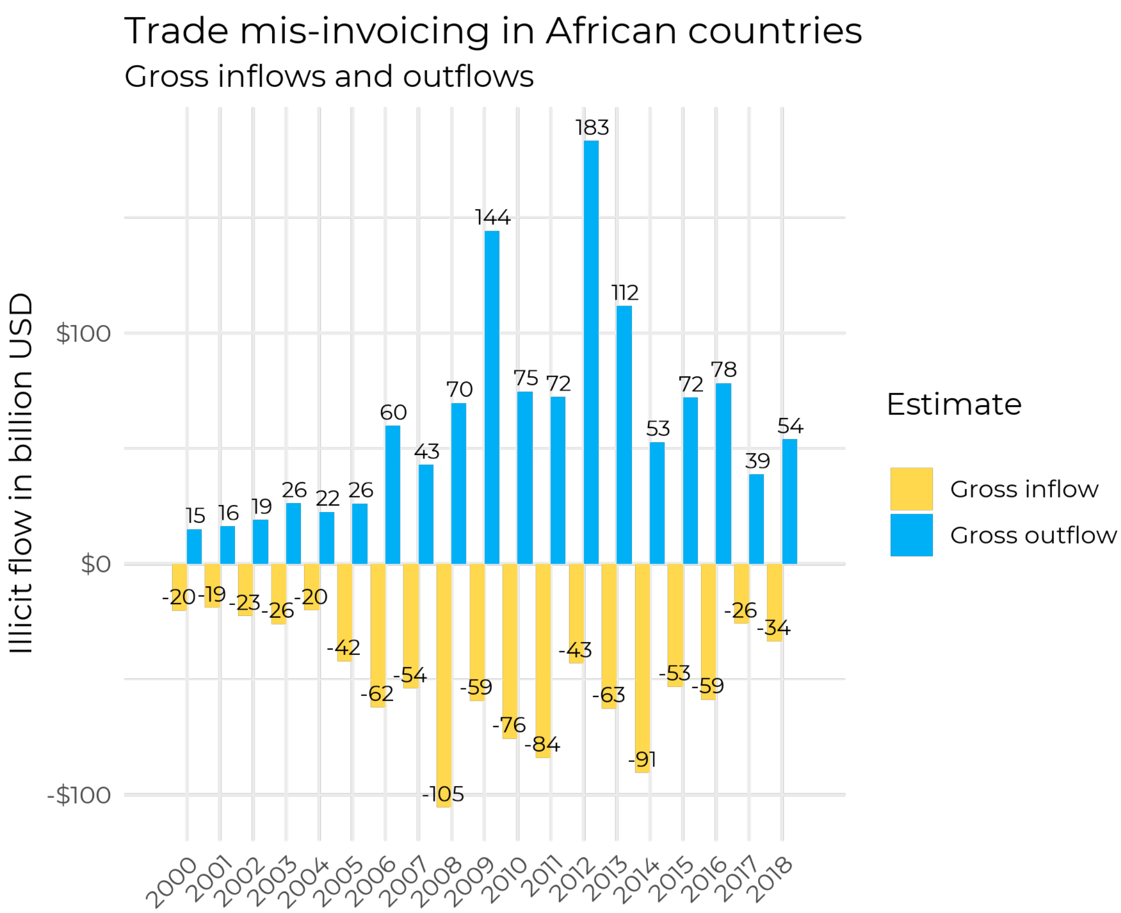


Figure 1: Yearly gross inflows (negative value) and gross outflows in Africa.

The incidence of outflows and inflows varies substantially across countries in the continent, as shown in figures 2 and 3. On a dollar basis, South Africa experienced the greatest amount of both outflows and inflows, with average annual flows of $24 billion and $13 billion, respectively. Following South Africa, Angola, Nigeria, and the Republic of Congo lead the continent in average yearly outflows. The pattern of inflows differs: in addition to South Africa and Nigeria, countries in the Maghreb are responsible for the most inflows, in particular Algeria, Egypt, Morocco, and Tunisia.

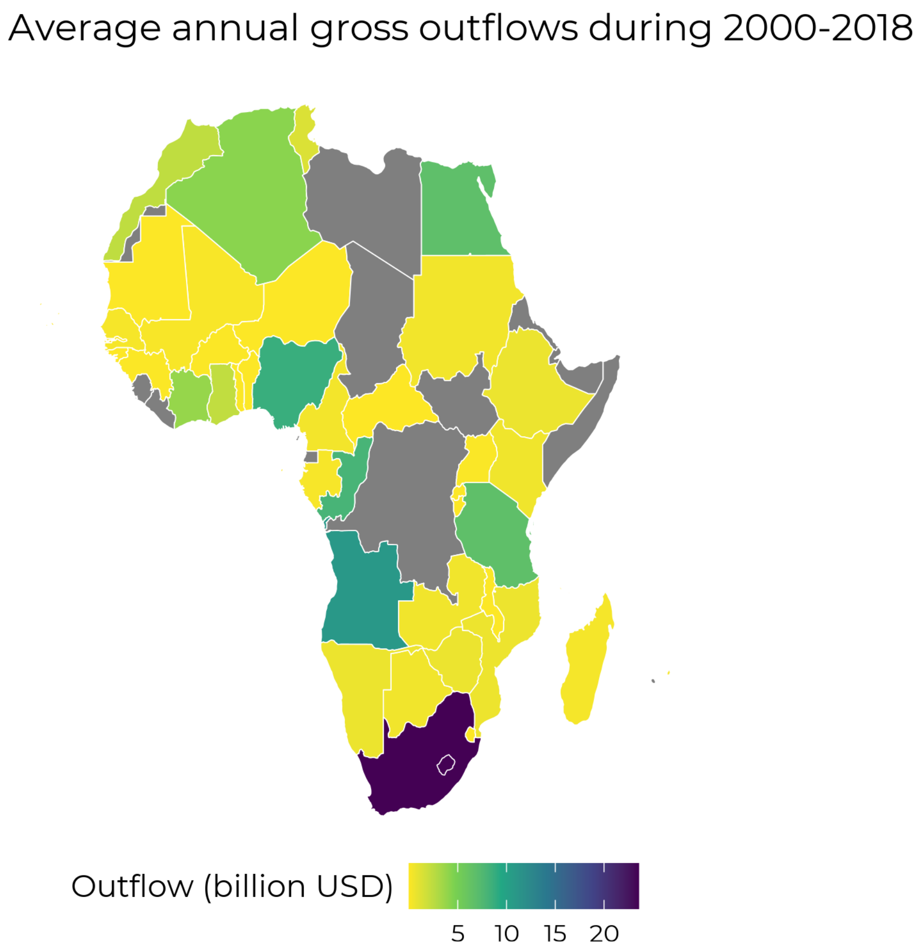


Figure 2: Average annual gross (without reversals) outflows during 2000-2018. Countries with missing data are in grey.

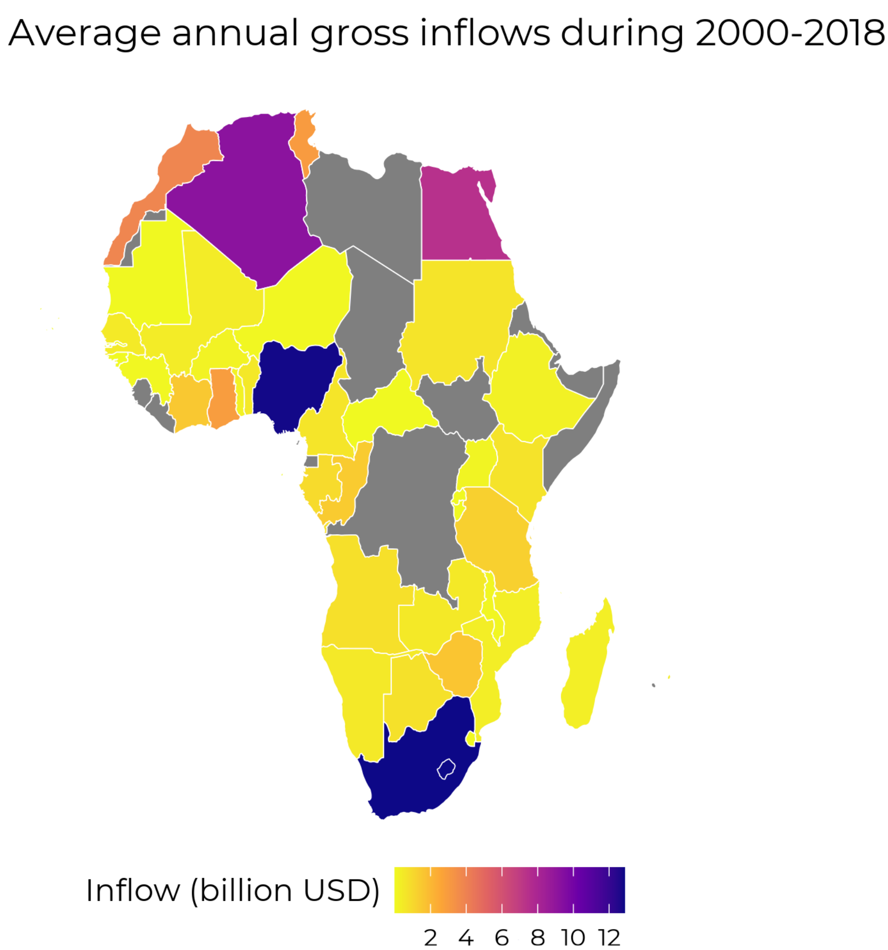


Figure 3: Average annual gross (without reversals) inflows during 2000-2018. The legend refers to the absolute value of inflows. Countries with missing data are in grey.

The countries in grey are the 10 African countries that are missing from the "atlas" because they do not report trade data to Comtrade, or do not provide trade data disaggregated by partner (instead reporting the aggregate amount of trade with the rest of the world). Details on the missing data for these countries are provided in table 2.

Country	Reason for missing data
Chad	No reports to Comtrade since 1995
Democratic Republic of Congo	Does not report to Comtrade
Equatorial Guinea	No disaggregated records by partner
Eritrea	No reports to Comtrade since 2003
Liberia	No reports to Comtrade since 1984
Libya	No disaggregated records by partner
Sierra Leone	No disaggregated records by partner
Somalia	No reports to Comtrade since 1982
South Sudan	Does not report to Comtrade
Western Sahara	Does not report to Comtrade (disputed territory by Morocco)

Table 2: Nature of the missing trade records for the 10 African countries with no data in the "atlas" database.

5.1.2 Transformations

Pre-processing data is an important step in the analytical pipeline of machine learning. The distributions of gross outflows and gross inflows are highly skewed, so the variables are taken in natural logs. Since inflows are represented by negative values in the dataset, the absolute value of the variable `In_GER_Tot_IFF` is taken before logging it. After logging, the distributions of both outcome variables look approximately normal, as shown in figure 4.

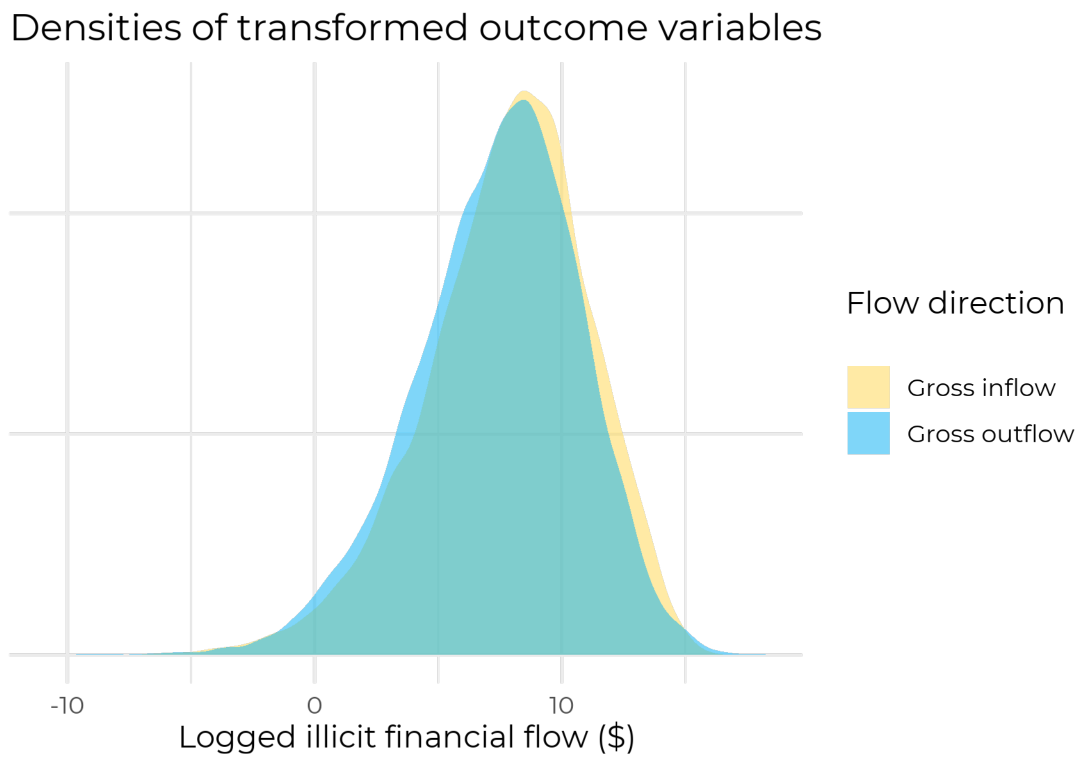


Figure 4: Distribution of logged outcome variable for gross illicit outflows and gross illicit inflows. Data are pooled for all African countries and years 2000-2018.

The distribution of the predictor variables was examined, and skewed continuous variables were transformed to increase the normality of the data. Either the log or the inverse hyperbolic sine transformations were applied. The inverse hyperbolic sine[19] is a transformation that can be used to reduce the skew in data when variables cannot be logged due to presence of zeroes (since $\ln(0)$ is undefined), such as in the case of tariffs or costs to entry. Note that transformation of the predictors is not strictly necessary for tree-based methods since they are scale-invariant methods. However, the features are still scaled to obtain a set of predictors that are consistent across predictive models so that a linear regression model can be estimated as a robustness. Moreover, the log and the inverse hyperbolic sine are monotone transformations, so this will not affect the results of the Random Forest models.

The dichotomous variables (dummy indicators taking a value of either 0 or 1) in the data are:

[19]Defined as $ihs(x) = \ln(x + \sqrt{x^2 + 1})$.

`contig`, `comlang`, `comcol`, `col45`, `RTA`, `FATF`, `cc`, `cci`, `cco`, `di`, `dii`, and `dio`. The other categorical variables in the data are the trichotomous variables `cc` and `di` that measure average restrictions on commercial credits for international trade and average restrictions on direct investment accounts, respectively. Decision trees can handle categorical variables and do not require using one-hot encoding to convert them to a set of dummies (James et al., 2013).

Figure 5 plots the correlation matrix of the continuous variables in the feature space after transformation. Unsuprisingly, the governance variables (i.e., control of corruption, quality of private sector regulatory environment, and respect for the rule of law) are strongly correlated with each other. In a regression setting, this would manifest as a problem of multicollinearity. The Random Forest algorithm implicitly deals with highly correlated variables by selecting a random subset of features each time a tree is grown. Moreover, high scores on the governance variables (which correspond to better governance outcomes) in the *partner* country are strongly negatively correlated with the measures of capital controls (both on inflows and outflows) in destination countries. The entry costs of business (which imposes frictions on trade) are negatively correlated with good governance measures, but positively correlated with the measures of financial secrecy in the partner country.

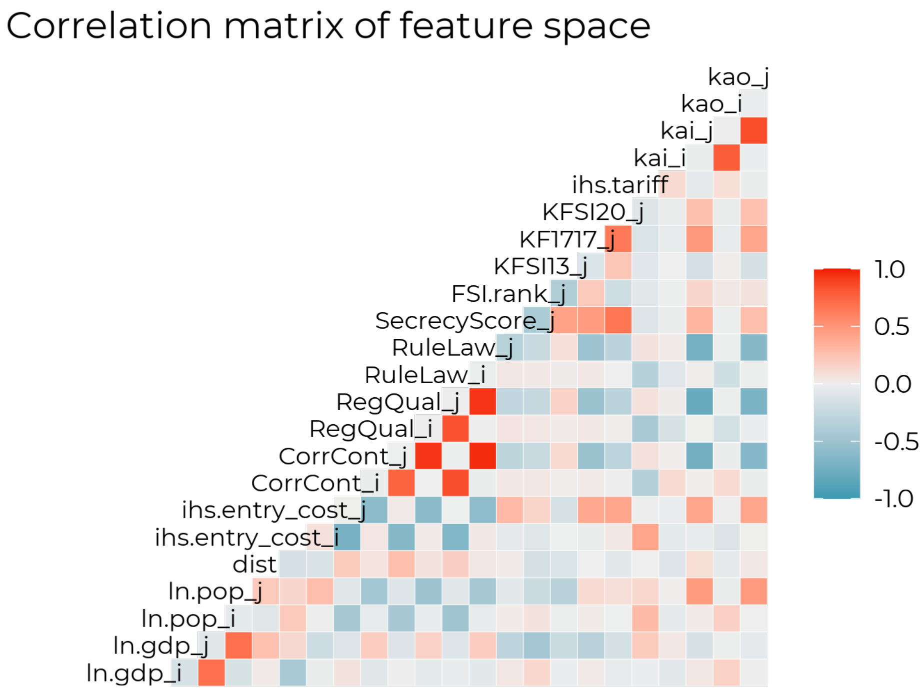


Figure 5: Correlation matrix of the feature space for continuous predictors, after transformations. Unilateral variables relating to reporters and partners have the suffixes `_i` and `_j` respectively. Variables that do not have a suffix are bilateral features.

The approach presented here shows that bilateral illicit trade outcomes can be reliably predicted without using the observed licit trade flow during training, instead using a combination of unilateral (for both reporters and partners) and bilateral country-level characteristics. The observational unit of the features is summarized in table 3.

Reporter	Partner	Brief description
ln.gdp_i	ln.gdp_j	Logged GDP
ln.pop_i	ln.pop_j	Logged population
ihs.entry_cost_i	ihs.entry_cost_j	Costs to enter market
CorrCont_i	CorrCont_j	Control of corruption
RegQual_i	RegQual_j	Quality of private sector regulations
RuleLaw_i	RuleLaw_j	Respect for the rule of law
kai_i	kai_j	Capital controls on inflows
kao_i	kao_j	Capital controls on outflows
cc_i	cc_j	Controls on commercial trade (aggregate)
cci_i	cci_j	Controls on commercial trade inflows
cco_i	cco_j	Controls on commercial trade outflows
di_i	di_j	Controls on direct investment (aggregate)
dii_i	dii_j	Controls on direct investment inflows
dio_i	dio_j	Controls on direct investment outflows
FATF_i	FATF_j	Member of Financial Action Task Force
	SecrecyScore_j	Financial secrecy score
	FSI.rank_j	Rank on FSI
	KFSI13_j	Promotion of tax evasion
	KFSI17_j	Weak anti-money laundering laws
	KFSI20_j	Uncooperative on AML judicial matters
Bilateral		**Brief description**
dist		Distance (km) between countries
contig		Countries share a border
comlang		Common official language
comcol		Share a common colonizer
col45		In a colonial relationship post-1945
rta		Have a Regional Trade Agreement
tariff		Average tariff on imports

Table 3: Observational unit of the unilateral and bilateral country-level features.

5.2 Tuning and training the machine learning models

5.2.1 Random Forest algorithm

Here, a Random Forest (RF) algorithm is used to generate predictions of illicit trade. The constituent element of a RF is an individual decision tree (or regression tree, in this context) (see Breiman (2001)). Regression trees are highly flexible estimators that partition the feature space into distinct and non-overlapping regions, and make predictions based on the mean response of the observations contained in a terminal node, or leaf (Hastie et al., 2017; James et al., 2013). Internal nodes are created by partitioning a specific feature on a specific threshold; parameters which are learned during model training. However, while regression trees can be grown to be very deep in order to fit the training data well, they also tend to be non-robust to making predictions in an unseen test set, since small perturbations to the input data might lead to significantly different predictions.

The innovation of RF rests on averaging predictions from several regression trees that have been grown using bootstrapped samples, and on decorrelating the trees by only considering a random sample of features that can be used to create splits when building the individual trees. By aggregating predictions from bootstrapped trees ("bagging"), this reduces the overall variance of the Random Forest estimator. Furthermore, restricting the (random) number of features that can be used to grow any given individual tree mitigates a potential problem where, in the case that one variable is a strong predictor for the outcome, that variable is repeatedly used in the top split of each tree, which would thus yield a forest of highly correlated trees. Averaging predictions from correlated trees would not result in as much of a reduction in overall variance, and so the RF estimator might still perform poorly on a new test set (Breiman, 2001; Hastie et al., 2017). When growing a forest, the number of random features that should be considered for splitting each tree is a hyperparameter that can be empirically tuned. Here, the tuned hyperparameter governing the maximum number of candidate features to consider when growing each tree is the entire feature set. Therefore, in this application, the tuned RF estimator amounts to a collection of bagged trees.

5.2.2 Parameter tuning and cross-validating

There are 44 African countries in the "atlas" database; at a bilateral level that is aggregated over commodities, the sample size is n = 13,030. After examining the missing data patterns, only the variables relating to the partner-side were retained for the financial integrity variables sourced from the *Financial Secrecy Index* (FSI);[20] which has data coverage for 112 countries. In other words, variables capturing the financial secrecy of the countries that transact with the African countries in the "atlas" are used. The FSI is designed to rank and identify the biggest secrecy jurisdictions responsible for a large share of offshore finance, and thus only 9 African countries appear in the FSI; though it should be noted that it includes Mauritius and Seychelles which have been identified as important conduits of IFFs in Africa (Abugre et al., 2020; High Level Panel on Illicit Financial Flows from Africa, 2015). In total, the feature space contains $K = 42$ predictors.

Observations which do not have data on the features described above are dropped, in order to obtain a complete data-set of $n =$ 5,333, corresponding to 17 African countries. This is not a particularly large number of observations, yet the RF models still achieve good out-of-sample performance. This underscores the advantage of using RF over a more complex, data-hungry, algorithm as a Neural Network (NN). Since the distribution and amount of misinvoiced trade differs across Africa for outflows and inflows, the analysis considers two distinct illicit trade outcomes of interest: the dollar value of misinvoiced trade that results in gross outflows on the one hand, and gross inflows on the other hand. Therefore, there are two outcome vectors: Y_{ijt}^{OUT} contains the labels on outflows, and Y_{ijt}^{IN} contains the labels on inflows.

The sample-splitting approach is described next. The procedure employed here combines a hold-out approach with inner cross-validation. First, the full dataset is split into disjoint training and test sets. The training set will be used for model tuning and evaluation, while the test is used exactly once in the final step to get an estimate of the model's performance on new, unseen, data. The test set is never used for tuning or training – it is held out and set aside until the very end. The data is split into training and test sets by randomly sampling without replacement 80% of the data into the training set, and reserving the other observations for the test set; yielding $n =$ 4,256 for the training sample and $n =$ 1,077 for the test sample.[21] The samples consist of reporter-partner-year observations that are pooled over the years 2000-2018.

[20] The variables are `SecrecyScore, FSIRank, KFSI13, KFSI17, KFSI20`.

[21] The sample-splitting procedure is done with a seed (1509) to ensure reproducibility.

Next, the RF estimator is tuned using k-fold cross-validation – a general procedure where the training set is split into k folds, the model is fit in on $k-1$ folds, and is evaluated on the held-out kth fold. The procedure is repeated k times and the error metric in the held-out sets is averaged to provide an estimate of the model's test error rate on new, unseen, data. The process of tuning the model is accomplished using inner cross-validation on the training set, where the best estimator is the one that maximizes the proportion of variance explained on the held-out validation sets. Then, the *tuned* models are trained on the pooled sample of observations of illicit trade at the reporter-partner-year level in the training set. The final step is to use the trained model exactly once on the previously reserved test set to assess the model's generalization performance. Therefore, model tuning and training is conducted on different data than the data used for overall model assessment.

In this paper, hyperparameters were tuned with 5-fold cross-validation on the training set using a randomized search strategy that randomly sampled, for 100 trials, the hyperparameter space in order to obtain the configuration of RF settings that yields the best performance on the hold-out set. The procedure was repeated twice: once for outflows and once for inflows, and in both cases yielded the same optimized tuning for the RF estimator. The tuning procedure was conducted on the training sample in order to preserve the integrity of the test set. Details on the procedure employed to tune the hyperparameters are provided in section B of the appendix. The Random Forest model was fit using the *scikit-learn* library.[22]

6 Results

6.1 Performance of the models

The tuned Random Forest models on both inflows and outflows were able to predict between 71% and 73% of the variation in illicit trade outcomes in an unseen test set. Table 4 reports two types of performance metrics for both the error of the model (using the Mean Square Error)

[22] Using *RandomForestRegressor* with a random seed of 1509.

and its explanatory power (using R^2): the cross-validation (CV) results obtained during model selection, and the results obtained on the independent test set. As mentioned above, the sample-splitting approach first involves splitting the data in distinct training and test sets, and then using cross-validation for model selection on the *training* set, i.e., conducting inner cross-validation. Model selection was accomplished with 5-fold cross-validation and 100 trials of randomly sampled hyperparameters from the search space, in order to select the model with the best performance on the held-out validation sets (i.e., the k folds used as validation sets during the cross-validation procedure) – and was conducted on the training set. That is, 500 candidate models with different tuning configurations were fit, and the model with the best cross-validated score was chosen.

The cross-validated scores are an estimate of the *tuned* models' expected generalization performance in the population. Since the CV scores were used to tune hyperparameters and choose the best model, the final performance of the model is evaluated on the unseen test set that was reserved at the beginning. When the CV error is used for model selection, it is likely to be a biased estimate of the true error on an independent test set (Varma & Simon, 2006). Therefore, results are also reported for the models' out-of-sample accuracy on the test set – which was never used for training or tuning.

	Gross outflows	**Gross inflows**
Cross-validated R^2 of tuned model	68%	70%
R^2 on unseen test set	71%	73%
Cross-validated MSE of tuned model	3.23	3.04
MSE on unseen test set	3.00	2.87

Table 4: Predictive performance of the RF models on illicit trade outcomes.

The models deliver high statistical performance for both outflows and inflows. The predictive power of the models on the unseen test set is slightly higher than the cross-validated R^2 scores, which suggests that the CV scores have a slightly pessimistic bias in estimating the generalization error. In this case, the true test error is the error that the model chosen through cross-validation would give in an infinite test dataset, i.e., the population. Since cross-validation was used for parameter tuning, the held-out validation samples then become part of the model, since the tuned model is fit on the whole training set. Therefore, an independent test sample is required to correctly measure the models' final performance.

Next, figures 6 and 7 display cross-validated predictions of outflows and inflows for the four countries with the greatest number of observations in the African sample, trained on the pooled sample of African countries. They represent out-of-fold predictions, where each point belongs to exactly one test set, and its prediction is computed with an estimator fitted on the corresponding training set. In other words, these are the predictions that are made on the held-out test folds during cross-validation. The R^2 values displayed in the figures are the scores of the cross-validated predictions, that is, they are the square of the correlation coefficients between the cross-validated predictions and the observed value – though it should be noted that they are not a valid way to measure generalization performance.

The superior predictive performance of the model for South Africa ($R^2 =$ 0.89 for outflows; $R^2 =$ 0.84 for inflows) is striking, even relative to the good performance of the other countries. This might be explained by the fact that South Africa is the most represented country in the training sample ($n =$ 616), and so the model might have trained with a greater emphasis on South Africa.

The cross-validated predictions of outflows and inflows for the remaining 13 countries in the African sample are provided in section C of the appendix. The accuracy of the cross-validated predictions for individual countries is highly dependent on the number of observations available for training. For example, the cross-validated R^2 scores for Angola are 0, because the entire African sample only contains 74 observations for Angola, and only 46 of them were randomly selected in the training set, so it follows that the model will not be able to explain any variation for Angola. Out-of-fold predictions for the remaining countries explain up to 60% and 64% of the variation in country-level illicit outflows and inflows, respectively.

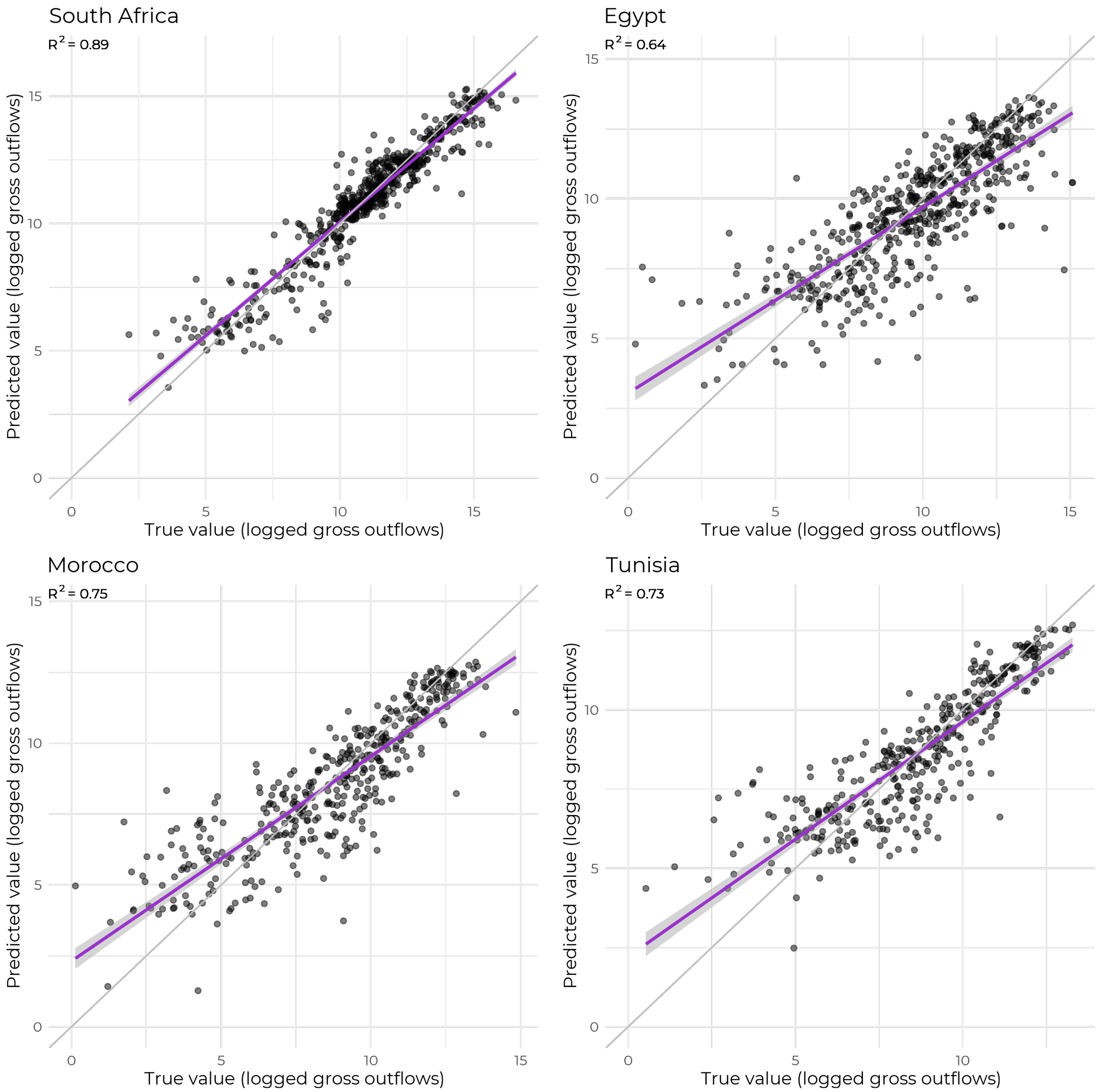


Figure 6: Cross-validated out-of-fold predictions of logged outflows for countries by Random Forest model trained on pooled model of African countries over 2000-2018.

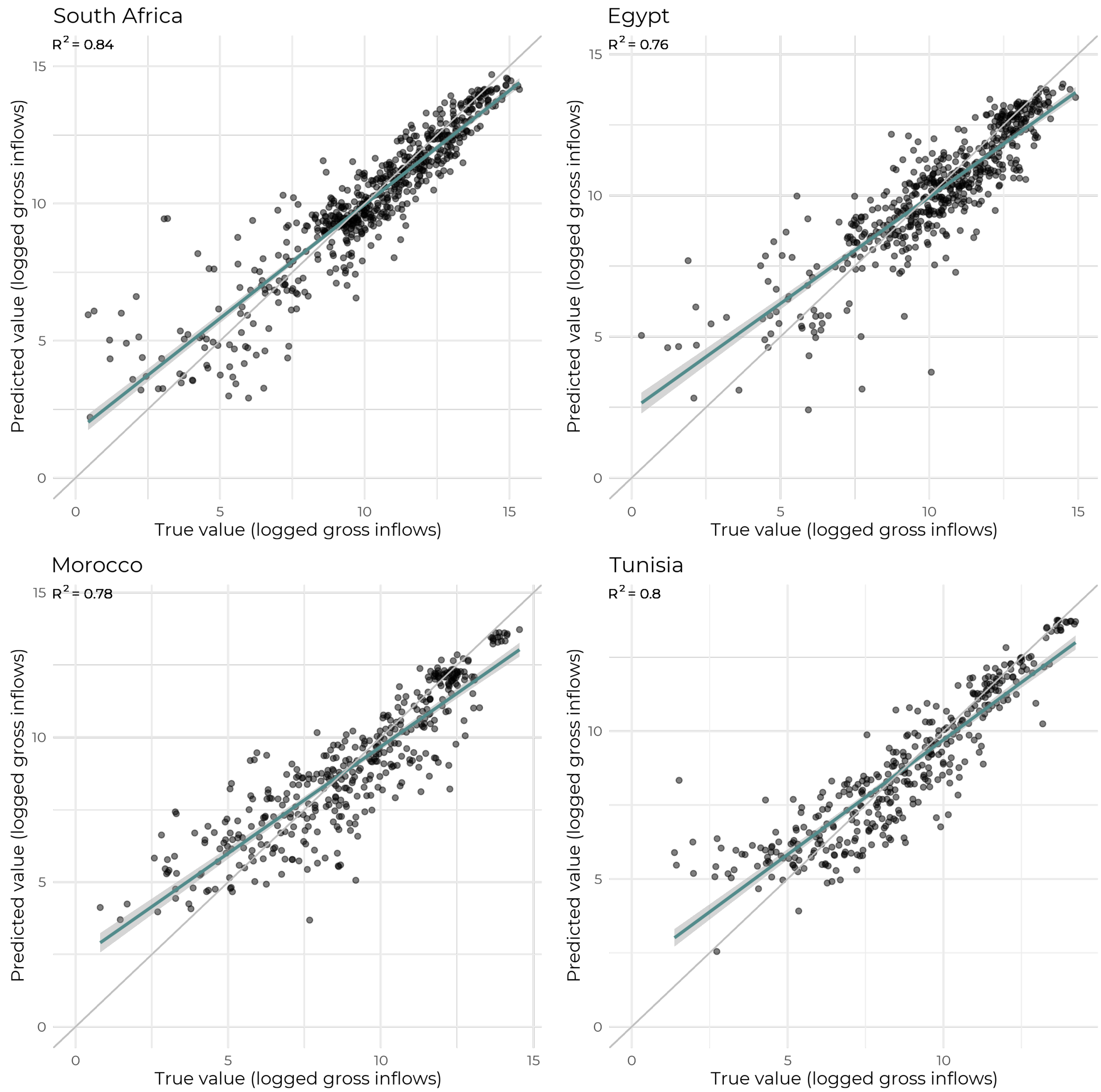


Figure 7: Cross-validated out-of-fold predictions of logged inflows for countries by Random Forest model trained on pooled model of African countries over 2000-2018.

6.2 Assessing statistical significance with placebo trials

Next, an experiment is conducted to assess whether the results reported here are the product of chance. Using a type of randomization inference, the experiment runs placebo trials to evaluate the statistical significance of the results. The individual bilateral transactions are randomly re-assigned to an illicit trade label; that is, the rows of the design matrix X are reshuffled and randomly paired with the vector of illicit trade outcomes Y. The randomization preserves the nature of the bilateral partnership, that is, the observational unit of the shuffled data retains the same given reporter, partner, and year as in the true data.

The RF model is retrained on the placebo bilateral identities, and is evaluated on the independent test set. This experiment is repeated 100 times, where in each trial the identity of the transacting countries is shuffled and the model is re-trained on the fake data. The Mean Square Error denoting the accuracy of these placebo models on the true test sets are collected, and their distribution is displayed in figure 8.

The MSE in the test set of the models trained on the correct data are indicated by the vertical lines on the graph. The fact that their MSEs are in the tails of the distribution of the placebo scores suggest that the results presented in this paper are unlikely to have arisen by chance. This suggests that the specific bilateral identifiers in the data capture some structure of the patterns of illicit trade. A specific combination of transacting partners – encapsulating the specific unilateral characteristics of each country (e.g., GDP, entry costs, etc.) – is thus highly predictive of illicit trade outcomes. In other words, there is some underlying structure, perhaps regarding the relative development level of each partner (e.g., countries in different income brackets), or the relative attractiveness of countries for conducting illicit affairs (e.g., the Walker type variables), that explains much of the variation in illicit trade. While it would not be prudent to infer the specific structure about the types of country combinations that are most associated with illicit trade, this finding nonetheless provides suggestive evidence that the variables relating to push-pull gravitational factors, the illicit premium dimension, and the regulatory environment that were identified earlier collectively have high predictive power for illicit trade.

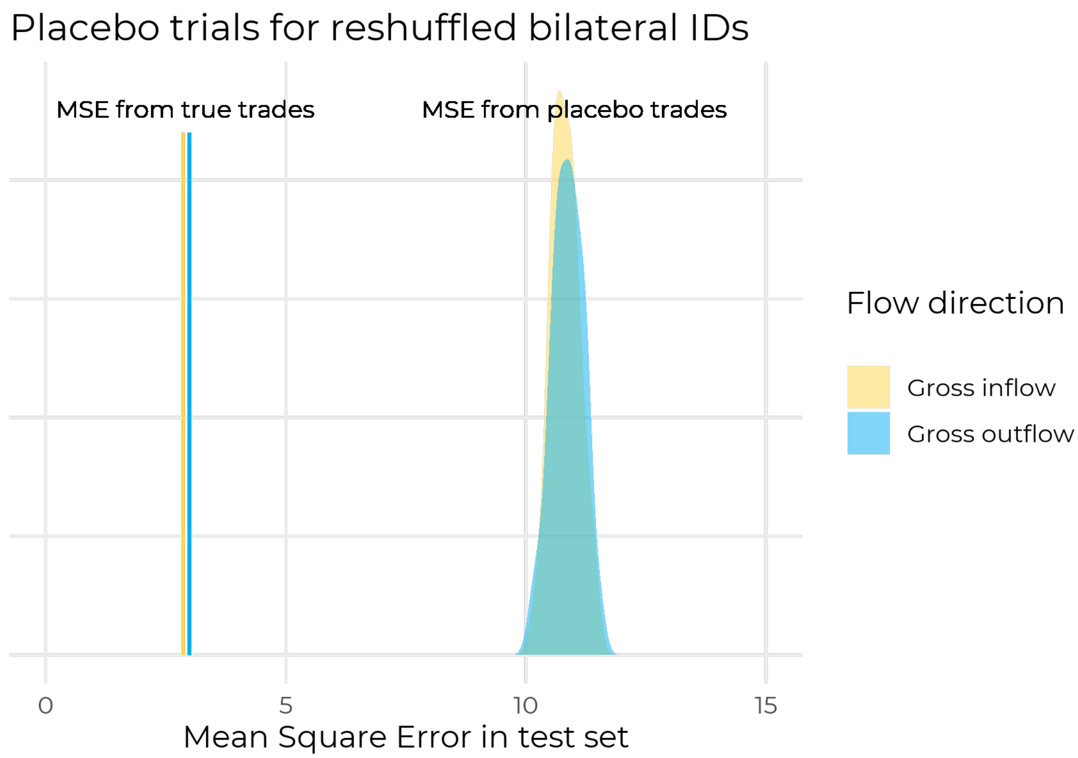


Figure 8: Mean Square Error on the independent test set of model trained on true bilateral transactions, and of 100 placebo models trained on randomly reshuffled bilateral trades.

6.3 Assessing the models' generalization performance

The degree to which the models "travel across borders" is now evaluated for other income groups. Observations from the "atlas" database for reporter countries in different income brackets are used as new test sets. Using the World Bank's 2020 classification for income groups, countries in the "atlas" database are classified as either low income or lower-middle countries (LMIC) or high income (HIC) countries.[23] The performance of the models trained on African countries is then evaluated on these different income group samples. The LMIC sample and the HIC sample represent tests of increasing difficulty because most African countries are classified as LMIC countries. Thus, the test on the LMIC sample – which contains 63 reporters – will evaluate

[23]According to this classification, low income countries are defined as those with a GNI per capita of $1,035 or less in 2019, lower-middle income countries are those with a GNI per capita between $1,036 and $4,045, and high income countries are those with a GNI per capita above $12,536. Low and lower-middle income countries are grouped together in the LMIC set.

the performance of models that were partly trained on countries from that group. By contrast, Mauritius is the only country that appears both in the African sample and in the HIC sample, which contains 58 countries.[24] Therefore, evaluating the models that were trained on African countries by using the sample of high income countries as a test set provides an indication of the extent to which the performance of the models can be expected to generalize to new data on other countries.

The models are evaluated directly on the LMIC and HIC country group samples, and also using 5-fold cross-validation; results are reported in table 5. The R^2 scores that are obtained by evaluating the models directly on the HIC sample can broadly be interpreted as R^2 scores on an unseen test set (notwithstanding the information leakage cause by the inclusion of Mauritius). By contrast, the R^2 scores that are obtained from evaluating the model on the LMIC set would probably overestimate the predictive performance of the models on new data because they would overfit to the training data (which includes African LMICs). Indeed, the cross-validated R^2s for the LMIC group are much lower than the R^2s obtained by direct evaluation on the LMIC set. The opposite happens for the group of high income countries: the tuned models applied directly to the HIC sample explain a lower proportion of the data (giving a one-time snapshot of the performance) but the average R^2s over cross-validation folds (which give a fuller picture of the models' potential to generalize) is higher, suggesting that the method presented in this paper is robust and can be scaled to other countries.

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

Table 5: Estimates of the models' ability to generalize across borders.

[24] There are other African countries that are classified as high income, but these are not present in the complete sample that is used to train the models.

6.4 Robustness checks

A robustness check is conducted to verify that a simpler linear regression model could not have done as well or better than the RF models presented in this paper. Linear regression models are estimated on the training data, and their performance is evaluated on the test set. When the entire feature set presented in this paper ($K = 42$) is used as the set of explanatory variables in a regression model, the fully specified linear model of misinvoiced trade is:

$$\begin{aligned} \log Y_{ijt} = \alpha + &\boldsymbol{Gravity\beta} + \boldsymbol{Governance\gamma} \\ &+ \boldsymbol{RegulEnvironment\lambda} + \boldsymbol{FinancialIntegrity\pi} + \epsilon_{ijt} \end{aligned} \tag{1}$$

where $\log Y_{ijt}$ is the outcome variable, e.g., gross outflows from i to j (`GER_Tot_IFF`), $\boldsymbol{\beta}$ is a vector of parameters on traditional gravity variables, which includes proxies for size, geographical distance, cultural distance, barriers to trade, and trade facilitation; $\boldsymbol{\gamma}$ is a vector of parameters on governance variables; $\boldsymbol{\lambda}$ is a vector of parameters on variables related to the regulatory environment that capture potential incentives to misinvoice trade; and $\boldsymbol{\pi}$ is a vector of parameters on variables related to the integrity of a country's financial system and the government's tolerance for money laundering and tax evasion.

However, due to the multicollinearity of the covariates, the design matrix of this model is rank deficient, that is, there is not enough information contained in the data to meaningfully estimate the full model. This further suggests that the implicit regularization that occurs with machine learning models is valuable. Nonetheless, the fully specified model is estimated and used to make predictions in order to have a benchmark – however, predictions from such as model will be misleading as they will overfit the training data.

Therefore, two additional reduced form models are estimated using a subset of variables that are theoretically important and empirically relevant, for gross outflows (`GER_Tot_IFF`) and gross inflows (`In_GER_Tot_IFF`) separately. These variables were chosen using domain knowledge and the theoretical insights developed in section 3; but there is no guarantee that selecting these variables *ex ante* will produce a model with high explanatory power. The model specifications differ for outflows and inflows, given that directionality is important. The reduced form models are presented below:

$$\begin{aligned}\log \texttt{GER\_Tot\_IFF}_{ijt} = \alpha &+ \beta_1 \log \texttt{GDP}_{it} + \beta_2 \log \texttt{GDP}_{jt} + \\ &+ \beta_3 \texttt{comlang}_{ij} + \beta_4 \texttt{col45}_{ij} + \beta_5 \texttt{RTA}_{ij} \\ &+ \gamma_1 \texttt{CorrCont}_{it} + \gamma_2 \texttt{CorrCont}_{jt} + \gamma_3 \texttt{RegQual}_{jt} \\ &+ \lambda_1 \texttt{FATF}_{it} + \lambda_2 \texttt{FATF}_{jt} \\ &+ \pi_1 \texttt{tariff}_{ijt} + \pi_2 \texttt{kao}_{it} + \pi_3 \texttt{kai}_{jt} + \epsilon_{ijt}\end{aligned} \tag{2}$$

$$\begin{aligned}\log \texttt{In\_GER\_Tot\_IFF}_{ijt} = \alpha &+ \beta_1 \log \texttt{GDP}_{it} + \beta_2 \log \texttt{GDP}_{jt} + \\ &+ \beta_3 \texttt{comlang}_{ij} + \beta_4 \texttt{col45}_{ij} + \beta_5 \texttt{RTA}_{ij} \\ &+ \gamma_1 \texttt{CorrCont}_{it} + \gamma_2 \texttt{CorrCont}_{jt} + \gamma_3 \texttt{RegQual}_{it} \\ &+ \lambda_1 \texttt{FATF}_{it} + \lambda_2 \texttt{FATF}_{jt} \\ &+ \pi_1 \texttt{tariff}_{ijt} + \pi_2 \texttt{kai}_{it} + \pi_3 \texttt{kao}_{jt} + \epsilon_{ijt}\end{aligned} \tag{3}$$

Both models control for gravity variables that proxy the size of an economy and cultural distance, but not geographical distance since is expected to matter less for flows of money than of merchandise. If the invoice associated with a commodity shipment that is presented to customs is manipulated to illicitly transfer money, then the illicit part of that payment should not reflect transport costs.

Further variables are added to capture the illicit premium of a jurisdiction as an attractive haven to conceal funds. The regulatory quality of the *destination* country's private sector (`RegQual`) can offer a prospective misinvoicer relative financial stability to safeguard their assets, and the existence of a Regional Trade Agreement between countries will facilitate transactions. The variable on corruption, `CorrCont`, is included for both reporter and partner countries since it can influence the propensity for outflows and inflows of illicit trade, since customs officials can be suborned to either inflate or deflate an invoice. Likewise, dummies (`FATF`) are included denoting whether each of the reporter or partner are members of the Financial Action Task Force, since the legal recommendations the FATF makes and the policy framework it encourages are designed to halt illicit finance in any direction.

Finally, variables that capture the incentives to misinvoice trade in order to commit regulatory or market abuse are included. The variable `tariff` represents the average tax rate imposed by country i on imports from country j. For the model that estimates outflows, an aggregate measure of capital controls on outflows (`kao`) in the reporter country (the source of IFFs), and capital controls on inflows (`kai`) in the partner country (the destination of IFFs), are used. The converse is applied for the model that estimates inflows, since the reporter country i is now a destination for IFFs coming from partner j (the source).

The reduced models are estimated on the training set – a pooled sample of bilateral illicit trade from African countries during 2000-2018 – using Ordinary Least Squares (OLS). The model coefficients are reported in table 8 of the appendix. Performance metrics are reported in table 6 below for the reduced form model and the fully specified model. The more complex model performs better than the reduced form model, both in the training set and in the test set. However, there is no performance improvement for the full model between the training and the test set. The baseline linear model was estimated on the variables that were identified above as theoretically and empirically important using theoretical insights from the literature and knowledge of the policy space in IFFs. By contrast, the full linear model was estimated on all of the predictors in the covariate space in order to benchmark the performance of the Random Forest algorithm. Both linear models performed worse than the RF models, which explained between 71% and 73% of the variation of illicit trade in an independent test set.

Tree-based methods are highly flexible regressors compared to linear regression models. In the case of trade-based misinvoicing, they provide superior predictive performance and are better able to recover the underlying structure of the data. The fact that tree-based methods outperform classical regression methods is indicative of a complex and non-linear relationship between illicit trade flows outcomes and the features. Consequently, Random Forest models are better-suited than linear models to provide predictions of illicit trade outcomes in the absence of data on the underlying trade flow.

	Reduced model		**Full model**	
	Gross outflows	Gross inflows	Gross outflows	Gross inflows
R^2 on training set	43%	40%	58%	53%
R^2 on test set	44%	39%	58%	57%
MSE on training set	6.00	6.34	4.20	4.73
MSE on test set	5.73	6.44	4.28	4.61

Table 6: Predictive performance of the linear regression models on illicit trade outcomes. The full model is rank deficient.

7 Discussion

In this paper, the predictive performance of machine learning models is assessed as a proof-of-concept, and the paper shows that machine learning models that are not trained on trade data are nonetheless able to account for much of the variation in illicit trade outcomes. The predictive performance of the models is estimated on held-out test datasets in order to assess how these models would perform on new, unseen, data. One specific application of this method is as follows. A researcher interested in predicting illicit trade for a low income country that does not report to Comtrade could assemble available data on that country's unilateral characteristics as a first step. Yet, the models presented here are also trained with dyadic features that require knowledge of the trading partner's characteristics – which would of course not be provided by the Comtrade non-reporter. In other words, if trade data is not reported to Comtrade by the low income country of interest, then the identity of that country's trading partners is also not reported. However, a crucial feature of the "double entry" accounting system of international trade statistics can be exploited to work around this problem and to learn the identity of the non-reporting country's trading partners. Even if some countries might not (consistently) report to Comtrade, the partner country on the other side of the trade might, since bilateral trade transactions should be recorded twice.

As an example, the Democratic Republic of Congo (DRC) does not provide declarations to Comtrade, and so it will not report its imports of commodities from a trading partner, say, France. But the value of French exports to the DRC in any given year is observable, since the declarations from the other side of the transaction are provided by French customs authorities, and France is a reporter to Comtrade. Thus, this strategy provides information on the partner's unilateral characteristics (e.g., France's financial secrecy score) and on the bilateral features of the dyad (e.g., whether France and DRC share a common official language). Therefore, a researcher could, for any given country i with missing data from the "atlas" database, use Comtrade to find the mirror declarations from countries that report imports from or exports to the missing country i. These mirror declarations then yield the specific dyads (e.g., USA and DRC, South Africa and DRC, etc.) that i is a member of. Then, information on the unilateral and bilateral characteristics of the dyad can be collected and used as the features of an out-of-sample test set, and can then be used to generate predictions by fitting the tuned models presented here. Therefore, the method described in this paper not only demonstrates high predictive potential, but it can also be used in a specific application to generate country-specific results for the countries that are missing from the "atlas" database because they do not report to Comtrade. The task of augmenting the "atlas" database using this method is left to future work. The focus of the paper here, as a necessary preliminary, is to demonstrate that this method can be expected to perform well.

This paper contributes to a broader literature that seeks to measure missing outcomes on economic development by using innovative quantitative methods that exploit already available data. Studies have focused on capturing development-related outcomes such as poverty and economic growth to mitigate the problems of data scarcity in developing countries by using data on luminosity that is passively collected by satellites (see, e.g., Chen and Nordhaus (2011), Henderson et al. (2012), N. Jean et al. (2016), and Pinkovskiy and Sala-i-Martin (2014)). Here, I show that missing outcomes on illicit trade – another important measure for economic well-being – can also be reliably recovered using available country-level characteristics that have relatively low collection costs for researchers.

The approach presented here has several limitations. First, contrary to data that is passively recorded by satellites or to financial data that is routinely collected by financial actors, the features employed here require some assembling by researchers, and some measures like Gross Domestic Product can also be affected by the weaker statistical capacities of developing countries. Nonetheless, some variables have already been collected and are time-invariant (e.g., distance) and others are provided by publicly available databases with broad country coverage that are updated yearly (e.g., *Worldwide Governance Indicators*). Moreover, in the category of variables collected by national statistical offices in poor countries, information on GDP is arguably the least likely to be missing – certainly compared to customs declarations.

Second, the mirror strategy that I describe above to predict missing data for Comtrade non-reporters from their partners' declarations will underestimate the true extent of intra-African illicit trade in this case, and illicit trade between developing countries in general. Importantly, the data cannot be assumed to be missing at random. Here we must distinguish between two types of biases occasioned by missing data. Data will be missing from the "atlas" database because some countries do not provide customs declarations to Comtrade, but also due to unobservable parameters that cannot be fully accounted for even if there were complete information on trade flows. Thus, in the case of illicit trade flows, the value of the data that is missing (trade flows) will be related to the reason why it is missing (trade misinvoicing). This is a conspicuous and pervasive problem across studies of illicit economic activity more broadly.

Third, caution should be exercised when employing this approach as a method for unit-level imputation and when interpreting the resulting predictions of specific reporter-partner-year illicit trade transactions. A more prudent strategy would be to first use this method to fill out bilateral gaps and second to aggregate the predictions of illicit trade for reporters over partners, years, or both; this approach is likely to be more robust and to enable greater confidence in the resulting interpretations.

8 Conclusion

This paper presents a new strategy to address the problems of missing data on economic outcomes in data-constrained developing countries, by using machine learning algorithms to predict bilateral illicit trade outcomes without requiring the underlying customs declarations of the observed trade flow. Missing or poor quality data in low income countries is a persistent problem due to weak administrative systems for statistical reporting. This complicates the analysis of development-related outcomes that depends on data from national statistical offices in poor countries, including the study of trade misinvoicing – the illicit practice of manipulating trade invoices to obscure transfers of money – which relies on recorded trade declarations by national customs authorities. The paucity of available data on commodity trade flows compounds the prejudice for African countries who are particularly vulnerable to illicit financial flows.

The "atlas" database, which offers the widest existing country coverage of trade misinvoicing estimates, is missing data for 10 African countries who do not report international trade statistics. Here, I originate an approach to predicting illicit trade that does not require official statistics compiled by governments in low income countries for training. A Random Forest algorithm is used to train models on a sample of African countries to predict trade misinvoicing using only data on country-level characteristics that are readily available. The models are trained using unilateral and bilateral features that are either easily observed or in publicly available databases. Results show that the models are able to explain between 70% and 73% of the variation in illicit trade outcomes. Placebo trials are conducted to demonstrate the statistical significance of the results, and the generalization performance of the models is characterized using an experiment that tests how well the models "travel" beyond Africa. The results show that the superior predictive performance of the machine learning models is unlikely to be the product of chance, suggesting instead that the machine learning models are able to detect meaningful structure in the dyadic nature of countries' bilateral relationships that is predictive of illicit trade. The paper substantively contributes to scholarship on illicit finance by developing a novel application of machine learning based on researcher-compiled aggregate economic data instead of routinely collected transaction-level financial data. Finally, the results demonstrate the promise of machine learning as an imputation tool to augment existing measures of development-related outcomes in the data-scarce settings of developing countries.

References

Abugre, C., Cobham, A., Lépissier, A., Etter-Phoya, R., Meinzer, M., Monkam, N., & Mosioma, A. (2020). *Vulnerability and Exposure to Illicit Financial Flows risk in Africa* (tech. rep.). Tax Justice Network. Chesham, UK.

African Union. (2015). *Decisions, Declarations, and Resolutions of the Twenty-Fourth Ordinary Session of the African Union*. Assembly/AU/Dec. 546-568(XXIV).

Alexandre, C., & Balsa, J. (2016). Client Profiling for an Anti-Money Laundering System.

Andersen, J. J., Johannesen, N., Lassen, D. D., & Paltseva, E. (2017). Petro Rents, Political Institutions, and Hidden Wealth: Evidence from Offshore Bank Accounts. *Journal of the European Economic Association*, *15*(4), 818–860. https://doi.org/10.1093/jeea/jvw019

Anderson, J. E. (1979). A Theoretical Foundation for the Gravity Equation. *The American Economic Review*, *69*(1), 106–116.

Anderson, J. E., & Wincoop, E. V. (2003). Gravity with Gravitas: A Solution to the Border Puzzle. *The American Economic Review*, *93*(1).

Athey, S., Imbens, G., Pham, T., & Wager, S. (2017). Estimating Average Treatment Effects: Supplementary Analyses and Remaining Challenges. *American Economic Review*, *107*(5), 278–281. https://doi.org/10.1257/aer.p20171042

Badal-Valero, E., Alvarez-Jareño, J. A., & Pavía, J. M. (2018). Combining Benford's Law and machine learning to detect money laundering. An actual Spanish court case. *Forensic Science International*, *282*, 24–34. https://doi.org/10.1016/j.forsciint.2017.11.008

Baker, R. W. (2005). *Capitalism's Achilles Heel: Dirty Money and How to Renew the Free-Market System*. John Wiley & Sons.

Batarseh, F. A., Gopinath, M., & Monken, A. (2020). Artificial Intelligence Methods for Evaluating Global Trade Flows. *International Finance Discussion Paper*, *2020*(1296). https://doi.org/10.17016/IFDP.2020.1296

Batarseh, F. A., Gopinath, M., Nalluru, G., & Beckman, J. (2019). Application of Machine Learning in Forecasting International Trade Trends.

Beegle, K., Christiaensen, L., Dabalen, A., & Gaddis, I. (2016). The State of Data for Measuring Poverty. In *Poverty in a Rising Africa* (pp. 25–56). World Bank Group. https://doi.org/10.1596/978-1-4648-0723-7_ch1

Bergstrand, J. H. (1985). The Gravity Equation in International Trade: Some Microeconomic Foundations and Empirical Evidence. *The Review of Economics and Statistics*, *67*(3), 474. https://doi.org/10.2307/1925976

Bhagwati, J. (1964). On the Underinvoicing of Imports. *Bulletin of the Oxford University Institute of Economics & Statistics*, *27*(4), 389–397. https://doi.org/10.1111/j.1468-0084.1964.mp27004007.x

Blankenburg, S., & Khan, M. (2012). Governance and Illicit Flows. In P. Reuter (Ed.), *Draining Development? Controlling Flows of Illicit Funds from Developing Countries* (pp. 21–68). World Bank Group.

Brainard, L. (2021). Supporting Responsible Use of AI and Equitable Outcomes in Financial Services.

Breiman, L. (2001). Statistical Modeling: The Two Cultures. *Statistical Science*, *16*(3), 199–215.

Buehn, A., & Eichler, S. (2011). Trade Misinvoicing: The Dark Side of World Trade: TRADE MISINVOICING. *The World Economy*, *34*(8), 1263–1287. https://doi.org/10.1111/j.1467-9701.2011.01375.x

Canhoto, A. I. (2020). Leveraging machine learning in the global fight against money laundering and terrorism financing: An affordances perspective. *Journal of Business Research*, *131*, 441–452. https://doi.org/10.1016/j.jbusres.2020.10.012

Carrère, C., & Grigoriou, C. (2015). *Can mirror data help to capture informal international trade?* (Tech. rep.). Fondation pour les Études et Recherches sur le Développement International. Clermont-Ferrand, France.

Chandler, D., Levitt, S. D., & List, J. A. (2011). Predicting and Preventing Shootings among At-Risk Youth. *American Economic Review*, *101*(3), 288–292. https://doi.org/10.1257/aer.101.3.288

Chen, X., & Nordhaus, W. D. (2011). Using luminosity data as a proxy for economic statistics. *Proceedings of the National Academy of Sciences*, *108*(21), 8589–8594. https://doi.org/10.1073/pnas.1017031108

Christensen, J. (2012). The hidden trillions: Secrecy, corruption, and the offshore interface. *Crime, Law and Social Change*, *57*(3), 325–343. https://doi.org/10.1007/s10611-011-9347-9

Christian Aid. (2009). *False profits: Robbing the poor to keep the rich tax-free*.

Clausing, K. A. (2003). Tax-motivated transfer pricing and US intrafirm trade prices. *Journal of Public Economics*, *87*(9-10), 2207–2223. https://doi.org/10.1016/S0047-2727(02)00015-4

Cobham, A., Bernardo, J. G., Palansky, M., & Mansour, M. B. (2020). *The State of Tax Justice 2020: Tax Justice in the time of COVID-19*. Tax Justice Network.

Cobham, A., & Janský, P. (2020). *Estimating illicit financial flows: A critical guide to the data, methodologies and findings*.

Cobham, A., Janský, P., & Prats, A. (2014). Estimating Illicit Flows of Capital via Trade Mispricing: A Forensic Analysis of Data on Switzerland. *Center for Global Development Working Paper*, *350*.

Collin, M. (2019). Illicit Financial Flows: Concepts, Measurement, and Evidence. *The World Bank Research Observer*, *35*(1), 44–86. https://doi.org/10.1093/wbro/lkz007

Conte, M., Cotterlaz, P., & Mayer, T. (2021). The CEPII Gravity Database.

Davies, R. B., Martin, J., Parenti, M., & Toubal, F. (2018). Knocking on Tax Haven's Door: Multinational Firms and Transfer Pricing. *The Review of Economics and Statistics*, *100*(1), 120–134. https://doi.org/10.1162/REST_a_00673

Deardorff, A. V. (1998). Determinants of Bilateral Trade: Does Gravity Work in a Neoclassic World? In J. A. Frankel (Ed.), *The Regionalization of the World Economy* (pp. 7–32). Chicago University Press.

Deloitte. (2018). *The case for artificial intelligence in combating money laundering and terrorist financing - A deep dive into the application of machine learning technology*.

Devarajan, S. (2013). Africa's Statistical Tragedy. *Review of Income and Wealth*, *59*, S9–S15. https://doi.org/10.1111/roiw.12013

Disdier, A. C., & Head, K. (2008). The puzzling persistence of the distance effect on bilateral trade. *The Review of Economics and Statistics*, *90*(1), 37–48. https://doi.org/10.1162/rest.90.1.37

Efimov, D., Xu, D., Kong, L., Nefedov, A., & Anandakrishnan, A. (2020). Using generative adversarial networks to synthesize artificial financial datasets.

Eichengreen, B., & Irwin, D. A. (1995). Trade blocs, currency blocs and the reorientation of world trade in the 1930s. *Journal of International Economics*, *38*(1-2), 1–24. https://doi.org/10.1016/0022-1996(95)92754-P

FACTI. (2021). *Report of the High Level Panel on International Financial Accountability, Transparency and Integrity for Achieving the 2030 Agenda*. Financing for Sustainable Development Office, Department of Economic and Social Affairs, United Nations.

FATF. (2019). *Terrorist Financing Risk Assessment Guidance* (tech. rep.). Financial Action Task Force. Paris, France.

FATF. (2021). *Opportunities and Challenges of New Technologies for AML/CFT* (tech. rep.). Financial Action Task Force. Paris, France.

FATF & Edgmont Group. (2020). *Trade-Based Money Laundering Risk Indicators* (tech. rep.). Financial Action Task Force. Paris, France.

Feenstra, R. C. (2004). *Advanced international trade: Theory and evidence*. Princeton University Press.

Fernández, A., Klein, M. W., Rebucci, A., & Schindler, M. (2015). Capital Control Measures: A New Dataset. *IMF Working Paper*, *WP/15/80*, 2–30.

Ferwerda, J., Kattenberg, M., Chang, H.-H., Unger, B., Groot, L., & Bikker, J. A. (2013). Gravity models of trade-based money laundering. *Applied Economics*, *45*(22), 3170–3182. https://doi.org/10.1080/00036846.2012.699190

Ferwerda, J., van Saase, A., Unger, B., & Getzner, M. (2020). Estimating money laundering flows with a gravity model-based simulation. *Scientific Reports*, *10*(1), 1–11. https://doi.org/10.1038/s41598-020-75653-x

Findley, M., Nielson, D., & Sharman, J. (2020). *Anti-corruption measures* (FACTI Background Paper 5). United Nations High Level Panel on Financial Accountability Transparency and Integrity.

Fisman, R., & Wei, S. J. (2004). Tax rates and tax evasion: Evidence from "missing imports" in China. *Journal of Political Economy*, *112*(2), 471–496. https://doi.org/10.1086/381476

Fisman, R., & Wei, S. J. (2009). The smuggling of art, and the art of smuggling: Uncovering the illicit trade in cultural property and antiques. *American Economic Journal: Applied Economics*, *1*(3), 82–96. https://doi.org/10.1257/app.1.3.82

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, *486*(3-5), 75–174. https://doi.org/10.1016/j.physrep.2009.11.002

Gara, M., Giammatteo, M., & Tosti, E. (2019). Magic mirror in my hand... How trade mirror statistics can help us detect illegal financial flows. *World Economy*, *42*(11), 3120–3147. https://doi.org/10.1111/twec.12840

Ge, Q., Huang, X., Fang, S., Guo, S., Liu, Y., Lin, W., & Xiong, M. (2020). Conditional Generative Adversarial Networks for Individualized Treatment Effect Estimation and Treatment Selection. *Frontiers in Genetics*, *11*, 585804. https://doi.org/10.3389/fgene.2020.585804

Ghosh, A., & Qureshi, M. (2016). What's In a Name? That Which We Call Capital Controls. *IMF Working Papers*, *16*(25), 2–44. https://doi.org/10.5089/9781498332835.001

Gonzalez-Lira, A., & Mobarak, A. M. (2019). Slippery Fish: Enforcing Regulation under Subversive Adaptation. *IZA Discussion Paper Series*, *12179*, 1–64.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems* (Vol. 27). Curran Associates, Inc.

Gopinath, M., Batarseh, F. A., & Beckman, J. (2020). Machine learning in gravity models: An application to agricultural trade. *NBER Working Paper Series*, *27151*, 2–36.

Gupta, M., Gao, J., Aggarwal, C. C., & Han, J. (2014). Outlier Detection for Temporal Data: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, *26*(9), 2250–2267. https://doi.org/10.1109/TKDE.2013.184

Haberly, D., & Wójcik, D. (2014). Regional blocks and imperial legacies: Mapping the global offshore FDI network. *Economic Geography*, *91*(3), 251–280. https://doi.org/10.1111/ecge.12078

Haberly, D., & Wójcik, D. (2015). Tax havens and the production of offshore FDI: An empirical analysis. *Journal of Economic Geography*, *15*(1), 75–101. https://doi.org/10.1093/jeg/lbu003

Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The Elements of Statistical Learning - Data Mining, Inference, and Prediction*. Springer Series in Statistics.

Head, K., & Mayer, T. (2014). Gravity Equations: Workhorse, Toolkit, and Cookbook. In G. Gopinath, E. Helpman, & K. Rogoff (Eds.), *Handbook of International Economics* (pp. 131–195, Vol. 4). North Holland. https://doi.org/10.1016/B978-0-444-54314-1.00003-3

Henderson, J. V., Storeygard, A., & Weil, D. N. (2012). Measuring Economic Growth from Outer Space. *American Economic Review*, *102*(2), 994–1028. https://doi.org/10.1257/aer.102.2.994

High Level Panel on Illicit Financial Flows from Africa. (2015). *Illicit Financial Flows: Report of the High Level Panel on Illicit Financial Flows from Africa*. United Nations Economic Commission for Africa and African Union.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning - with Applications in R* (Vol. 102). Springer Texts in Statistics.

Javorcik, B. S., & Narciso, G. (2008). Differentiated products and evasion of import tariffs. *Journal of International Economics*, *76*(2), 208–222. https://doi.org/10.1016/j.jinteco.2008.07.002

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, *353*(6301), 790–794. https://doi.org/10.1126/science.aaf7894

Jean, S., & Mitaritonna, C. (2010). Determinants and Pervasiveness of the Evasion of Customs Duties. *CEPII Working Paper*, *26*, 1–57.

Jerven, M. (2009). The relativity of poverty and income: How reliable are African economic statistics? *African Affairs*, *109*(434), 77–96. https://doi.org/10.1093/afraf/adp064

Jerven, M. (2013). *Poor Numbers - How We Are Misled by African Development Statistics and What to Do about It*. Cornell University Press.

Jerven, M. (2016). Trapped between tragedies and miracles: Misunderstanding African economic growth. *Development Policy Review*, *34*(6), 911–915. https://doi.org/10.1111/dpr.12176

Jerven, M., & Johnston, D. (2015). Statistical Tragedy in Africa? Evaluating the Data Base for African Economic Development. *Journal of Development Studies*, *51*(2), 111–115. https://doi.org/10.1080/00220388.2014.968141

Joaristi, M., Serra, E., & Spezzano, F. (2019). Detecting suspicious entities in Offshore Leaks networks. *Social Network Analysis and Mining*, *9*(1), 1–15. https://doi.org/10.1007/s13278-019-0607-5

Jullum, M., Løland, A., Huseby, R. B., Ånonsen, G., & Lorentzen, J. (2020). Detecting money laundering transactions with machine learning. *Journal of Money Laundering Control*, *23*(1), 173–186. https://doi.org/10.1108/JMLC-07-2019-0055

Kar, D., & Cartwright-Smith, D. (2008). *Illicit Financial Flows from Developing Countries: 2002-2006*. Global Financial Integrity.

Kaufmann, D., Kraay, A., & Mastruzzi, M. (2010). The worldwide governance indicators: Methodology and analytical issues. *World Bank Policy Research Working Paper*, *5430*, 2–28.

Kellenberg, D., & Levinson, A. (2019). Misreporting trade: Tariff evasion, corruption, and auditing standards. *Review of International Economics*, *27*(1), 106–129. https://doi.org/10.1111/roie.12363

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2017). Human decisons and machine predictions. *NBER Working Paper Series*, *23180*, 3–76. https://doi.org/S1793557119500840

Kleinberg, J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). Prediction policy problems. *American Economic Review*, *105*(5), 491–495. https://doi.org/10.1257/aer.p20151023

Labib, N. M., Rizka, M. A., & Shokry, A. E. M. (2020). Survey of Machine Learning Approaches of Anti-money Laundering Techniques to Counter Terrorism Finance (A. Z. Ghalwash, N. El Khameesy, D. A. Magdi, & A. Joshi, Eds.). *Internet of Things—Applications and Future*, 73–87.

Lépissier, A., & Cobham, A. (2019). Risk Measures for Illicit Financial Flows Dataset. https://doi.org/10.5281/zenodo.3371739

Li, Y., Duan, D., Hu, G., & Lu, Z. (2009). Discovering hidden group in financial transaction network using hidden markov model and genetic algorithm. *6th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2009*, *5*, 253–258. https://doi.org/10.1109/FSKD.2009.592

McCallum, J. (1995). National Borders Matter: Canada-U.S. Regional Trade Patterns. *The American Economic Review*, *85*(3), 615–623.

Molenberghs, G., Fitzmaurice, G., Kenward, M. G., & Tsiatis, A. (2014). *Handbook of Missing Data Methodology*. Chapman and Hall/CRC.

Morse, J. C. (2019). Blacklists, Market Enforcement, and the Global Regime to Combat Terrorist Financing. *International Organization*, *73*(3), 511–545. https://doi.org/10.1017/S002081831900016X

Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87–106. https://doi.org/10.1257/jep.31.2.87

Nelleman, C., & INTERPOL Environmental Crime Programme. (2012). *Green Carbon, Black Trade: Illegal logging, tax fraud and laundering in the world's tropical forests*. A Rapid Response Assessment. United Nations Environment Programme, GRID-Arendal.

OECD. (2017). *OECD Transfer Pricing Guidelines for Multinational Enterprises and Tax Administrations*. https://doi.org/10.1787/9789264192218-sum-pt

O'Hare, B., Makuta, I., Bar-Zeev, N., Chiwaula, L., & Cobham, A. (2014). The effect of illicit financial flows on time to reach the fourth Millennium Development Goal in Sub-Saharan Africa: A quantitative analysis. *Journal of the Royal Society of Medicine*, *107*(4), 148–156. https://doi.org/10.1177/0141076813514575

O'Neil, C. (2016). *Weapons of Math Destruction: How big data increases inequality and threatens democracy*. Broadway Books.

Paige, J., Fuglstad, G.-A., Riebler, A., & Wakefield, J. (2020). Design- and Model-Based Approaches to Small-Area Estimation in A Low- and Middle-Income Country Context: Comparisons and Recommendations. *Journal of Survey Statistics and Methodology*, 1–31. https://doi.org/10.1093/jssam/smaa011

Patnaik, I., Sen Gupta, A., & Shah, A. (2012). Determinants of Trade Misinvoicing. *Open Economies Review*, *23*(5), 891–910. https://doi.org/10.1007/s11079-011-9214-4

Paula, E. L., Ladeira, M., Carvalho, R. N., & Marzagão, T. (2017). Deep learning anomaly detection as support fraud investigation in Brazilian exports and anti-money laundering. *Proceedings - 2016 15th IEEE International Conference on Machine Learning and Applications, ICMLA 2016*, 954–960. https://doi.org/10.1109/ICMLA.2016.73

Piermartini, R., & Yotov, Y. (2016). Estimating Trade Policy Effects with Structural Gravity. *World Trade Organization Working Paper*, *ERSD-2016-*, 1–63.

Piketty, T., & Goldhammer, A. (2014). *Capital in the Twenty-first Century*. The Belknap Press of Harvard University Press.

Pinkovskiy, M., & Sala-i-Martin, X. (2014). Lights, camera,... income! Estimating poverty using national accounts, survey means, and lights. *Federal Reserve Bank of New York Staff Reports*, *669*, 1–91. https://doi.org/10.1111/j.1467-629X.1980.tb00220.x

Quimba, F. M. A., & Barral, M. A. A. (2018). Exploring Neural Network Models in Understanding Bilateral Trade in APEC: A Review of History and Concepts. *Philippine Institute for Development Studies Discussion Paper*, *33*, 1–20.

Raza, S., & Haider, S. (2011). Suspicious activity reporting using Dynamic Bayesian Networks. *Procedia Computer Science*, *3*, 987–991. https://doi.org/10.1016/j.procs.2010.12.162

Reuter, P. (2012). *Draining development? Controlling flows of illicit funds from developing countries*. World Bank Group.

Rijkers, B., Baghdadi, L., & Raballand, G. (2017). Political Connections and Tariff Evasion Evidence from Tunisia. *World Bank Economic Review*, *31*(2), 459–482. https://doi.org/10.1093/wber/lhv061

Rose, A. K., & Spiegel, M. M. (2007). Offshore Financial Centres: Parasites or Symbionts? *The Economic Journal*, *117*(523), 1310–1335. https://doi.org/10.1111/j.1468-0297.2007.02084.x

Rose, G. (2014). *Following the Proceeds of Environmental Crime: Fish, Forests and Filthy Lucre*. Routledge.

Salomon, M. (2019). *Illicit Financial Flows to and from 148 Developing Countries: 2006-2015*. Global Financial Integrity.

Sandefur, J., & Glassman, A. (2015). The Political Economy of Bad Data: Evidence from African Survey and Administrative Statistics. *Journal of Development Studies*, *51*(2), 116–132. https://doi.org/10.1080/00220388.2014.968138

Santos Silva, J., & Tenreyro, S. (2006). The Log of Gravity. *The Review of Economics and Statistics*, *88*(4), 641–658.

SAS. (2019). *What is next-generation AML? The fight against financial crime fortified with robotics, semantic analysis and artificial intelligence*.

Savage, D., Wang, Q., Chou, P., Zhang, X., & Yu, X. (2016). Detection of money laundering groups using supervised learning in networks. *arXiv*, *1608.00708*(cs.SI), 1–11.

Schuster, C., & Davis, J. (2020). *Old dog, new tricks? The fitness of mirror trade analysis to detect illicit financial outflows from Africa*. United Nations Conference on Trade and Development.

Senator, T. E., Goldberg, H. G., Wooton, J., Cottini, M. A., Khan, A. F. U., Klinger, C. D., Llamas, W. M., Marrone, M. P., & Wong, R. W. H. (1995). Financial Crimes Enforcement Network AI System (FAIS) Identifying Potential Money Laundering from Reports of Large Cash Transactions. *AI Magazine*, *16*(4), 21–39.

Shaxson, N. (2011). *Treasure islands: Uncovering the damage of offshore banking and tax havens*. Palgrave Macmillan.

Shaxson, N., & Christensen, J. (2013). *The Finance Curse*. Tax Justice Network.

Spanjers, J., & Salomon, M. (2017). *Illicit Financial Flows to and from Developing Countries: 2005-2014*. Global Financial Integrity.

Storm, H., Baylis, K., & Heckelei, T. (2020). Machine learning in agricultural and applied economics. *European Review of Agricultural Economics*, *47*(3), 849–892. https://doi.org/10.1093/erae/jbz033

Sudjianto, A., Nair, S., Yuan, M., Zhang, A., Kern, D., & Cela-Díaz, F. (2010). Statistical methods for fighting financial crimes. *Technometrics*, *52*(1), 5–19. https://doi.org/10.1198/TECH.2010.07032

Tax Justice Network. (2020). *Financial Secrecy Index 2020 Methodology*.

Tinbergen, J. (1962). *Shaping the world economy; suggestions for an international economic policy*. The Twentieth Century Fund.

Tiwari, M., Gepp, A., & Kumar, K. (2020). A review of money laundering literature: The state of research in key areas. *Pacific Accounting Review*, *32*(2), 271–303. https://doi.org/10.1108/par-06-2019-0065

Tørsløv, T. R., Wier, L. S., & Zucman, G. (2018). The Missing Profits of Nations. *National Bureau of Economic Research Working Paper Series*, *24701*, 1–29. https://doi.org/10.3386/w24701

UNCTAD. (2018). *Trade Analysis Information System (TRAINS)*. United Nations Conference on Trade and Development.

UNCTAD. (2020). *Tackling Illicit Financial Flows for Sustainable Development in Africa*. United Nations Conference on Trade and Development.

UNECA. (2017). *Impact of illicit financial flows on domestic resource mobilization: Optimizing revenues from the mineral sector in Africa*. United Nations Economic Commission for Africa. https://doi.org/10.5040/9781350220652.ch-004

UNECA. (2018a). *Base Erosion And Profit Shifting In Africa: Reforms to Facilitate Improved Taxation of Multinational Enterprises*. United Nations Economic Commission for Africa.

UNECA. (2018b). *A study on the global governance architecture for combating illicit financial flows*. United Nations Economic Commission for Africa.

UNECA. (2019). Multinational corporations, tax avoidance and evasion and natural resources management. In *Economic Report on Africa 2019* (pp. 118–137). United Nations Economic Commission for Africa.

United Nations. (2020). Commodities Trade (Comtrade) Database.

UNODC. (2011). *Estimating Illicit Financial Flows Resulting from Drug Trafficking and other Transnational Organized Crimes*. United Nations Office of Drugs and Crime.

UNODC & UNCTAD. (2020). *Conceptual framework for the Statistical Measurement of Illicit Financial Flows*. United Nations Office of Drugs and Crime and United Nations Conference on Trade and Development.

van der Does de Willebois, E., Halter, E. M., Harrison, R. A., Park, J. W., & Sharman, J. (2011). *The Puppet Masters: How the Corrupt Use Legal Structures to Hide Stolen Assets and What to Do About It*. The World Bank and UNODC. https://doi.org/10.1596/978-0-8213-8894-5

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC bioinformatics*, *7*, 91. https://doi.org/10.1186/1471-2105-7-91

Vézina, P. L. (2015). Illegal trade in natural resources: Evidence from missing exports. *International Economics*, *142*, 152–160. https://doi.org/10.1016/j.inteco.2014.09.001

Walker, J. (1999). How Big is Global Money Laundering? *Journal of Money Laundering Control*, *3*(1), 25–37. https://doi.org/10.1108/eb027208

Walker, J., & Unger, B. (2009). Measuring global money laundering: "The Walker Gravity Model". *Review of Law and Economics*, *5*(2), 821–853. https://doi.org/10.2202/1555-5879.1418

West, J., & Bhattacharya, M. (2016). Intelligent financial fraud detection: A comprehensive review. *Computers and Security*, *57*, 47–66. https://doi.org/10.1016/j.cose.2015.09.005

Wohl, I., & Kennedy, J. (2018). Neural Network Analysis of International Trade. *Office of Industries, US International Trade Commission*, *ID-049*, 1–17.

Worku, T., Mendoza, J. P., & Wielhouwer, J. L. (2016). Tariff evasion in sub-Saharan Africa: The influence of corruption in importing and exporting countries. *International Tax and Public Finance*, *23*(4), 741–761. https://doi.org/10.1007/s10797-016-9407-2

World Customs Organization. (2018). *Illicit Financial Flows via Trade Mis-invoicing*.

Yikona, S., Slot, B., Geller, M., & Hansen, B. (2011). *Ill-gotten Money and the Economy - Experiences from Malawi and Namibia*. World Bank Group.

Yotov, Y. V. (2012). A simple solution to the distance puzzle in international trade. *Economics Letters*, *117*(3), 794–798. https://doi.org/10.1016/j.econlet.2012.08.032

Yotov, Y. V., Piermartini, R., Monteiro, J.-A., & Larch, M. (2016). *An Advanced Guide to Trade Policy Analysis: The Structural Gravity Model*. World Trade Organization and United Nations Conference on Trade and Development. https://doi.org/10.30875/abc0167e-en

Zucman, G. (2013). The missing wealth of nations: Are Europe and the U.S. net debtors or net creditors? *Quarterly Journal of Economics*, *128*(3), 1321–1364. https://doi.org/10.1093/qje/qjt012

Appendices

A Codebook

A.1 Outcome variables

Code	Manipulation	Direction	Aggregation
`GER_Tot_IFF`	Gross outflows	Total outflows from i to j	Gross
`In_GER_Tot_IFF`	Gross inflows	Total inflows from j to i	Gross

A.2 Predictors

Gravity variables

Code name	Description	Data source	Type	Unit of observation
`GDP`	Gross domestic product (thousands, current US$)	*Gravity Database*, CEPII	Continu-ous	it, jt
`pop`	Population (thousands)	*Gravity Database*, CEPII	Continu-ous	it, jt
`dist`	Distance between most populated city of each country (km)	*Gravity Database*, CEPII	Continu-ous	ij
`contig`	Countries are contiguous	*Gravity Database*, CEPII	Dummy	ij
`comlang`	Share a common official language	*Gravity Database*, CEPII	Dummy	ij
`comcol`	Share a common colonizer	*Gravity Database*, CEPII	Dummy	ij
`col45`	In a colonial relationship post-1945	*Gravity Database*, CEPII	Dummy	ij
`entry_cost`	Cost of business start-up procedures (% of GNI per capita)	*Gravity Database*, CEPII	Continous	it, jt
`RTA`	Countries have a regional trade agreement	*Gravity Database*, CEPII	Dummy	ijt

Governance variables

Code name	Description	Data source	Type	Unit of observation
`CorrCont`	Control of corruption, percentile rank	*Worldwide Governance Indicators*, Kaufmann and Kraay	Continu-ous	it, jt
`RegQual`	Regulatory quality, percentile rank	*Worldwide Governance Indicators*, Kaufmann and Kraay	Continu-ous	it, jt
`RuleLaw`	Rule of law, percentile rank	*Worldwide Governance Indicators*, Kaufmann and Kraay	Continu-ous	it, jt

Financial integrity variables

Code name	Description	Data source	Type	Unit of observation
`SecrecyScore`	Secrecy score on the Financial Secrecy Index	*Financial Secrecy Index*, Tax Justice Network	Continuous	j
`FSIRank`	Rank on the Financial Secrecy Index (low: more secretive)	*Financial Secrecy Index*, Tax Justice Network	Continuous	j
`KFSI13`	Avoids promoting tax evasion	*Financial Secrecy Index*, Tax Justice Network	Continuous (100: fully secretive, 0: fully transparent)	j
`KFSI17`	Meets anti-money laundering FATF recommendations	*Financial Secrecy Index*, Tax Justice Network	Continuous (100: fully secretive, 0: fully transparent)	j
`KFSI20`	Engages in international judicial cooperation on money laundering	*Financial Secrecy Index*, Tax Justice Network	Continuous (100: fully secretive, 0: fully transparent)	j
`FATF`	Financial Action Task Force (FATF) membership	FATF	Dummy	i, j

Regulatory environment variables

Code name	Description	Data source	Type	Unit of observation
`tariff`	Average tariff across HS 2-digit commodities applied by i on imports from j	UNCTAD TRAINS	Continuous	ijt, $ijtc$
`kai`	Average capital controls on inflows	*Capital Control Measures*, Fernández et al. 2021	Continuous	it, jt
`kao`	Average capital controls on outflows	*Capital Control Measures*, Fernández et al. 2021	Continuous	it, jt
`cc`	Average restrictions on commercial credits for international trade	*Capital Control Measures*, Fernández et al. 2021	Categorical (0, 0.5, 1)	it, jt
`cci`	Commercial credits inflow controls	*Capital Control Measures*, Fernández et al. 2021	Dummy	it, jt
`cco`	Commercial credits outflow controls	*Capital Control Measures*, Fernández et al. 2021	Dummy	it, jt
`di`	Average restrictions on direct investment accounts	*Capital Control Measures*, Fernández et al. 2021	Categorical (0, 0.5, 1)	it, jt
`dii`	Direct investment inflow controls	*Capital Control Measures*, Fernández et al. 2021	Dummy	it, jt
`dio`	Direct investment outflow controls	*Capital Control Measures*, Fernández et al. 2021	Dummy	it, jt

B Procedure for tuning hyperparameters

While model parameters themselves are learned by the algorithm, hyperparameters are those parameters that can be manipulated by the analyst in order to improve predictive performance. These tuning parameters govern how severely the *parameters* of the final estimator will penalize flexibility. For example, in the case of a Random Forest (RF), the individual parameters that are *learned* by the model from the data are the features and the thresholds that are used to split each node during training. By contrast, hyperparameters must be set before by the analyst; they are knobs to be turned before training occurs.

Several hyperparameters were tuned to identify a sensible way to configure the Random Forest estimator. Parameters that were tuned include the number of regression trees that will make up the forest, and the number of variables that will be taken into account by each tree. Increasing the number of trees in a forest will create a more robust aggregate model (since Random Forest is an ensemble learner), but will come at the cost of increased computational time. Moreover, reducing the number of features that the RF algorithm will use each time it grows a tree can further serve to decorrelate the individual trees and decrease the overall variance (though the individual trees will be more biased). Another hyperparameter that was tuned is the maximum depth of the individual trees: very deep trees will fit the training data well but will have high individual variance; though since RF aggregates the individual trees, overall variance of the ensemble is less of a concern. The remaining hyperparameters that were tuned are the minimum number of observations required in a node before a split can be considered, and the minimum amount of samples that must be placed in a leaf node (decreasing both of these parameters will result in more flexible, less biased, trees).

A randomized search strategy with 5-fold cross-validation was employed in order to tune the hyperparameters of the Random Forest estimator. Since it would be computationally prohibitive to consider every possible combination of the hyperparameters, a distribution of hyperparameters was provided instead. This defines the search space, and the tuning process involves randomly sampling a combination of those hyperparameters and evaluating the performance of the resulting RF configuration using cross-validation. The hyperparameter space was randomly sampled 100 times and evaluated in 5-fold cross-validation using *scikit-learn*.[25] In other words, each of the 100 trials corresponds to a candidate RF model that is tuned with a different configuration of hyperparameters, is trained on 4 folds, and then evaluated using the hold-out fold, resulting in 500 possible model configurations that were fitted. The tuning procedure was conducted on the training sample to preserve the integrity of the test set. The procedure was repeated twice:

[25]Implemented using the *RandomizedSearchCV* procedure of *scikit-learn* (random seed 1509).

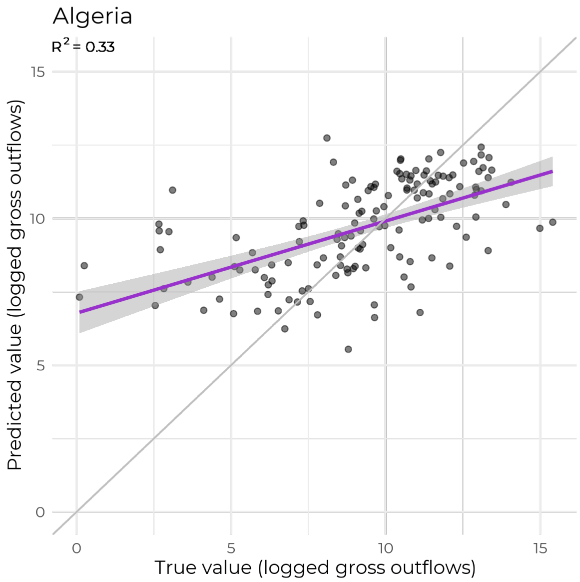
once on the training data for outflows and once on the training data for inflows; in both cases, the randomized search yielded the same tuning for the RF estimator. The best configuration of hyperparameters was identified as the one that leads to the highest R^2 on the hold-out sets during cross-validation, and is reported in table 7 below.

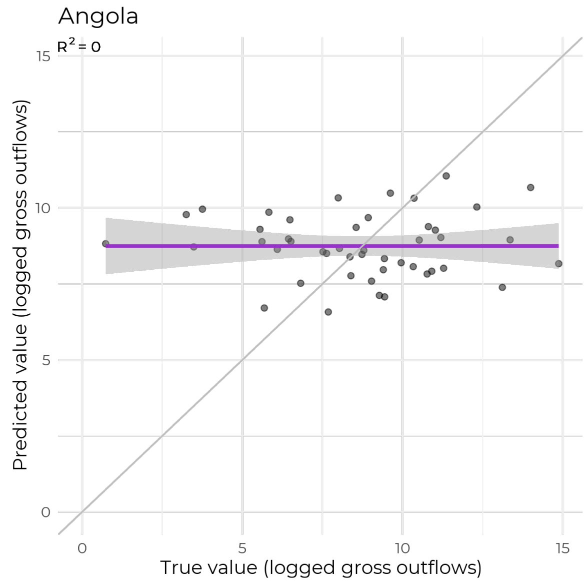
Hyperparameter	**Tuning**
Number of trees	1278
Maximum depth of individual trees	195
Minimum number of observations to split on at an internal node	12
Minimum number of observations in a leaf (terminal node)	1
Maximum number of random features to consider at each split	All features
Use bootstrapped samples to build the trees	Yes

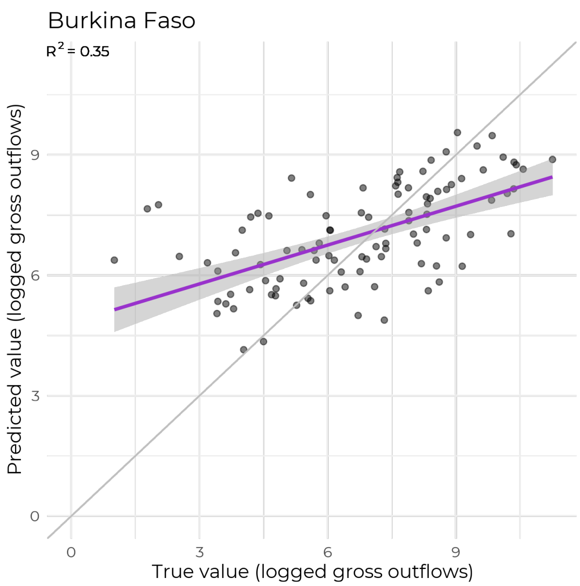
Table 7: Tuned hyperparameters for the Random Forest estimator following randomized search strategy with 5-fold cross-validation. The search resulted in the same configuration of hyperparameters for both outflows and inflows.

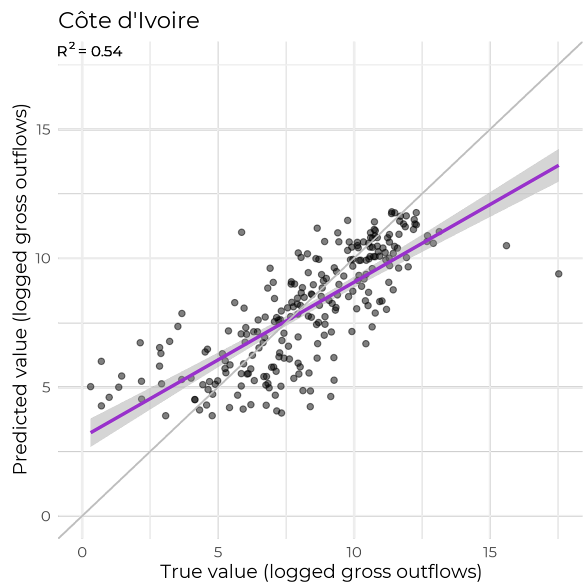
C Cross-validated predictions for all African countries

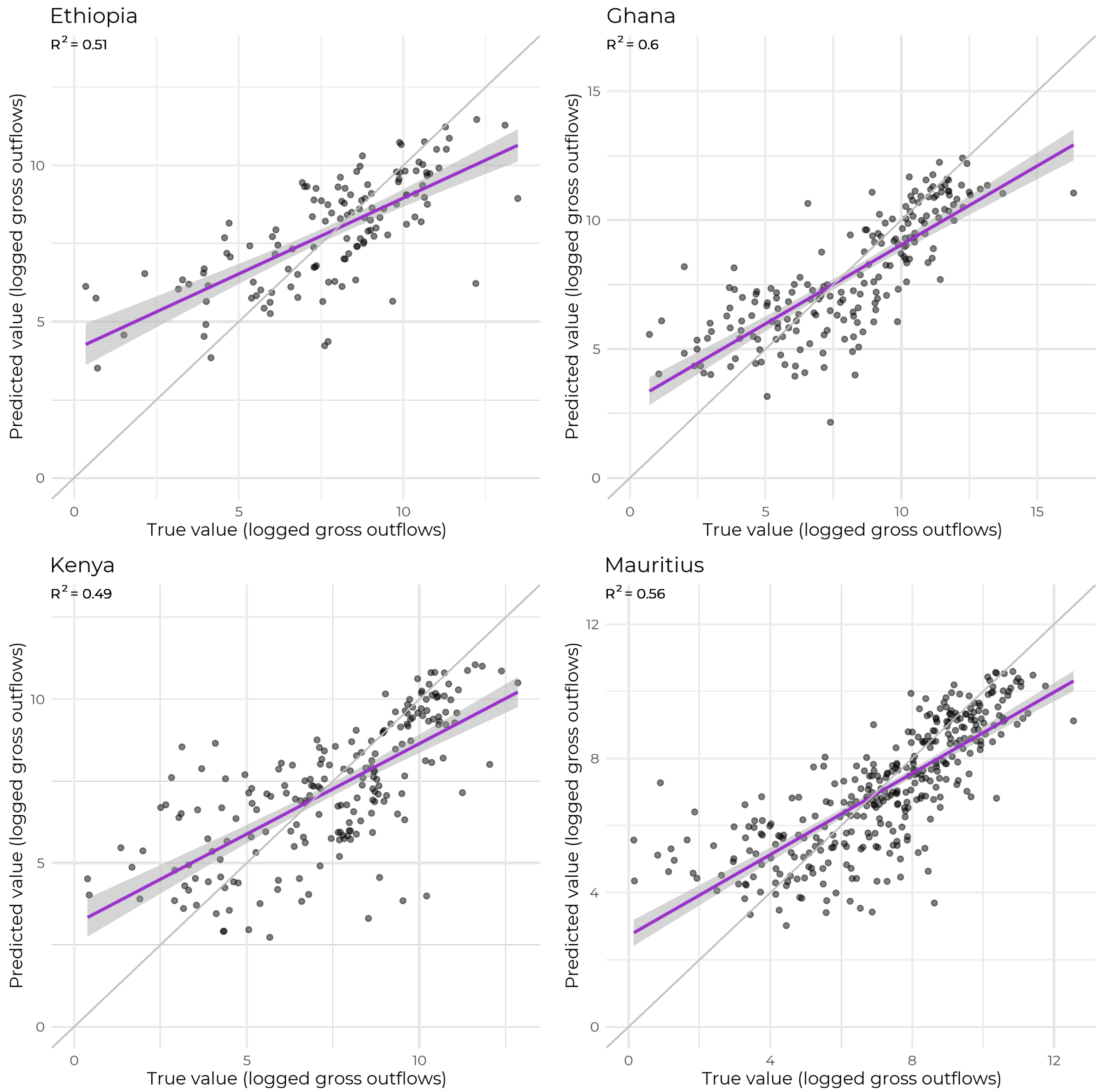
C.1 Gross outflows

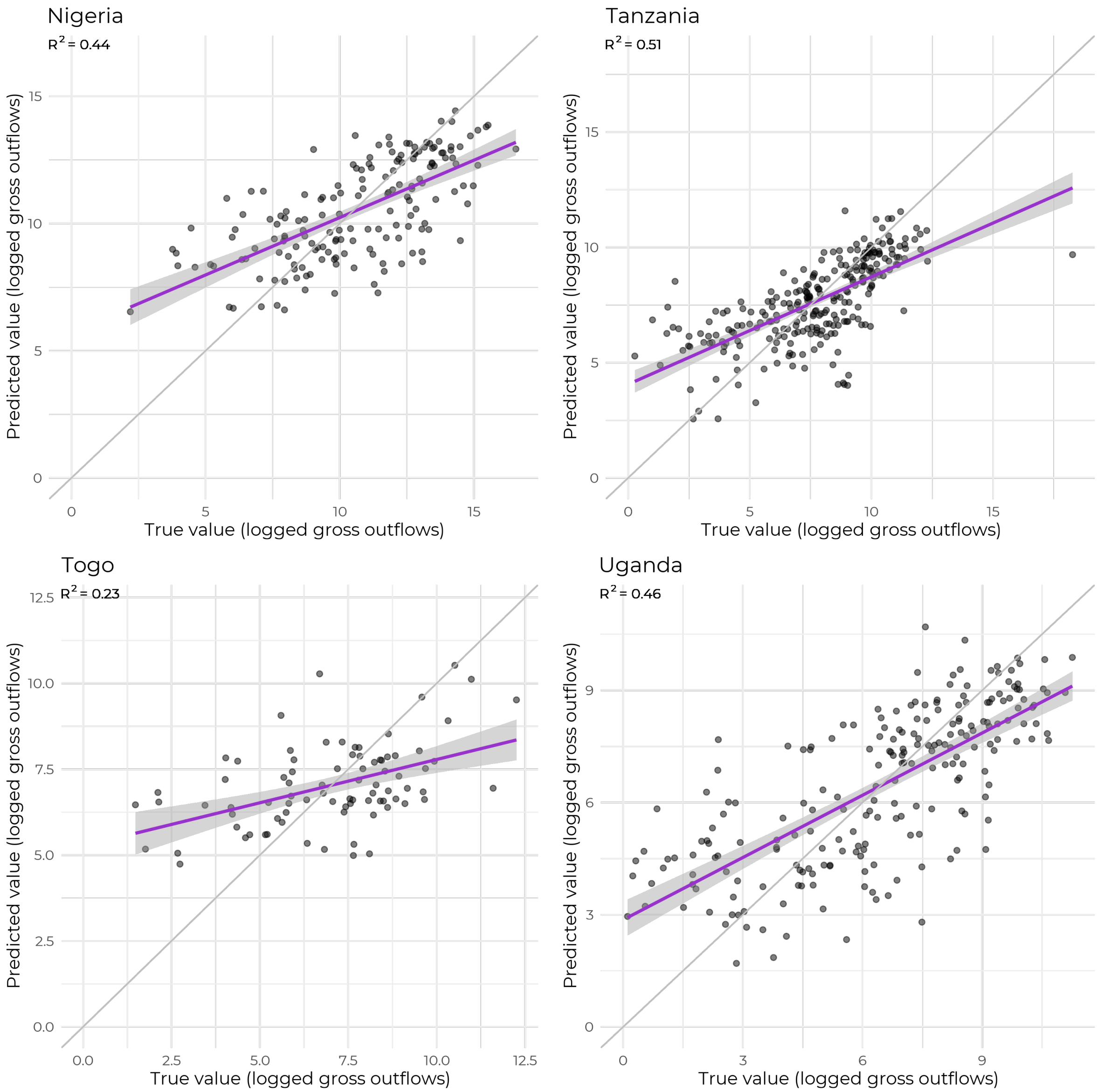


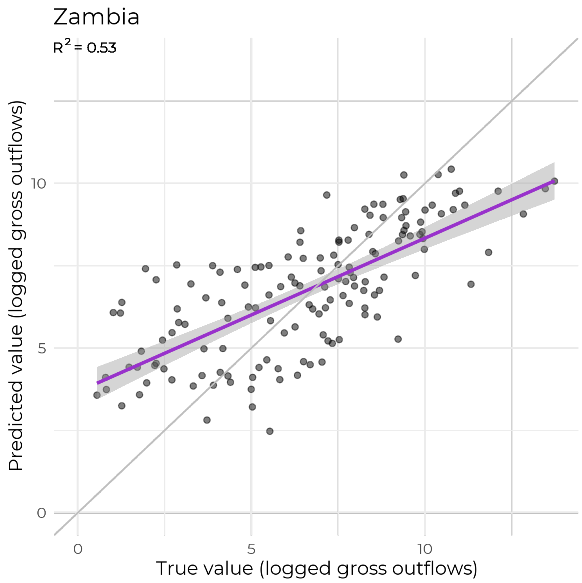






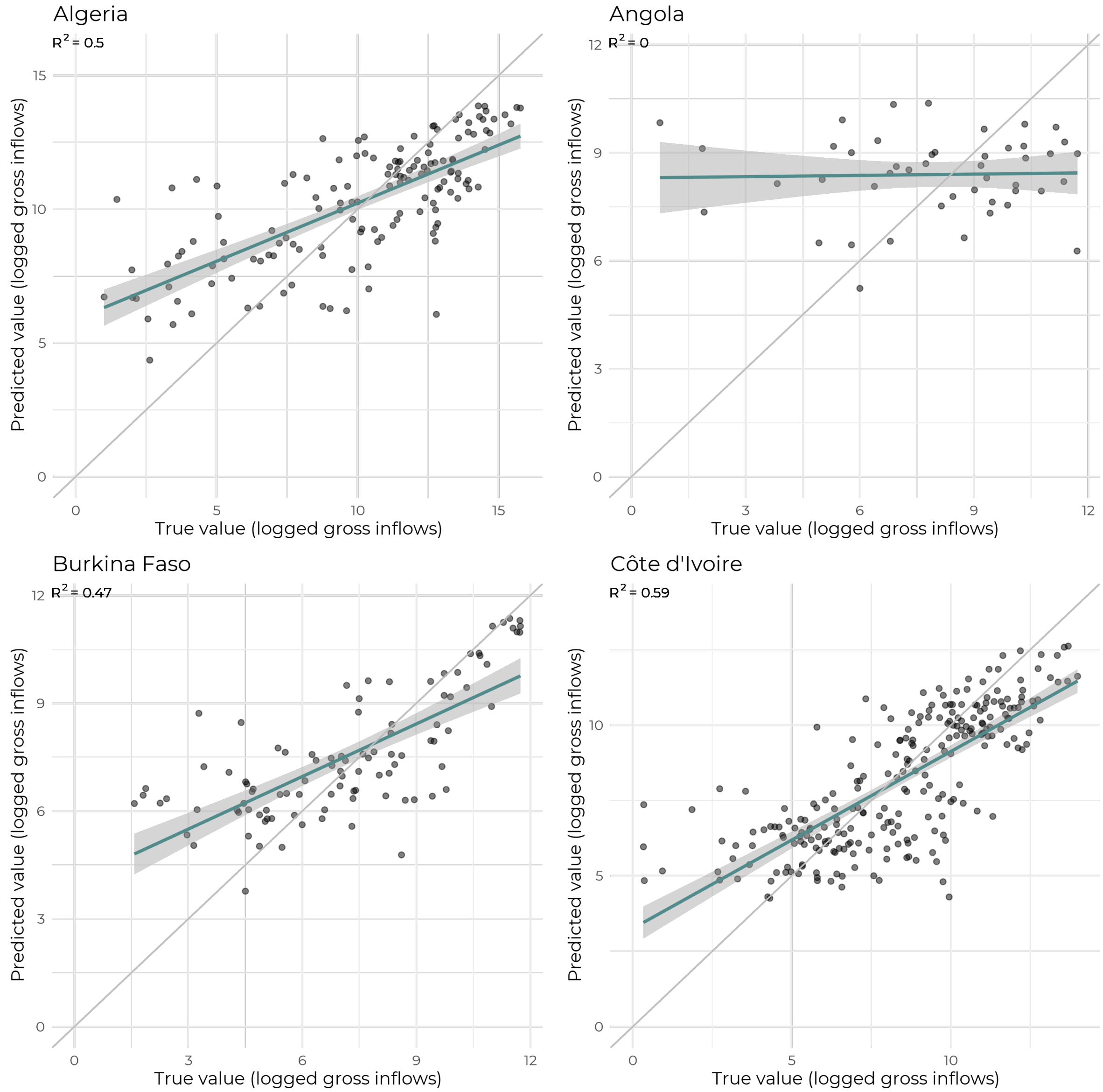

Ethiopia
R² = 0.51
Ghana
R² = 0.6
Kenya
R² = 0.49
Mauritius
R² = 0.56
Predicted value (logged gross outflows)
True value (logged gross outflows)

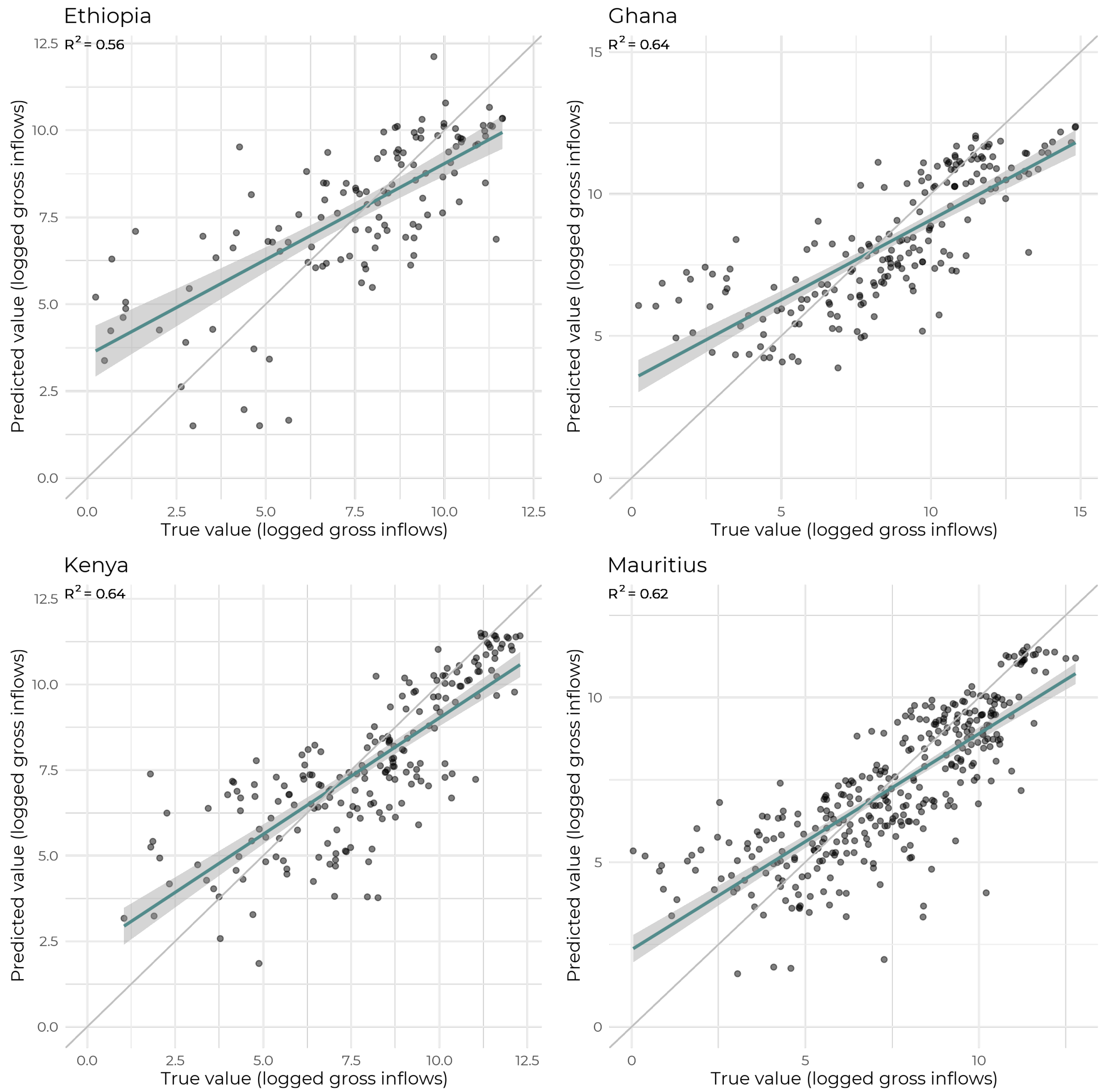

Nigeria
R² = 0.44
Predicted value (logged gross outflows)
True value (logged gross outflows)
Tanzania
R² = 0.51
Predicted value (logged gross outflows)
True value (logged gross outflows)
Togo
R² = 0.23
Predicted value (logged gross outflows)
True value (logged gross outflows)
Uganda
R² = 0.46
Predicted value (logged gross outflows)
True value (logged gross outflows)

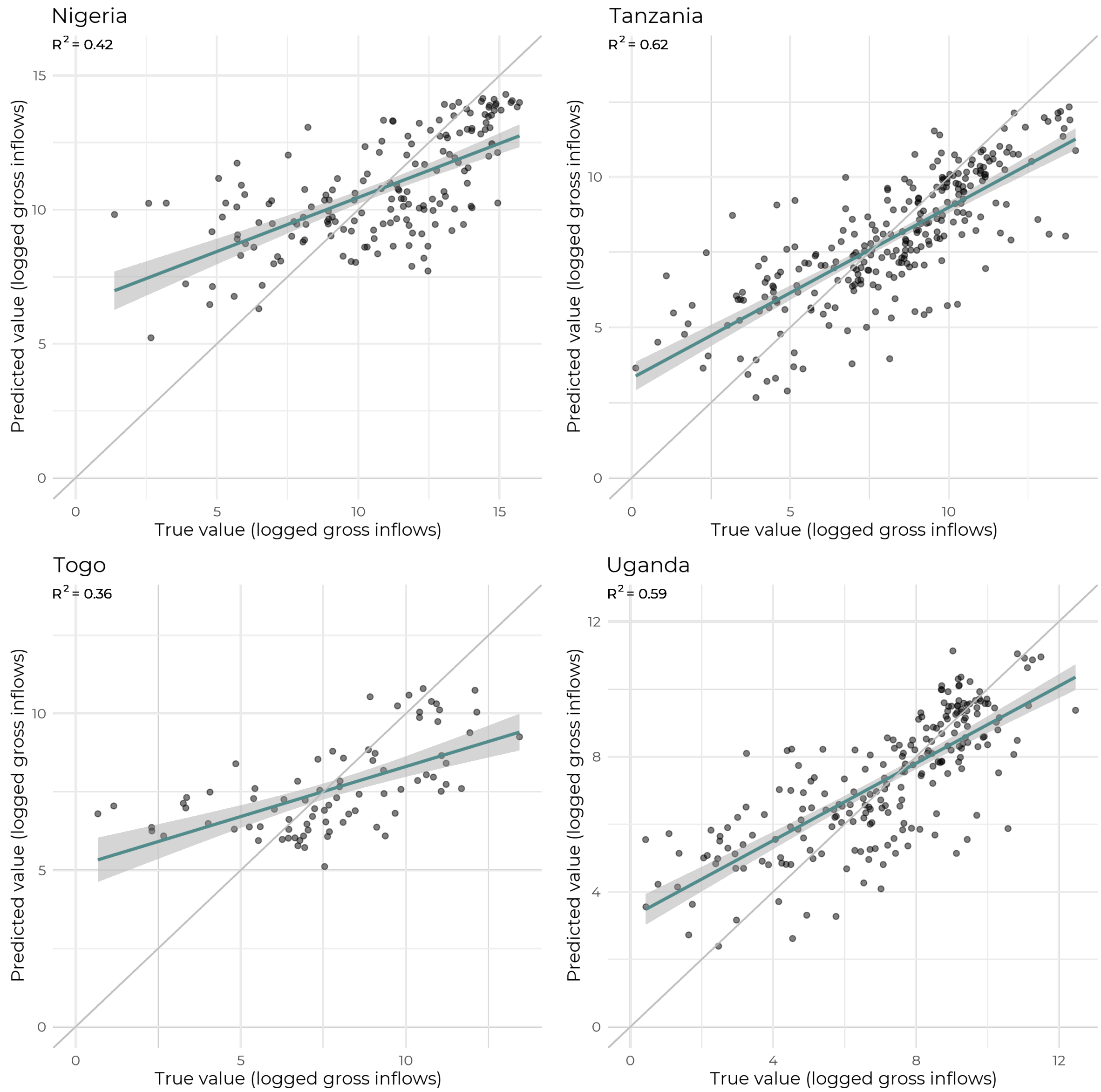


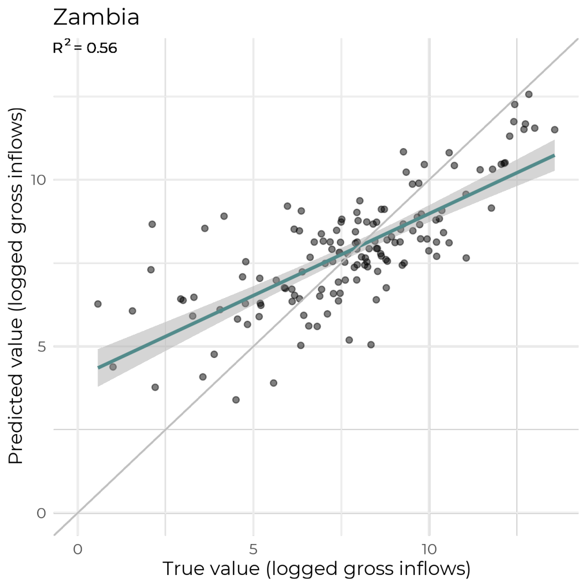
Zambia
R² = 0.53
Predicted value (logged gross outflows)
True value (logged gross outflows)
0
5
10
0
5
10

C.2 Gross inflows




Ethiopia
$R^2 = 0.56$
Ghana
$R^2 = 0.64$
Kenya
$R^2 = 0.64$
Mauritius
$R^2 = 0.62$
Predicted value (logged gross inflows)
True value (logged gross inflows)


Nigeria
R² = 0.42
Tanzania
R² = 0.62
Togo
R² = 0.36
Uganda
R² = 0.59
Predicted value (logged gross inflows)
True value (logged gross inflows)


Zambia
R² = 0.56
Predicted value (logged gross inflows)
True value (logged gross inflows)
0
5
10
0
5
10

D Robustness check: reduced form linear models

	Dependent variable	
	ln.Tot_IFF	ln.In_Tot_IFF
ln.gdp_i	0.748***	0.665***
ln.gdp_j	0.974***	0.885***
comlang	0.714***	0.976***
comcol	1.146***	1.036***
rta	1.869***	2.329***
CorrCont_i	−0.008***	−0.004
CorrCont_j	−0.001	0.0004
RegQual_j	−0.004	
RegQual_i		−0.014***
FATF_i	1.900***	1.328***
FATF_j	0.885***	1.194***
ihs.tariff	−0.034	0.035
kao_i	0.268***	
kai_j	0.412***	
kai_i		−0.434***
kao_j		0.864***
Constant	−25.759***	−22.539***
Observations	6,165	5,874
Adjusted R^2	0.430	0.402
Residual Std. Error	2.453 (df = 6151)	2.520 (df = 5860)
F Statistic	358.451*** (df = 13; 6151)	304.649*** (df = 13; 5860)
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 8: Estimates of gross inflows and gross outflows of misinvoicing (pooled over 2000-2018).

The coefficient estimates of the reduced form linear models are presented above. The predictor variables included in these baseline models have not been selected empirically, and instead were selected because they are likely to be theoretically important predictors. The parameter estimates should be interpreted as correlations and not causal estimates. Note that the number of observations available for estimating the reduced form linear models is greater than the training set of the

RF model because less features are used which results in fewer list-wise deletion of observations. The estimates presented above were obtained by fitting the linear models on a training sample; while predictive performance was evaluated on an independent test set.