Title
A methodological toolkit to understand complex policy problems: applications to climate change and illicit finance

Author
Lépissier, Alice

Publication Date
2021

Peer reviewed|Thesis/dissertation

University of California

Santa Barbara

A methodological toolkit to understand complex policy problems: applications to climate change and illicit finance

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy

in

Environmental Science and Management

by

Alice Lépissier

Committee in charge:

Professor Matthew Potoski, Chair
Professor Matto Mildenberger
Professor Robert Heilmayr
Professor Alexander Franks

December 2021

The Dissertation of Alice Lépissier is approved.

Professor Matto Mildenberger

Professor Robert Heilmayr

Professor Alexander Franks

Professor Matthew Potoski, Committee Chair

October 2021

A methodological toolkit to understand complex policy problems: applications to climate change and illicit finance

by

Alice Lépissier

À Max et à mes parents

Acknowledgements

Though pursuing a doctorate is ultimately an individual enterprise, my experience has been far from solitary, and I would like to acknowledge the colleagues, friends, and family who have supported me and cheered me on.

I have had the good fortune of benefitting from the guidance of an extraordinary committee. My advisor, Matt Potoski, pushed me to think bigger and clearer – and I thank him for his support over the years. My committee members, Matto Mildenberger, Robert Heilmayr, and Alexander Franks, have provided terrifically useful and strategic advice. I have learned a lot from all of my advisors, and I am deeply grateful for their kindness and mentorship. Knowing that my commmittee was always in my corner helped me reach my goals.

During my time at UCSB, I found various intellectual homes in the ENVENT, Conservation Economics, and Franks-Oh labs. I am also thankful for the environmental politics community at Bren and in the department of Political Science.

Faculty in the departments of Statistics & Applied Probability and Political Science have expanded my intellectual horizons in ways that I had not imagined. Thank you to Alex Franks and Sang-Yun Oh for opening up a world to me. Faculty at Bren and in Political Science have been very generous with me, and have provided support at crucial junctures in my PhD. In particular, I'd like to thank Leah Stokes, Mark Buntaine, Kyle Meng, and Sarah Anderson.

This work would not have been possible without the financial support of the Bren School and UC Santa Barbara. I also gratefully acknowledge the United Nations Economic Commission for Africa.

I thank my colleagues Will Davis and Gamal Ibrahim, and my former colleagues at the Center for Global Development, including Owen Barder, Michael Clemens, and Jonah Busch.

I would like to thank my students, and in particular my POLS 15 students. I taught this class in various capacities over the years and it turned out to be one of the most personally meaningful experiences of my PhD. Being able to encourage my students to believe in themselves and in their coding abilities, and to see them thrive as a result has been deeply rewarding. It made me so happy to have, without fail, a few students approach me at the end of every quarter, seeking to take their data science journeys further. I was honored to mentor several brilliant young women throughout my time here, and I can't wait to see what they achieve next.

Being in the trenches with my rockstar officemates – Sam Collie, Sallie Collie, Jason Maier, Niklas Griessbaum, Nākoa Farrant – made going to work every day much more fun. I couldn't imagine sweeter or lovelier officemates. Thanks for the jokes and for the brilliant conversations. Thanks to the 3L lab wing for the snacks, and for feeding me spoonfuls of peanut butter when I needed a pick-me-up.

Hands down the best experience at Bren was finding a community of people who are not only whip smart, but are also the kindest and the nicest people. I love and will miss my Brennie pals! Patrick Hunnicutt has been my brother-in-arms throughout this PhD – and has provided more support and solidarity than I can adequately express and thank him for. We are lucky to have a solid crew of social scientists at Bren who are always rooting for each other – thank you to Elliott Finn, Vincent Thivierge, Jacob Gellman, and Geoff Henderson, and many others. Thank you for your friendship, insightful comments, and all the fun times! Come see us in Rhode Island!

I would like to thank my parents, Bertrand and Monique Lépissier, for the unique gift they bestowed upon my sister and I of a childhood spent across different countries, without which I probably wouldn't have ended up doing a PhD in California. I am a product of the French, British, and sometimes Spanish, educational systems – and this has profoundly shaped my personal and intellectual development. My late grandfather Henri Lépissier is the person who inspired me to pursue a career in international development. I thank my sister Pauline, who I admire more than she knows, for her sweetness, goofiness, and our shared effortless understanding of each other. I thank my family-in-law for their love and affection.

I thank my girlfriends in London and Paris for their friendship. I've known some of them for more than 15 years, and I treasure them. Merci les filles!

Above all, the person that I would like to thank the most is my husband Max Wojcik. It is hard to overstate the impact that a supportive spouse has on all life outcomes. Max has been an unwavering source of support, in all ways. He is the person that kept me fed, watered, and told me when it was time to go to bed. And when I still decided to work late, he was the person that would bring me cups of tea. A million cups of tea later, here we are – and Max is still the best decision I have ever made.

Oso the puppy bounced into our lives during the last year of my dissertation and brought with him countless hours of laughter and joy. He has delighted us with his antics, his German Shepherd theatrics, and his deep abiding conviction that it is always time to play ball. Thank you Oso for only eating my computer mouse and not my dissertation.

Alice Lépissier

alicelepissier@ucsb.edu
www.alicelepissier.com
alicelepissier
walice

Education

2021 **PhD**, *Environmental Science and Management*, Bren School of Environmental Science and Management, University of California, Santa Barbara.

- Research on climate change, illicit financial flows, and illegal logging
- Qualifying fields: Comparative Politics and Applied Statistics

2020 **Master of Arts**, *Statistics*, University of California, Santa Barbara.

- Data Science specialization
- GPA: 4.0

2011 **Master of Science**, *Economic History*, London School of Economics.

- Awarded with Merit
- Dissertation: "Adaptation strategies and market responses to deal with uncertainty in rain-fed agriculture in Sub-Saharan Africa"

2010 **Master of Science**, *Economics and Public Policy*, Sciences Po Paris jointly with École Polytechnique.

- Quantitative economics program to elaborate and evaluate public policies in a wide range of contexts
- Events director for student organization on development in Sub-Saharan Africa · Managed 15 organization members

2009 **Bachelor of Arts**, *European Social and Political Studies*, University College London.

- Awarded with First Class Honors
- Year abroad at the *Universidad Complutense de Madrid*

Experience

2015 – 2020 **Consultant**, *Various organizations*.

- Development of open data tool for the **Tax Justice Network** quantifying risk exposure to illicit financial flows
- Quantitative modeling of foreign direct investment and illicit financial flows for the **United Nations Economic Commission for Africa**
- Scoping report for the **Global Commission on Business and Sustainable Development** on how the private sector can contribute to the Sustainable Development Goals
- Quantitative analysis and creation of user-friendly replication tool for the **Tax Justice Network** on how much capital is held offshore

2014 – 2015 **Research Associate**, *Center for Global Development*, London.

- Creator of tool SkyShares.org which allows policy-makers to compute economic costs of a climate deal
- Conceptualized and developed website with intuitive user interface for decision-makers
- Quantitative and qualitative research: economic modeling, drafting policy papers and blog posts
- Policy outreach and engagement on climate agreements, climate finance, and illicit financial flows

2012 – 2014 **Research Assistant**, *Center for Global Development*, London.
Evidence-driven research on illicit financial flows and climate change to develop practical policy proposals.

Summer 2009 **Consultant**, *LaquaR Consultants CC*, Cape Town.
Research for report on Namibian sea-level rise for the Namibian Ministry of Environment and Tourism.

Summer 2009 **Research**, *ICLEI Africa Secretariat*, Cape Town.
Research on climate change adaptation for the city of Cape Town.

2007 & 2006 **Summer Internships**, *BRED Banque Privée*, Paris, Private Wealth Management division.

Teaching

(x4) **POLS 15 Introduction to Research Methods**, *Department of Political Science*, Instructor of Record (Summer 2020), Teaching Assistant (Spring 2020, Spring 2019, Fall 2017).

(x2) **ESM 210 Business and the Environment**, *Bren School of Environmental Science and Management*, Teaching Assistant (Fall 2019), Grader (Spring 2017).

(x1) **ESM 269/ED 218 Survey Design and Environmental Public Opinion**, *Cross-listed Bren School and Department of Education*, Teaching Assistant (Spring 2018).

Working papers

A. Lépissier. “Machine learning for missing data on illicit trade”.

A. Lépissier, W. Davis, and G. Ibrahim. “Illicit financial flows from trade mis-invoicing and their implication for financing for development”.

Open datasets

2019 **A. Lépissier** and A. Cobham. “Risk Measures for Illicit Financial Flows Dataset”.
doi:10.5281/zenodo.3371738

2019 **A. Lépissier**. "Illicit Financial Flows from Trade Mis-Invoicing Dataset".
doi:10.5281/zenodo.3610557

Publications

2021 **Climatic Change**, *Unilateral climate policies can substantially reduce national carbon pollution*, 166:31.
A. Lépissier and M. Mildenberger

2021 **Tax Justice Network and Latindadd report**, *Vulnerability and Exposure to Illicit Financial Flows Risk in Latin America*.
A. Cobham, J. Garcia-Bernardo, M. Harari, A. Lépissier, S. Lima, M. Meinzer, L. Montoya Fernández and L. Moreno

2019 **Tax Justice Network report**, *Vulnerability and Exposure to Illicit Financial Flows Risk in Africa*.
C. Abugre, A. Cobham, R. Etter-Phoya, A. Lépissier, M. Meinzer, N. Monkam and A. Mosioma

2016 **Technical note**, "Chapter 5: Cheaper, Cooler, Faster: Reducing Tropical Deforestation for a More Cost-Effective Global Response to Climate Change" In *Why Forests? Why Now? The Science, Economics, and Politics of Tropical Forests and Climate Change*, Frances Seymour and Jonah Busch. Brookings Institution Press, Washington, DC.
J. Busch, J. Engelmann and A. Lépissier

2015 **CGD Policy Paper series**, *SkyShares: Modelling the Economic Implications of a Future Global Emissions Budget*, Policy Paper 067, Center for Global Development.
O. Barder, A. Evans and A. Lépissier

2015 **Technical paper**, *Modelling SkyShares: Technical background*, Center for Global Development.
A. Lépissier, O. Barder and A. Evans

2015 **UNA-UK report**, "A global emissions budget." In *Climate 2020 - Facing the Future*, Report by the United Nations Association - UK.
O. Barder, A. Lépissier and A. Evans

2015 **Contributed chapters**, *Track it! Stop it! Get it! Report of the High Level Panel on Illicit Financial Flows from Africa*, Commissioned by the AU/ECA Conference of Ministers of Finance, Planning and Economic Development.
African Union and United Nations Economic Commission for Africa

2013 **Journal of International Development**, *Europe Beyond Aid: Assessing European countries' individual and collective commitment to development*, Vol. 25, Issue 6.
O. Barder, J. Clark, A. Lépissier, L. Reynolds and D. Roodman

2012 **CGD Working Paper series**, *Europe Beyond Aid: Assessing Europe's commitment to development*, Working Paper 313, Center for Global Development.
O. Barder, J. Clark, A. Lépissier, L. Reynolds and D. Roodman

Presentations

2019

- "Mapping developing countries' exposure to illicit financial flows: a new data-set of risk measures", Dark Architectures: Advancing Research on Illicit Global Wealth Chains, *University of Sussex*, UK
- "Assessing the causal impact of UK climate policy using synthetic controls: an impact analysis of the 2001 Climate Change Programme", 5th Annual Conference on Environmental Politics and Governance, *University of California, Santa Barbara*, USA
- "Mapping developing countries' exposure to illicit financial flows: a new data-set of risk measures", Development Studies Association Annual Conference 2019, *The Open University*, UK
- "The environmental effectiveness of UK climate reforms: An impact analysis of the 2001 Climate Change Programme", Duck Family Graduate Workshop on Environmental Politics and Governance, *University of Washington*, USA
- "Assessing the causal impact of carbon pricing using synthetic control methods: an impact analysis of the UK's Climate Change Programme", Bren PhD Symposium, *University of California, Santa Barbara*, USA
- "Trade mis-invoicing estimates in Africa", Inception Meeting on Preventing Trade Mis-invoicing in Selected African Countries, *UN Economic Commission for Africa*, Ethiopia

2018

- Expert meeting on statistical methodologies for measuring illicit financial flows, Sustainable Development Goal 16.4.1, *UN Conference on Trade and Development*, Switzerland

2015

- "Ice and Climate Science - Question Time with the Experts", *University of Newcastle*, UK
- "Economics of climate change: Extension or Reinvention?", Warwick Climate Forum, *University of Warwick*, UK

2014

- "How can financial markets help us tackle climate change?", *University of Warwick*, UK
- "Stakeholder perspectives on climate change adaptation", JPI Climate, *Euro-Mediterranean Center on Climate Change*, Italy

2013

- "Exposure to illicit flows in Africa", Technical Committee of the High Level Panel on Illicit Financial Flows from Africa, *UN Economic Commission for Africa*, Ethiopia
- "An illicit flows target for post-2015", Conference on Illicit Financial Flows, *Academics Stand Against Poverty*, UK

2010

- Interview with Youssou N'Dour about African development, *Sciences Po Paris*, France
  Event filmed for Senegalese television · Organized the event & negotiated free concert · 200 attended

Professional courses

2014 Practical General Equilibrium Modeling with GAMS *EcoMod, Boston, MA*

2013 SQL Data Management *Impartica IT Training, Leeds, UK*

2013 Panel Data Analysis using Stata *Timberlake, London, UK*

Computing

Skills econometrics · machine learning · Computable General Equilibrium modeling · website development · user experience design · scientific reproducibility

Languages **R** · **Python** · **Stata** · **JavaScript** · **Visual Basic** · **TEX** · **HTML/CSS** · **Docker** *main*

Google Cloud Platform · Linux · MATLAB · SQL *secondary*

Languages

Fluent **French** (native) · **English** · **Spanish** *written & spoken*

Italian *conversational*

Other writing and service activity

Op-ed "Global goals can deliver on 2°C and new development finance - here's how", *The Guardian*, 9/24/15.

Blogging Various posts on *Views from the Center for Global Development*, including "Setting a Price on Carbon Immediately Is Necessary to Avoid Dangerous Climate Change" and "A Global Carbon Tax or Cap-and-Trade? Part 1: The Economic Arguments".

Referee *Journal of Environment and Development*, *Climatic Change*

Service Coordinator of the "Sustaining Movement Momentum and Building Political Power" conference, June 2017, *University of California Santa Barbara*

Service Advisor to Master's group project: "Prioritizing restorative wood products by market readiness, climate impact and carbon storage potential", *Bren School of Environmental Science and Management*

Service Search committee for tenure-track Lecturer in Environmental Data Science, 2020, *Bren School of Environmental Science and Management*

Service Search committee for Assistant Professor in Environmental Data Science, 2020, *Bren School of Environmental Science and Management*

Awards and fellowships

2020 Central Campus Dissertation Fellowship, Graduate Division, *University of California Santa Barbara*

2020 Nominee for Outstanding Teaching Assistant Award, Academic Senate, *University of California Santa Barbara*

2020 Graduate Student Research Fellow, *National Socio-Environmental Synthesis Center (SESYNC)*

2019 H. William Kuni Fellowship, *Bren School of Environmental Science and Management*

2019 Winner of "People's Choice" award for best talk at PhD Symposium, *Bren School of Environmental Science and Management*

2017 Energy and Climate Graduate Fellow of the *Scholars Strategy Network*

2016 Recruitment Fellowship, *Bren School of Environmental Science and Management* (for 3 years)

2016 International Doctoral Recruitment Fellowship, *University of California Santa Barbara* (for 3 years)

2006 Sessional Prize for academic merit, *University College London*

Grants

2020 "Financial Opacity and Challenges to Forest Governance in Indonesia and Malaysia", *National Socio-Environmental Synthesis Center (SESYNC)*, co-PI on Graduate Pursuit

2016 "The economic prize for business of achieving the Sustainable Development Goals", *Business and Sustainable Development Commission*, research grant

2016 "The Price of Offshore, robust replication", *Tax Justice Network*, research grant

2013 "The Political Economy of Illicit Financial Flows in Africa and their Impact on Africa's Development", *United Nations Economic Commission for Africa*, grant with Alex Cobham

Abstract

A methodological toolkit to understand complex policy problems: applications to climate change and illicit finance

by

Alice Lépissier

Complex policy problems like climate change and illicit finance require a diverse methodological repertoire and an agnostic approach to selecting the appropriate analytical tool to accomplish discrete inferential tasks. Drawing from the disciplines of political science, economics, and statistical data science, this dissertation tackles three distinct problems on causal evaluation, measurement, and missing data.

The first paper evaluates the causal effect of a climate mitigation policy on the carbon emissions of the UK. Using a synthetic control estimator, this chapter finds that post-treatment emissions in the UK were 10% lower than what they would have been without the climate policy. The results imply that voluntary climate reforms that make concessions to domestic producers are still able to meaningfully reduce emissions, even in the absence of a legally binding global climate treaty.

The second paper presents a novel methodology to measure illicit trade flows and originates the "atlas of misinvoicing", the first database to provide comprehensive bilateral estimates of the dollar amount of misinvoiced trade disaggregated by commodity sector for 167 countries during 2000-2018. Results show that African countries lost on average $86 billion a year in gross illicit outflows during that period, and that the biggest source

of illicit trade on the continent was the natural resources sector. The findings suggest that combating illicit financial flows will be crucial to providing finance for sustainable development and to promoting domestic resource mobilization in poor countries.

The third paper proposes a machine learning approach to ameliorate the problem of missing data from developing countries, where administrative systems for data collection tend to be weaker. Some African countries do not provide customs declarations, which the "atlas" method requires as input data. This chapter predicts 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. This demonstrates the promise of predictive approaches to augment existing measures of illicit finance in data-constrained settings.

Broadly, the chapters in this dissertation can be understood as operating in the different scientific frameworks of causal, descriptive, and predictive inference, respectively. Tackling difficult environmental and developmental problems will require a willingness to traverse methodological siloes in order to identify the best tool for the job – this dissertation contributes to pushing the search for solutions forward.

Contents

Chapter 1

Prologue

1.1 Introduction

Complex policy problems like climate change and illicit finance require a diverse methodological repertoire and an agnostic approach to formulating the research design that will be most appropriate to addressing the specific question at hand. Climate change and illicit finance fall under the class of "wicked problems": problems that resist systematic *a priori* formulations and that seem impossible to solve for the social planner due to their social complexity. Wicked problems have no optimal set of solutions, only specific aspects of the problem can be ameliorated. However, certain dimensions of wicked problems can be broken down into "tame" problems that can be solved with discrete inferential tasks. Drawing from the disciplines of political science, economics, and statistical data science, this dissertation addresses three distinct problems in the study of climate change and illicit finance: a causal evaluation problem, a measurement problem, and a missing data problem.

The target of analytical inquiry differs in each case; consequently, this dissertation uses an assortment of methods to answer the questions in a principled way, while proposing specific innovations to correct for methodological difficulties pertaining to each task. In other words, the problems presented here have different estimands that call for separate estimators. In terms of substantive contributions, chapter 2 of this dissertation advances our understanding of climate change policies, while chapters 3 and 4 contribute to the scholarship on illicit finance. Chapter 2 shows how to conduct *ex post* impact evaluations of climate reforms that do not rely on simplistic Business As Usual scenarios to generate a baseline level of carbon emissions. Chapter 3 presents the construction of a new proxy measure for illicit financial flows that addresses long-standing methodological concerns about the extant methods used to detect illicit activity from discrepancies in mirror trade statistics. Finally, chapter 4 proposes a predictive approach to deal with the paucity of data on economic outcomes in developing countries, and demonstrates that machine learning models can be used to credibly augment the database presented in chapter 3. Broadly, the chapters in this dissertation can be understood as operating in the different scientific frameworks of causal, descriptive, and predictive inference, respectively. Brief abstracts of the chapters are provided next.

Chapter 2 evaluates the causal effect of a climate mitigation policy on the carbon dioxide emissions of the United Kingdom. Here, the estimand is the Average Treatment Effect (ATE) of the climate policy, that is, the mean difference in outcomes between CO_2 emissions once the policy was in place and hypothetical emissions if the climate reform had not been passed. Impact evaluation of climate change policies is difficult because the underlying drivers of carbon emissions are complex latent factors, and the adoption of climate policies by governments is not random. Since the failure of climate governance regimes that sought to impose legally binding treaty-based obligations, the Paris Agreement relies on voluntary actions by individual countries. Yet, there is no guarantee that unilateral policies will lead to a decrease in carbon emissions. Critics worry that voluntary climate measures will be weak and ineffective, and insights from political economy imply that regulatory loopholes are likely to benefit carbon-intensive sectors. The chapter empirically evaluates whether unilateral action can still reduce carbon pollution by estimating the causal effect of the United Kingdom's 2001 Climate Change Programme (CCP) on the country's carbon emissions. Existing efforts to evaluate the overall impact of climate policies on national carbon emissions rely on Business-As-Usual (BAU) scenarios to project what carbon emissions would have been without a climate policy. Instead, the chapter uses a synthetic control estimator to undertake an *ex post* national-level assessment of the UK's CCP without relying on parametric BAU assumptions, by constructing a plausible counterfactual for the emissions trajectory of the UK in a world where the Climate Change Programme (CCP) would not be in place. Despite setting lax carbon targets and making substantial concessions to producers, the resulting estimate is that, post-treatment, the UK's CO_2 emissions per capita in 2005 were 9.8% lower relative to what they would have been if the CCP had not been passed. The findings offer empirical confirmation that unilateral climate policies can still reduce carbon emissions, even in the absence of a binding global climate agreement and in the presence of regulatory capture by industry.

Turning now to the emerging academic scholarship in illicit finance, Chapter 3 accomplishes the elemental task of creating a new proxy measure for illicit financial flows from trade misinvoicing – an illicit practice that is used to clandestinely shift money in and out of a country by manipulating the trade invoices presented to customs. Here, the estimand is the population-level quantity of misinvoiced trade – which remains unobservable because illicit financial flows are deliberately hidden. The secrecy of illicit financial flows is an emblematic characteristic of the problem and, as such, much of the academic effort in the literature is geared towards developing credible methods to detect illicit activity from official economic statistics. Thus, the problem can be reduced to asking what can be learned from an unobservable random variable given an observable one. This chapter presents an original methodology and database – the "atlas of misinvoicing" – that provides bilateral estimates of illicit trade, disaggregated by sector, for 167 countries between 2000 and 2018. Existing methods of estimating trade misinvoicing look for discrepancies in mirror trade statistics to locate instances of trade misinvoicing. Yet, these methods have been faulted for uncritically equating trade irregularities with illicit misinvoicing, and have been accused of generating phantom estimates of illicit financial flows that are an illusion created by the statistical artefacts of how countries record international trade transactions. The chapter approaches the problem in a principled way by deriving the properties that a persuasive measure of trade misinvoicing must possess in order to be both theoretically cogent and practically applicable. Then, the "atlas" method proposes several innovations designed to fulfill these criteria and to ameliorate long-standing problems in the literature; including by providing an empirical way to ascertain the trade gaps that result from benign, non-illicit, factors. The "atlas" measure estimates that developing countries lose $500 billion of dollars in gross outflows a year, with illicit trade costing Africa $86 billion a year, where trade in natural resource commodities is heavily misinvoiced. The implication of these findings is that combating illicit financial flows from trade misinvoicing will be crucial to allow poor countries to mobilize domestic resources to finance their own sustainable development goals.

While the "atlas" database introduced in chapter 3 is the first of its kind to provide broad country coverage, it is still missing data from countries who do not report international trade statistics, including 10 African countries. Chapter 4 presents a strategy to address the problem of missing data on economic measures in developing countries, by using machine learning models to predict illicit trade outcomes without requiring data on the observed trade flow for training. Here, the "atlas" database is taken as a measure of ground truth, and the estimand is the amount of trade misinvoicing conditional on observed country-level features denoting unilateral and bilateral characteristics. Missing or poor quality data is a prevalent problem in developing countries due to weak administrative systems for statistical reporting. This hinders the study of trade misinvoicing, which relies on recorded trade declarations by national customs authorities. The patchiness of data on commodity trade flows compounds the prejudice for African countries who are particularly vulnerable to illicit financial flows. Random Forest machine learning models are trained to predict misinvoiced trade on a sample of African countries using only variables that are easily observed, such as distance between countries, or that are readily compiled by researchers and publicly available, such as perceptions of good governance, without requiring data on the observed trade flow. The Random Forest estimators are able to explain 70% to 73% of the variation in illicit trade outcomes on an unseen test set. 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 models are able to detect meaningful structure in the dyadic nature of countries' bilateral relationships that is predictive of illicit trade. The findings 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.

The remainder of this prologue introduces the broad analytical lens of the dissertation and offers a perspective on empirical research. Analysts who study complex real-world problems and who seek to use rigorous empirical research to drive evidence-based decision-making forward might encounter the disconcerting trade-off between the need to provide relevant and timely insights that were obtained in a principled manner, and the reluctance to make assumptions beyond what can be credibly accepted in their field. Instead, an agnostic point of view of policy-relevant empirical research emphasizes the value of producing reasoned insights while eschewing making strong assumptions on the data-generating process in nature. Recognizing the "wicked" nature of many complex policy problems suggests that there is no need to accept that there is a simple generative model for the problem that can be known to be true. Wicked problems defy a definitive formulation, and so it follows that different epistemologies will exist about the way to reach a reasoned conclusion. Instead, by breaking off facets of the wicked problem into discrete "tame" problems, this dissertation advances substantive knowledge on climate change and illicit finance using methods that are appropriate to the inferential target of the "tame" problem. The emphasis in this dissertation is on generating credible inferences that are based on substantive assumptions, derived from theory and domain knowledge, and rigorously estimated within the strictures of statistical inference. Moreover, any normative positions and assumptions relating to the social welfare consequences of these problems are transparently and clearly articulated.

Climate change is an existential peril where the urgent imperative of drastically reducing carbon emissions will necessitate a complete transformation of the ways in which our economies organize production and consumption (Intergovernmental Panel on Climate Change, 2014; Rogelj et al., 2018), thus generating inherent conflicts around the distribution of the costs and benefits of climate action (Mildenberger, 2020; Genovese, 2019). Moreover, the glacial pace of progress globally in the last 40 years relative to the starkness of the scientific record reflects disagreements over the meaning of global and intergenerational equity (Barder et al., 2015; Pickering et al., 2015). Similarly, the proliferation of illicit finance both reflects and exacerbates the unequal distribution of the gains of globalization. On the one hand, deeper financial integration has led to the emergence of secrecy jurisdictions – countries whose comparative advantage in the global marketplace is to provide legalized financial opacity, loose regulations, and low taxes – that have provided fertile ground for abuses such as concealing corruption, tax avoidance, and money laundering by powerful actors such as multinational companies, wealthy individuals, and other elites (Shaxson, 2011; Shaxson and Christensen, 2013; Christensen, 2012). On the other hand, illicit finance threatens to disrupt the fabric of society by entrenching disparities and inflaming the political problems associated with inequality, and continues to jeopardize the prospects for sustainable development in poor countries

(Lonergan and Blyth, 2020; Baker, 2005; Reuter, 2012). Therefore, the symptoms of climate change and illicit finance are related to their genesis, a hallmark of "wicked problems". These complex policy problems seem intractable due to the complexity of competing interests from different stakeholders. Researchers wishing to make progress, piece by piece, grapple with various types of questions to ameliorate different facets of the problem. Therefore, tackling wicked problems will require a good dose of humility and a willingness to traverse methodological siloes in order to identify the best tool for the job at hand.

While shunning absolutist claims to knowledge – some tools will be useful some times to solve some aspects of the problem – it is nonetheless possible to distinguish the type of methodological instruments that are well-suited to address a specific manifestation of the problem, and some iconoclasm might even be warranted to outline the limitations of popular approaches in empirical social sciences and suggest specific ways in which other approaches from statistical data science might fill those gaps. This dissertation is an ecumenical collection of papers where the different methodological traditions in empirical social sciences and statistical data science are appreciated for the distinctive value added they confer to the analysis of climate change and illicit finance, and where the comparative advantage of each approach is correctly identified in order to select the inferential framework that will yield the most analytical leverage for the specific "tame" problem at hand. Faced with the real-world urgency of climate change and illicit finance, researchers must know how to direct a suitable methodology to the target of inquiry.

Illustrating this point, the next sections develop the concrete argument that machine learning approaches are uniquely suited to the study of wicked problems. Section 1.2 presents the attributes of wicked problems and demonstrates that climate change and illicit finance exhibit all the features of wicked problems. In turn, understanding climate change and illicit finance as wicked problems reveals the epistemological limitations of the common inferential framework that much of applied social sciences is predicated on; one rooted in causal inference where the task of the analyst is to credibly estimate the causal links between independent variables and an outcome. Yet, predictive questions are conspicuous in the field of illicit finance, and as such they require a different mode of statistical inference. While econometric techniques specialize in the consistent estimation of parameters and the interpretability of the resulting coefficients, machine learning excels at predictive tasks. Section 1.3 further reflects on the two cultures of traditional econometrics approaches and machine learning approaches and sketches their limits and areas of complementarity. Finally, section 1.4 derives the unique value proposition of machine learning for the analysis of wicked problems from the features of its inferential machinery, and offers some caveats.

1.2 The "wicked problems" of climate change and illicit finance

Climate change and illicit financial flows (IFFs) fall under the class of "wicked problems", that is, problems that are difficult or impossible to solve for the social planner (Rittel and Webber, 1973; Conklin, 2006). Wicked problems are hard to define due to their social complexity and are resistant to solutions. Rittel and Webber (1973) propose ten characteristics of wicked problems, which broadly map to either the problem's formulation or the problem's solutions. First, there is no definitive formulation of a wicked problem (Rittel and Webber, 1973). There are numerous explanations that can be provided for why a discrepancy representing a wicked problem exists. Since there is always more than one explanation for a phenomenon, the choice of explanation will determine the nature of the problem's resolution (Rittel and Webber, 1973). The other definitional attributes of a wicked problem are that each problem is essentially unique,[1] and that every wicked problem can be considered to be a symptom of another problem.

Second, the defining traits of a wicked problem that pertain to the nature of its solutions are: wicked problems have no stopping rule; there are no true-or-false solutions to wicked problems, only better or worse ones; there is no immediate and no ultimate test of a solution to a wicked problem; and finally, there is not an enumerable set of solutions to attempt (Rittel and Webber, 1973). Furthermore, Rittel and Webber (1973) theorize that every attempt at a solution to a wicked problem is consequential and leaves traces that cannot be undone. There is no opportunity to learn by trial-and-error; every effort is a "one-shot operation" that changes the policy space and future solution set. Wicked problems come with an intrinsic challenge to designing policy solutions to combat them, because the process of conceiving a solution is identical to the process of understanding the nature of the problem; in Rittel and Webber (1973)'s view, "[t]he formulation of a wicked problem is the problem!" (p. 161).

In addition to complicating the design of practical policy solutions, wicked problems also entail a more subtle epistemological implication that is worth highlighting. In the definition of wicked problems proposed by Rittel and Webber (1973), the social planner "has no right to be wrong" (p. 166), which is taken to mean that the planner is liable for the consequences of the solutions they enact. A Popperian approach to social policy would construe solutions to problems as hypotheses that are put forward for refutation and thus, it follows that they should be falsifiable with evidence (Rittel and Webber,

[1] It is always possible to find a distinguishing property for any two problems that is trivially unique, but the authors who formalized the theory hold that a wicked problem is *essentially* unique, because one can never be sure that the idiosyncracies of the problem are not more important than any common features that it may share with a similar looking problem (Rittel and Webber, 1973).

1973). Should it be repeatedly demonstrated that the implemented solution fails to reject the null hypothesis of no effect, then the planner's confidence in this policy solution would decrease as a result; and alternative policy interventions can then be designed and their performance can be similarly evaluated using this approach. By contrast, wicked problems do not have falsifiable solutions where the aim is to find out the truth; rather, the goal is to ameliorate some of the characteristics of the problem. Since the boundaries of wicked problems are hard to delineate, "[t]he planner who works with open systems is caught up in the ambiguity of their causal webs" Rittel and Webber (1973, p. 167). This suggests that the common epistemological approach that underpins much of applied social sciences and public policy work – one rooted in causal inference where the main exercise consists of evaluating either the determinants or consequences of social phenomena – has some limitations. Indeed, Rittel and Webber (1973) argue that "[i]n dealing with wicked problems, the modes of reasoning used in the argument are much richer than those permissible in the scientific discourse" (p. 166). Interpreting the former as an invitation, this dissertation seeks to offer modest contributions to the analysis of climate change and illicit finance by leveraging a range of inferential tools, each designed to address specific methodological challenges that arise from the type of question at hand, grappling with policy evaluation, measure construction, and missing data in turn.

The phenomena of climate change and IFFs demonstrably meet all the criteria of wicked problems. The irreversibility of certain courses of climate action and the potential for unintended consequences, the absence of a discrete set of permissible solutions, and the multiplicity of stakeholders with competing views of the problem and of how to solve it, all conspire to make climate change a wicked problem. Climate change is perhaps the "super wicked" problem of our time, given that time is also running out, those who caused the problem are now seeking to solve it, and policy interventions irrationally discount the future (Levin et al., 2012). The analytical position that climate change is a wicked problem has long been understood, and I do not linger on it further (see, e.g., Haug et al. (2010); Hildén (2011)). However, the academic field of illicit finance is still in its infancy, and conceptual treatments of the problem are sorely needed (Cobham and Janský, 2020; Reuter, 2012). Moreover, two out of the three chapters in this dissertation pertain to illicit finance, so I outline below the features of IFFs that identify it as a wicked problem.

There is no definitive formulation of illicit financial flows: unresolved debates remain on whether the definition of IFFs should be a narrow legalistic one (i.e., restricting the categorization to flows that stem from activities in direct contravention of laws) or a broader one where normative considerations are included (e.g., taking the position that while aggressive tax *avoidance* by multinational corporations is legal, unlike tax *evasion* which is illegal, it is harmful and so it should be treated as an IFF, see Blankenburg and Khan (2012); Cobham (2014)). This in turn has practical implications for how policy-makers respond to the phenomenon; in fact, the setting of the United Nations Sustainable Development Goals (SDGs) left open the question of definition and measurement, simply concluding that "by 2030, [the world should] significantly reduce illicit financial flows and arms flows, strengthen the recovery and return of stolen assets and combat all forms of organized crime" (SDG 16.4, UN General Assembly (2015)).

Moreover, the fact that observed irregularities can be interpreted as representing the wicked problem differently is a particularly apt descriptor of IFFs such as trade mis-invoicing. For any given gap in bilateral trade statistics that is observed, there exist plausible alternative explanations for the gap that cannot easily be parsed. This is the subject of chapter 3 which proposes a methodology to overcome this problem, by using a systematic way of distinguishing between benign discrepancies in trade gaps and ones that can be ascribed to illicit activity. The various possible ways of representing the problem in turn inform the nature of the proposed solutions. For instance, an outflow from country i to country j could be conceptualized as a problem of cash smuggling or of the embezzlement of public funds, or both. These two ways of problematizing the phenomenon would tend to suggest different policy interventions: the first one emphasizing the role of currency and capital controls, while the second would put the onus on boosting good governance.

As outlined above, wicked problems can always be viewed as symptomatic of other problems. The case of trade-based money laundering could be viewed, for example, as a symptom of grand corruption, which is the abuse of office committed by high-level public officials in pursuit of illicit enrichment. This inherent difficulty is exemplified in most of today's international anti-money laundering regimes, including the Warsaw Convention, by the fact that predicate offenses (that is, the underlying crimes) may be used to establish a charge of money laundering.[2] In turn, the subversion of the functions of the state associated with grand corruption can be viewed as a symptom of the erosion of the social contract between a government and its citizens. Therefore, as soon as one causal explanation is proposed, other causes must be accounted for; and it is hard to dispute that the causal arrow between variables goes in both directions, when we note that the depletion of state coffers and looting of domestic resources occasioned by IFFs further weaken the social contract. As a result, usual tools of causal inference, such as Directed Acyclic Graphs (DAGs) for example, cannot be used since they presuppose unidirectional causal relationships (a *directed* graph) and the absence of feedback loops between variables (an *acyclical* graph) (Pearl, 1995, 2009; Imbens, 2020).

The endogeneity of IFFs complicates the prescription of appropriate solutions, and ex-

[2]FATF Recommendation 3 calls on countries to criminalize the laundering of proceeds of all serious offences, with a view to including the widest range of predicate offences. The Warsaw Convention of 2005 (CETS 198) provides a legal framework for charging predicate offenses.

isting policy interventions on IFFs exhibit the same features as those of the solutions to a wicked problem. Combating illicit finance is akin to playing the arcade game "whack-a-mole", where attempts to solve the problem are often piecemeal and only result in temporary improvements: as soon as one mole is disposed of, another one will emerge from its hole. This can be seen most clearly in the case of trade misinvoicing to launder the proceeds of transnational organized crime: if one accepts the premise that crime will always exist (since most societies organize their activities around laws dictating permissible and impermissible behaviors), then there must always be some irreducible amount of illicit finance that is not caught since illicit activity, by definition, seeks to remain hidden.

Moreover, Lowery and Ramachandran (2015) have provided evidence of the unintended consequences of anti-money laundering policies, which is congruent with the idea that attempted solutions for a wicked problem are consequential. Currently, banks are required to put in place good faith efforts to stall sanctions violations, money laundering, and terrorist financing. Collectively, these rules are commonly referred to as AML/CFT (Anti-Money Laundering/Countering Financing of Terror) regimes. Given that banks are subject to large fines if they fail to perform their due diligence, many banks are altogether exiting certain sectors they deem too risky in a process known as "de-risking", which has had dire consequences for poor countries in particular. For example, widespread denials of banking accounts to money transfer organizations have increased the cost of remittances (the money that migrants send home) (Lowery and Ramachandran, 2015). Given that remittance flows are the largest source of external financing for low- and lower-middle income countries, above foreign direct investment and official development assistance (Ratha et al., 2016), the AML/CFT procedures put in place that were supposed to combat illicit financial outflows have aggravated the draining of resources in developing countries.

Problematizing illicit financial flows and climate change as wicked problems suggests that the way social science research can generate meaningful insights is to divide the wicked problem into discrete "tame" problems to ameliorate a specific dimension of the problem; this is the strategy that underpins the analytical enterprise of this dissertation. In some cases, the study of a wicked problem will benefit from working on a causal problem; this is typically the preferred approach to conducting policy impact evaluation or to analyzing the determinants of a specific manifestation of the wicked problem. Chapter 2 sets out to causally evaluate the effectiveness of a policy designed to reduce carbon emissions. Similarly, in the field of illicit finance, Allred et al. (2017) have employed the reputed gold standard in the causal inference toolkit (Gerber and Green, 2012) – a randomized experiment – to show that firms in OECD countries (compared to those in tax havens) were more willing to provide anonymous incorporations, i.e., create shell companies, and to flout international rules on financial transparency (Allred et al., 2017). However, it is not always possible for social scientists to either carefully design experiments or to identify natural experiments that provide a source of exogenous variation and allow for causal identification. When the latent factors driving a problem are complex and multiple, it is hard to provide actionable insights to policy-makers regarding the policy interventions that they can undertake to combat the problem. Pol (2020) blames the failure of existing anti-money laundering (AML) regimes, dubbed the "the world's least effective policy experiment" (p. 73), on the mismatch between the way that outcomes are understood and the design principles of current policy prescriptions. Likewise, isolating the marginal causal effect of a particular determinant on the outcome is of limited value for public decision-making if it cannot be manipulated by the social planner. In other cases, however, there exist discrete tasks associated with wicked problems that can be usefully accomplished in the different frameworks of descriptive and predictive inference, as demonstrated by chapters 3 and 4 of this dissertation, respectively.

1.3 Revisiting a tale of two cultures

In a provocative paper, Breiman (2001) issued a polemic against the overreliance of statisticians on data models which assume that the data are generated according to a stochastic model. Instead, he argues, statisticians should make space for algorithmic approaches to learning from data, where the data mechanism is treated as unknown. According to him, the almost exclusive dependence of statisticians on data models, predicated on the belief that the analyst "by imagination and by looking at the data, can invent a reasonably good parametric class of models for a complex mechanism devised by nature" (Breiman, 2001, p. 202), has led to irrelevant theory and questionable scientific conclusions. This is because any quantitative conclusion that is drawn from fitting a model to data will be a conclusion about the model's mechanism and not about nature's mechanism, and hence it follows that "[i]f the model is a poor emulation of nature, the conclusions may be wrong" (Breiman, 2001, p. 202). Breiman's prescient piece foresaw the emergence of machine learning (ML) into the mainstream of many of today's scientific enterprises.

At around the same time, a critique reminiscent of Breiman's appeared in political science with Brady and Collier (2004)'s pushback on the "quantitative imperialism" of King et al. (1994)'s *Designing Social Inquiry*, a treatise on how qualitative research should follow the precepts of quantitative social science. Brady and Collier (2004) resisted the dogmatism of mainstream quantitative methods by pointing out that regression analysis relies on the difficult-to-test assumption that the model being estimated is correct. Though the depth of the qualitative-quantitative divide in political science is sometimes overstated when considering that both still belong to a positivist framework of social science (cf. Mildenberger (2016); Mahoney and Goertz (2006)), this type of criticism from the qualitative culture of research is a salutary reminder that econometric work packs a lot of assumptions on the nature of the problem in the course of its efforts to understand it.

Economists have taken up Breiman's challenge and an influential literature is emerging in applied economics that bridges the two cultures, notably by applying machine learning algorithms to causal inference problems (Athey, 2019, 2017; Mullainathan and Spiess, 2017; Athey and Imbens, 2015; Varian, 2014; Athey and Imbens, 2019; Storm et al., 2020; Imbens and Athey, 2021). Political science is also increasingly receptive to this view (Grimmer et al., 2021; Radford and Joseph, 2020; Brady, 2019; Diamond and Sekhon, 2013; Abadie et al., 2010), and some authors have proposed an agnostic analytical framework that centers on what is learnable from the world without assuming that there exists a simple generative model that can be known to be true (Aronow and Miller, 2019; Grimmer et al., 2021).

How can machine learning provide added value in social scientific settings, particularly for the study of wicked problems? As a first stage, it is useful to draw some distinctions between machine learning approaches and traditional econometric approaches in applied social sciences. To some extent, applied empirical researchers in political science and economics share much of the same methodological toolkit (cf., for example, King et al. (1994); Angrist and Pischke (2009); Wooldridge (2010)). Though there are areas of discord and some turf wars between the two camps, applied economists and political scientists have more in common than separates them, and with the indulgence of the reader, I will refer to this as the econometric approach hereafter.

The econometric approach tends to be concerned with parameter estimation while the machine learning approach focuses on prediction (Mullainathan and Spiess, 2017; Athey and Imbens, 2019; Athey, 2019). Supervised machine learning techniques aim to estimate the conditional expectation $\mathbb{E}[Y|X]$ of a response Y given a set of variables X (called "features" in the ML literature). By contrast, econometric techniques seek to provide estimates of parameters β that govern the relationship between the explanatory variables X and the outcome Y. In the canonical use case, the analyst will first specify a functional form for the underlying assumed "true" relationship between Y and X, of the general form $\mathbb{E}[Y|X] = X^{\mathrm{T}}\beta + \epsilon$, and then seek to estimate the parameters $\hat{\beta}$ using a linear regression model. It is important to notice that, although the predictors X can enter the relationship in non-linear forms (including interactions with other variables or as higher-order polynomials X^k of the kth degree),[3] the models are *linear in the parameters* β. This is also the case for Generalized Linear Models (GLM), another commonly used class of models in social sciences designed to deal with limited dependent variables.[4] This already imposes a great deal of structure on the problem. Much effort in econometrics is spent on deploying estimators that have useful asymptotic properties such as consistency and normality.[5] Statistical inference is then performed on the estimated $\hat{\beta}$ under some regularity conditions.

[3]With the only mechanical limitations on estimating complexity – absent the common sense of the analyst – being computational power and the number of degrees of freedom available. By contrast, machine learning techniques have explicit procedures to choose the amount of complexity in a model.

[4]The linear regression model is not appropriate when the range of the outcome Y is restricted, as is often the case with the variables that social scientists work with, such as count (e.g., population) or categorical (e.g., "voted", "did not vote") data. In that case, GLMs allow the analyst to work with limited dependent variables by relaxing the constraints on the mean of the dependent variable $\mathbb{E}[Y_i|X_i] = \mu_i$. This is accomplished by using a possibly non-linear link function $g(\mu_i) = \eta_i$ that specifies how the mean relates to a *linear* function of the explanatory variables and of the parameters $\beta_1, \ldots, \beta_k$: $\eta_i = \sum_{k=1}^{K} \beta_k x_{ik}$. Since the linear predictor η_i can take any value in $(-\infty, \infty)$ while the range of Y is limited, the goal of the link function is to relax the constraints on the mean so that it maps onto $\Re$ and to define the scale over which η_i is *additive*. The point here is to note that these models once again estimate a linear function of the parameters β.

[5]An estimator is consistent when, as sample size grows to infinity, its sampling distribution converges in probability to the true parameter value β^*. Normality of the errors is required to provide valid statistical inferences.

By contrast, machine learning approaches tend to make less assumptions about the data-generating process that is responsible for the phenomenon we observe. Some ML techniques do make distributional assumptions (e.g., LDA or QDA), but most are non-parametric methods that allow flexible functional forms, and only require that observations are independent to work.[6] The machine learning approach specifies the relationship between X and Y in very general terms: $Y = f(X) + \epsilon$. The observed response Y is some unknown function f of the predictors plus some irreducible error ϵ. Most machine learning tasks are predictive and seek to generate predictions of the response by estimating f, but make no assumption about the functional form that f will have; the estimated function can end up being very complex. Thus, the problem is posed as estimating $\hat{Y} = \hat{f}(X)$. By contrast, econometric approaches impose many more assumptions on the functional form that f will take.

As I illustrated above with the canonical linear regression model, the assumption in many of the most commonly used econometric models in social sciences is that f is linearly additive in its parameters, yet that is not often surfaced. Specifically, linear regression models assume $\mathbb{E}[Y|X] = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon$. Many policy-oriented problems in social sciences revolve around identifying the causal effect β_k of a treatment X_k. Policy-oriented analyses are interested in providing estimates of the *marginal* effect of the treatment of interest. That is, we seek to know what the causal effect is of increasing X_k by one unit, holding other variables $X_1, \ldots, X_{k-1}$ constant. But it could be the case that the model f is misspecified, and the true data-generating process contains both non-linearities and interactions, which would complicate the interpretation of $\hat{\beta}_k$ as a marginal effect.[7] When policy-relevant estimands, e.g., the Average Treatment Effect

[6] The other substantive assumption that underpins machine learning is that the data in the training and test set are drawn i.i.d. from the same (unknown) distribution (Athey, 2019; Breiman, 2001), but I address this later.

[7] For example, it could be the case that the true data-generating process contains interactions, e.g.,

(ATE), are defined in terms of a linear model, the consequences of misspecifying the model f might lead to non-negligible welfare loss if treatment effect estimates are used to guide policy decisions.

The goals of the two frameworks are fundamentally different. Parameter estimation approaches set out to find an unbiased estimator $\hat{\beta}$ that minimizes in-sample error, while machine learning approaches aim to minimize the prediction error of $\hat{Y}_i$ for a *new* data point i. Thus, econometric approaches are focused on unbiasedness by construction whereas machine learning techniques provide an empirical way to manage the bias-variance trade-off.[8] The difference between both frameworks is visible in how they approach validation: econometric approaches will rely on in-sample goodness-of-fit measures to judge analytical power, while machine learning approaches assess performance by evaluating predictive accuracy on an out-of-sample observation. Since $\hat{Y} = \hat{f}(X)$ is a deterministic function of $\hat{f}$ given X, the problem of providing the best prediction of $\hat{Y}_i$ for a new observation i reduces to finding the functional form of f that minimizes a chosen loss function $\mathcal{L}(Y_i, \hat{f}(X_i))$ from a set of functions $\hat{f} \in \mathcal{F}$.[9] This is possible because we can observe prediction quality, whereas in econometric approaches we need to make assumptions about f to ensure consistency (Mullainathan and Spiess, 2017). In other words, econometric approaches *assume* the functional form of f, while machine learning approaches *learn* the representation of f from the data.

$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k(X_1, \ldots, X_{k-1})X_k + \epsilon$ where the treatment effect of interest β_k is a (possibly very complicated) function of all other covariates, or the true model could be non-linear where the effect of X_k changes with the level of X_k (so that not all unit changes are equal). I thank Alex Franks for pointing this out to me.

[8]Note that $\hat{\beta}_{OLS}$ is the Best Linear Unbiased Estimator and so by construction makes a choice in the bias-variance trade-off by ensuring zero bias; whereas ML optimizes with respect to both in-sample and out-of-sample error (Kleinberg et al., 2015).

[9]Crucially, the set of candidate functions $\mathcal{F}$ is not restricted to a set of linear predictors. The loss functions that are often picked are the Mean Square Error (MSE) in regression problems and the misclassification error rate in classification problems.

1.4 How machine learning can help study wicked problems

Machine learning (ML) approaches in general, and deep learning techniques such as neural networks in particular,[10] are well-suited to deal with problems where the underlying nature of the phenomenon is complex, dynamic, and resists *a priori* and systematic characterizations; all of which are properties of wicked problems. In other fields, machine learning has had spectacular success in accomplishing tasks where it is similarly difficult to come up with a set of hard-wired rules to follow: object recognition, translating natural languages, generating artificial but photo-realistic images, etc. What these tasks have in common is that they seem to require some degree of "intelligence" because it is not possible to pre-specify exhaustively and comprehensively the set of procedures that an agent must follow in order to correctly accomplish the task. For example, to accurately identify and distinguish between images of dogs and cats, there is no set recipe to follow that would, say, direct us to first look for triangular ears and then to look at the length of the snout before arriving at a decision. Mullainathan and Spiess (2017) contend that the real breakthrough occured when analysts stopped approaching intelligence tasks procedurally, but instead approached them empirically. For example, in the case of computer vision and image recognition – where models called Convolutional Neural Networks have distinguished themselves (LeCun et al., 2015; Goodfellow et al., 2016) – the algorithm will first learn a set of low-dimensional features such as edges

[10] In this paper, I do not spend time on definitional debates about what constitutes artificial intelligence (AI), machine learning (ML), and deep learning (DL), and what distinguishes them from statistics. Instead, I adopt the simplest operational definitions used across computer science (Goodfellow et al., 2016) and computational social science (Brady, 2019). AI is a broad concept that refers to computer-led intelligent tasks. ML is a subset of AI where tasks are accomplished by feeding data to an algorithm that improves the more data it is exposed to. DL is a subset of ML where learning occurs through a hierarchy of concepts, where each concept is defined through its relation to a simpler concept, and where the computer learns the complicated concept by building on the simpler one (LeCun et al., 2015; Goodfellow et al., 2016).

and shadows which are then encoded into progressively more complex elements (e.g., an ear). The point is that the analyst does not attempt to deduce the rules that would help discriminate between a dog and a cat; instead the algorithm lets the data find the rules that work best to accomplish the task at hand.

Though in social sciences we tend to work with structured tabular data rather than pixels, the principle is the same. Complex socio-political phenomena are difficult to capture and represent as a set of equations. Certainly, the role of theory is to come up with a parsimonious model of that phenomenon, but it can only ever be a stylized representation of the "true" state of the world, and models more often than not impose strong assumptions on the nature of the problem. However, with a wicked problem, we do not actually know the definitive nature of the problem until we have solved it. In other words, we struggle to specify the true data-generating process, because it might contain non-linearities and complex interactions in many dimensions. Yet machine learning techniques are exceptionally good at fitting flexible functional forms. A neural network, for instance, can approximate any continuous function arbitrarily well (Hastie et al., 2017). Machine learning tools estimate the conditional value of an outcome variable given a set of independent variables, but without making many assumptions about the structure of the relationship. And in the case of intractable wicked problems, I submit that this is a desirable attribute. More worryingly, by explicitly attempting to specify the complex relationships that underpin a wicked problem, we will – according to the definition of a wicked problem – misspecify the model. Machine learning techniques are thus particularly good at finding generalizable structure in a problem because the functional form is determined by the data. That is, model selection occurs automatically in the process of finding the best predictions of the outcome.

There are two other situations that are relevant to wicked problems where machine learning methods may be advantageous: (1) in the case of heterogeneous treatment effects, and (2) in high dimensional settings. Machine learning can illuminate heterogeneous "treatment" effects where there exists heterogeneity in responses with respect to observed covariates (Athey, 2019; Storm et al., 2020). In a linear regression setting, heterogeneous effects might be estimated via the coefficients on interaction terms (Mullainathan and Spiess, 2017), yet they can also be construed as a prediction problem of mapping unit-level attributes to individual effect estimates. This view underpins many existing marketing and recommendation systems in business applications (e.g., a web page renders customized advertisements according to the browsing habits of the internet user) and is being increasingly used in precision medicine to provide individualized treatment plans (Obermeyer and Emanuel, 2016; Ge et al., 2020). Predictive tasks can also be helpful to tackling policy problems. I demonstrate this in chapter 4 of the dissertation where the problem of missing data in developing countries can be abated by generating machine learning predictions that can be used to augment the measure of illicit finance that is developed in chapter 3.

Another attractive feature of machine learning is that it offers a principled way of dealing with high dimensional settings, which is where there are a large number of covariates K, sometimes to the point where $K >> N$ and the number of variables is much higher than the number of observations. High dimensionality poses several challenges. First, when $K > N$ estimation is infeasible, since we have more regressors than data points and hence negative degrees of freedom. Further, when there are many covariates that we think might interact to predict the outcome but we don't know how, it would not be sensible to, say, include all pairwise interactions of the explanatory variables in our model. By contrast, model selection with machine learning is data-driven and systematic, because ML searches for those interactions automatically. A key way in which machine learning deals with high dimensional settings is through regularization, a procedure that penalizes variables that are not informative. In standard econometric approaches, high dimensionality will also manifest as multicollinearity, where some variables are highly correlated with each other, which would be reflected by large standard errors on those coefficients. Because most ML techniques do not provide standard errors, multicollinearity in ML settings is effectively dealt with through regularization. The data-driven approach of ML in dealing with high dimensional settings is useful for wicked problems, where the covariate space is plausibly large. Given that in wicked problems, the formulation of the problem is inextricable from the planner's (subjective) view of how to solve it, it might be productive to let the data speak for itself and use a transparent and empirical approach to extracting the predictors that have the most informational value for explaining the variation in the outcome.

A particularly useful heuristic for the study of wicked problems that can be borrowed from machine learning approaches is the concept of regularization. Regularization in ML is the process of selecting important variables in a data-driven way. These techniques can help us extract information from a messy and noisy dataset and "distill the essence of the data" (Grimmer et al., 2021, p. 402). Implicit in this view is the notion that we can represent a high-dimensional covariate space in lower-dimensional space because there are some key latent factors that are responsible for the phenomenon we observe. The core idea behind regularization – that a high dimensional space can be expressed in low-dimensional terms – offers us a lot of analytical traction when dealing with wicked problems, because it allows analysts to delineate the scope of the problem by identifying the variables or dimensions that explain a lot of the variance in the outcome. Wicked problems encompass so much social complexity that it seems quixotic to attempt to completely characterize them with a parametric model, yet we can start to make progress by thinking of them as being constituted of latent factors that operate in unknown and complex ways to produce the phenomenon that we observe. Those latent factors operate in a low-dimensional space that can be interpreted as representing an underlying property of the wicked problem (Grimmer et al., 2021).[11] By definition, there is no definitive formulation of a wicked problem: we do not know f. The generative process in nature that gives rise to illicit finance Y is a black box that contains some predictive features, some parameters, and noise. Machine learning techniques are criticized for precisely this reason: the function $f(X)$ that maps the feature space X to the outcome Y is a black box. The way that ML deals with the "black box" problem is to let the data tell us what variables are important through regularization. This then allows for dimension reduction, which is critical to drawing meaningful conclusions from the data. Machine learning approaches take the position that it is better to use a very flexible model constrained by regularization than to constain the model *ex ante* by using few predictors (Grimmer et al., 2021).

The classic bias-variance trade-off means that a very complex model might overfit the training data and generalize poorly out-of-sample. ML methods deal with this by adding a penalty parameter to the model that discourages model complexity. The severity of the penalty parameter is determined empirically by tuning the model (usually through a process known as cross-validation) to find the amount of flexibility in a model that yields the best predictive performance. The data-driven way in which ML tunes model parameters is well-suited for wicked problems, since model selection is accomplished empirically, rather than by a subjective choice of the analyst. However, one common critique levelled against machine learning is that the estimated representation of the data may be difficult to interpret. However, there is no guarantee that the tuned model will be simple or easy

[11] Note that regularization on high-dimensional data does necessitate the assumption that the true model is "sparse" (Belloni et al., 2014): that it is possible to reconstruct the data relatively well using a low-dimensional representation of the covariate space.

to interpret. Occam's razor holds that there is a trade-off between parsimony and accuracy: more complex models are usually better predictors than simple and interpretable ones. This is a valid concern, but Breiman (2001) proposes an alternative outlook on the topic that can be a useful frame for the study of wicked problems: the goal is to extract information, and interpretability is the means to that end. A model does not have to be simple to provide reliable information about the relationship between X and Y.

Machine learning is not without its drawbacks. Next, I highlight three challenges in the application of machine learning to the study of wicked problems: the difficulty in incorporating uncertainty around parameter estimates, the fact that coefficient estimates $\hat{\beta}$ are rarely consistent, and the pitfalls of naively inferring structure about the data-generating model from machine learning outputs.

First, one drawback of ML techniques is that it is difficult to obtain correct standard errors on the *coefficient* estimates because the data was used for model selection. In the econometric approach, researchers will rely on statistical theory to estimate confidence intervals for their estimated parameters (Athey, 2019). By contrast, ML methods have struggled with providing valid confidence intervals, even if only asymptotically (Grimmer et al., 2021). This is because it is hard to know how to incorporate uncertainty when the data itself has been used for model selection (Athey, 2019; Mullainathan and Spiess, 2017; Athey and Imbens, 2015). Since ML techniques use the "training" data to learn the representation of f that yields the best predictive performance, any parameter estimates that are generated in the course of estimating the model will have to reflect uncertainty around model selection itself. Mullainathan and Spiess (2017) point out that this is a ubiquitous problem in machine learning: not only do the lack of standard errors make it hard to make inferences on parameters *after* model selection, but this is also a problem of the consistency of the model selection itself. Therefore, ML algorithms have to be modified to provide valid confidence intervals for estimated parameters when the data is used to select the model (Athey, 2019). One approach proposed by Athey and Imbens (2016) relies on what they call an "honest" approach to estimation, which is accomplished through sample-splitting. With an application to decision trees, they use one sample to construct the partition (i.e., fit the model), and the other sample to estimate treatment effects. They show that confidence intervals built around the estimates in the second sample will have nominal coverage (Athey and Imbens, 2016).

Even though the empirical approach to model selection in ML makes it a challenge to provide standard errors, there is one ancillary benefit that algorithmic model selection offers. It is a systematic, data-driven way of selecting a model that provides both superior performance and is reproducible (Athey, 2019). Athey (2019) makes the point that, in practice, applied researchers may test a variety of econometric specifications behind the scenes when they are performing model selection, yet only report a few specifications as a robustness check. According to Athey (2019), this practice is rampant yet researchers are not often honest about it because it would invalidate many of the reported confidence intervals around their coefficients due to the multiple comparison problem. Note that this is still a problem, even if researchers are otherwise honest and do not engage in p-hacking or cherry-picking specifications (Ferman et al., 2020). Compared to this, the algorithmic approach to specification searches has the advantage of being systematic, transparent, and reproducible.

Second, another drawback of machine learning concerns the consistency of the $\hat{\beta}$s. Even when a ML model produces coefficient estimates $\hat{\beta}$ in the course of making its predictions, these estimates are rarely consistent (Mullainathan and Spiess, 2017). Even if a model generates predictions $\hat{Y}$ of the outcome that are robust and of good quality (as measured by predictive accuracy in a test dataset), the coefficients $\hat{\beta}$ on the variables X used to arrive at these predictions can vary with small perturbations of the dataset. Mullainathan and Spiess (2017) perform an experiment that illustrates this problem. They set out to estimate house prices from a variety of predictors (e.g., square footage) from the American Housing Survey. They randomly split the sample into 10 different partitions of 5,000 observations. On each partition, they use a LASSO regression to estimate house prices with a fixed penalty parameter. The LASSO (Least Absolute Shrinkage and Selection Operator) (Tibshirani, 1996) is a regression model that induces sparsity by shrinking some of the coefficient estimates to exactly 0, with the amount of shrinkage determined by a given penalty parameter. Therefore, when estimating a LASSO model, not every variable ends up being explanatory of the outcome (if the corresponding estimated $\hat{\beta}$ is 0). By estimating a LASSO on 10 different sub-samples of the data, Mullainathan and Spiess (2017) show that different explanatory variables end up being used each time, such that no stable patterns are detected, even though the R^2 denoting prediction quality remains constant from partition to partition. This is a problem of consistency of the model selection itself. Breiman (2001) refers to this as the Rashomon Effect of the multiplicity of good models:[12] there exist multiple models of the data $f(X)$ that are just as good to predict the outcome Y. That is, different functional forms that combine the variables X in multiple ways can yield the same prediction error rate. This effect is likely to be at play with wicked problems too: there are multiple ways in which the problem can be characterized that are just as good, "good" depending on the conceptualization of the planner.

A final caveat with the application of ML to wicked problems is that it is hard to recover the structure of the model from the estimated coefficients. Mullainathan and Spiess

[12]Named after a movie where the plot revolves around 4 people testifying at a trial about their recollection of the same crime that they witnessed. Even though the facts of the crime that the 4 people report are the same, the stories that they recount about what happened are very different (Breiman, 2001).

(2017) warn against the temptation to use the estimated $\hat{\beta}$ parameters to try and learn something about the underlying data-generating processs. After all, when predictive performance is high, some structure in $\hat{Y}$ must have been found. However, the lack of consistency of the parameter estimates will prevent us from making credible inferences on the underlying structure. On the other hand, Mullainathan and Spiess (2017) make the point that making some assumptions on the data-generating process would allow us to take the $\hat{\beta}$ more seriously. Thus, the challenge of the analyst studying wicked problems is to decide whether to try and recover some of the structure of the problem by placing some assumptions on the data-generating process that delineate the problem *ex ante*, or whether to stick to predictive tasks only. Ultimately, this is a judgement call for the analyst depending on the question at hand. In chapter 4 of this dissertation, the latter option is chosen because the task is a purely predictive exercise that seeks to generate reliable predictions in order to address a missing data problem.

1.5 Permissions and Attributions

1. The contents of chapter 2 and appendix A are the result of a collaboration between Alice Lépissier and Matto Mildenberger, and have previously been published as: Lépissier, A. Mildenberger, M. (2021). "Unilateral climate policies can substantially reduce national carbon pollution". *Climatic Change*, 166:31. DOI:10.1007/s10584-021-03111-2. The article is made available under a Creative Commons Attribution 4.0 International License.[13]

2. Some of the contents of chapter 3 are the result of a collaboration between Alice Lépissier, William Davis, and Gamal Ibrahim. The methodology section of this

[13]To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

chapter draws on an unpublished methodological note prepared for the United Nations Economic Commission for Africa (UNECA) by Alice Lépissier. Part of this work was carried out while Alice Lépissier was employed as a consultant for UNECA during portions of 2018 and 2019. The methodology has since been substantially refined. Results based on an early version of this methodology are included in the *Financing for Sustainable Development Report 2019* of the Inter-agency Task Force on Financing for Development (New York, NY: United Nations), available at https://developmentfinance.un.org/sites/developmentfinance.un.org/files/FSDR2019.pdf. The contents of the methodology and findings sections in this chapter are reproduced with the permission of William Davis and Gamal Ibrahim. The methodology is developed in the authors' personal capacity and does not necessarily reflect the views of their respective institutions. Alice Lépissier gratefully acknowledges financial support from the United Nations Economic Commission for Africa.

1.6 Statement of Scientific Reproducibility

Every effort has been made to provide an entirely reproducible analytical pipeline; starting with the acquisition of raw data, data cleaning, statistical analyses, and generating results and data visualizations in the final step. The code base for each project in this dissertation is publicly avaible in online repositories. The input data required to produce results for each project is available online (data acquisition is either automated or details on the sources are provided in the relevant script files). Additional data products are available upon request.

1. The code for chapter 2 is available at https://github.com/walice/synth. Raw

data and results are available at https://doi.org/10.5281/zenodo.4566803.

2. The code for chapter 3 is available at https://github.com/walice/Trade-IFF. The full "atlas" database is available at https://doi.org/10.5281/zenodo.3610557.

3. The code for chapter 4 is available at https://github.com/walice/illicitAI.

Bibliography

Abadie, A., Diamond, A., and Hainmueller, J. (2010). Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. *Journal of the American Statistical Association*, 105(490):493–505.

Allred, B. B., Findley, M. G., Nielson, D., and Sharman, J. C. (2017). Anonymous shell companies: A global audit study and field experiment in 176 countries. *Journal of International Business Studies*, 48(5):596–619.

Angrist, J. D. and Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press, Princeton, NJ.

Aronow, P. M. and Miller, B. T. (2019). *Foundations of Agnostic Statistics*. Cambridge University Press, Cambridge, UK.

Athey, S. (2017). Beyond prediction: Using big data for policy problems. *Science*, 355:483–485.

Athey, S. (2019). The Impact of Machine Learning on Economics. In Agrawal, A., Gans, J., and Goldfarb, A., editors, *The Economics of Artificial Intelligence*, pages 507–547. University of Chicago Press.

Athey, S. and Imbens, G. (2015). Machine Learning Methods in Economics and Econometrics: A Measure of Robustness to Misspecification. *American Economic Review: Papers & Proceedings*, 105(5):476–480.

Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences of the United States of America*, 113(27):7353–7360.

Athey, S. and Imbens, G. W. (2019). Machine Learning Methods That Economists Should Know About. *Annual Review of Economics*, 11:685–725.

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

Barder, O., Evans, A., and Lépissier, A. (2015). SkyShares: Modelling the distributive and economic implications of a future global emissions budget. *Center for Global Development Policy Paper*, 067:1–25.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). High-Dimensional Methods and Inference on Structural and Treatment Effects. *Journal of Economic Perspectives*, 28(2):29–50.

Blankenburg, S. and Khan, M. (2012). Governance and Illicit Flows. In Reuter, P., editor, *Draining Development? Controlling Flows of Illicit Funds from Developing Countries*, chapter 2, pages 21–68. World Bank Group, Washington, DC.

Brady, H. E. (2019). The challenge of big data and data science. *Annual Review of Political Science*, 22:297–323.

Brady, H. E. and Collier, D. (2004). *Rethinking social inquiry: diverse tools, shared standards*. Rowman & Littlefield, Lanham, MD.

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

Christensen, J. (2012). The hidden trillions: Secrecy, corruption, and the offshore interface. *Crime Law and Social Change*, 57(3):325–343.

Cobham, A. (2014). Benefits and Costs of the IFF Targets for the Post-2015 Development Agenda. *Copenhagen Consensus Center Assessment Paper*, pages 1–31.

Cobham, A. and Janský, P. (2020). *Estimating Illicit Financial Flows - A Critical Guide to the Data, Methodologies, and Findings*. Oxford University Press, Oxford, UK.

Conklin, J. (2006). *Dialogue Mapping: Building Shared Understanding of Wicked Problems*. John Wiley & Sons, Chichester, UK.

Diamond, A. and Sekhon, J. S. (2013). Genetic Matching for Estimating Causal Effects: Balance in Observational Studies. *The Review of Economics and Statistics*, 95(3):932–945.

Ferman, B., Pinto, C., and Possebom, V. (2020). Cherry Picking with Synthetic Controls. *Journal of Policy Analysis and Management*, 39(2):510–532.

Ge, Q., Huang, X., Fang, S., Guo, S., Liu, Y., Lin, W., and Xiong, M. (2020). Conditional Generative Adversarial Networks for Individualized Treatment Effect Estimation and Treatment Selection. *Frontiers in Genetics*, 11(585804):1–17.

Genovese, F. (2019). Sectors, Pollution, and Trade: How Industrial Interests Shape Domestic Positions on Global Climate Agreements. *International Studies Quarterly*, 63(4):819–836.

Gerber, A. S. and Green, D. P. (2012). *Field Experiments: Design, Analysis, and Interpretation*. W. W. Norton & Company, New York, NY.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press, Cambridge, MA.

Grimmer, J., Roberts, M. E., and Stewart, B. M. (2021). Machine Learning for Social Science: An Agnostic Approach. *Annual Review of Political Science*, 24:395–419.

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

Haug, C., Rayner, T., Jordan, A., Hildingsson, R., Stripple, J., Monni, S., Huitema, D., Massey, E., van Asselt, H., and Berkhout, F. (2010). Navigating the dilemmas of climate policy in Europe: Evidence from policy evaluation studies. *Climatic Change*, 101(3):427–445.

Hildén, M. (2011). The evolution of climate policies - The role of learning and evaluations. *Journal of Cleaner Production*, 19(16):1798–1811.

Imbens, G. and Athey, S. (2021). Breiman's Two Cultures: A Perspective from Econometrics. *Observational Studies*, 7(1):127–133.

Imbens, G. W. (2020). Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics. *Journal of Economic Literature*, 58(4):1129–1179.

Intergovernmental Panel on Climate Change (2014). *Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. IPCC, Geneva, Switzerland.

King, G., Keohane, R. O., and Verba, S. (1994). *Designing social inquiry: Scientific inference in qualitative research*. Princeton University Press, Princeton, NJ.

Kleinberg, J., Ludwig, J., Mullainathan, S., and Obermeyer, Z. (2015). Prediction policy problems. *American Economic Review*, 105(5):491–495.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Levin, K., Cashore, B., Bernstein, S., and Auld, G. (2012). Overcoming the tragedy of super wicked problems: Constraining our future selves to ameliorate global climate change. *Policy Sciences*, 45(2):123–152.

Lonergan, E. and Blyth, M. (2020). *Angrynomics*. Agenda Publishing Limited, Newcastle upon Tyne, UK.

Lowery, C. and Ramachandran, V. (2015). *Unintended Consequences of Anti–Money Laundering Policies for Poor Countries - A CGD Working Group Report*. Center for Global Development, Washington, DC.

Mahoney, J. and Goertz, G. (2006). A tale of two cultures: Contrasting quantitative and qualitative research. *Political Analysis*, 14(3):227–249.

Mildenberger, M. (2016). Process-tracing, counterfactual comparisons and causal inference: A new bridge over the qualitative-quantitative divide.

Mildenberger, M. (2020). *Carbon Captured: How Business and Labor Control Climate Politics*. MIT Press, Cambridge, MA.

Mullainathan, S. and Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106.

Obermeyer, Z. and Emanuel, E. J. (2016). Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. *New England Journal of Medicine*, 375(13):1212–1216.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4):669–688.

Pearl, J. (2009). *Causality*. Cambridge University Press, Cambridge, UK, second edition.

Pickering, J., Jotzo, F., and Wood, P. J. (2015). Sharing the Global Climate Finance Effort Fairly with Limited Coordination. *Global Environmental Politics*, 15(4):39–62.

Pol, R. F. (2020). Anti-money laundering: The world's least effective policy experiment? Together, we can fix it. *Policy Design and Practice*, 3(1):73–94.

Radford, J. and Joseph, K. (2020). Theory In, Theory Out: The Uses of Social Theory in Machine Learning for Social Science. *Frontiers in Big Data*, 3(18):13.

Ratha, D., Plaza, S., and Dervisevic, E. (2016). *Migration and Remittances Factbook 2016*. World Bank Group, Washington, DC.

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

Rittel, H. W. J. and Webber, M. M. (1973). Dilemmas in a General Theory of Planning. *Policy Sciences*, 4:155–169.

Rogelj, J., Shindell, D., Jiang, K., Fifita, S., Forster, P., Ginzburg, V., Handa, C., Kheshgi, H., Kobayashi, S., Kriegler, E., Mundaca, L., Séférian, R., and Vilariño, M. (2018). Mitigation Pathways Compatible with 1.5°C in the Context of Sustainable Development. In Masson-Delmotte, V., Zhai, P., Pörtner, H.-O., Roberts, D., Skea, J., Shukla, P., Pirani, A., Moufouma-Okia, W., Péan, C., Pidcock, R., Connors, S., Matthews, J., Chen, Y., Zhou, X., Gomis, M., Lonnoy, E., Maycock, T., Tignor, M., and Waterfield, T., editors, *Global Warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to* *the threat of climate change, sustainable development, and efforts to eradicate poverty*, chapter 2, pages 93–174. IPCC.

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

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

Storm, H., Baylis, K., and Heckelei, T. (2020). Machine learning in agricultural and applied economics. *European Review of Agricultural Economics*, 47(3):849–892.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

UN General Assembly (2015). Transforming our world: the 2030 Agenda for Sustainable Development.

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28(2):3–28.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge, MA, second edition.

Chapter 2

Unilateral climate policies can substantially reduce national carbon pollution

2.1 Introduction

In recent years, policies to reduce greenhouse gas emissions (GHG) have been deployed at a rapid pace across the world. While scholars have extensively debated the theoretical merits of different types of policy instruments, we still do not know enough about the extent to which these policies work in the context of real-world, practical implementations. Identifying the specific impacts of climate policies on environmental outcomes is a difficult task. GHG emissions pervade industrial economies as the by-product of transportation, energy and manufacturing processes. As a result, nearly every significant economic trend shifts carbon pollution patterns (Schleich et al., 2001; Peters et al., 2012). Moreover, the adoption of carbon pricing policy is endogenous: countries might introduce climate mitigation policies as their emissions are already falling (Downs et al., 1996). Consequently, most efforts to identify the effect of specific climate reforms on carbon pollution levels are either *ex ante* economic simulations or *ex post* sectoral impact analyses. These models excel at simulating how policy instruments will affect different sectors of the economy, identifying economic trade-offs, and exploring sector-specific policy effects. However, existing approaches struggle to evaluate the net causal effect of national policies because they compare realized outcomes to business as usual scenarios rather than counterfactual outcomes in the absence of the specific policy.

In this article, our contribution is to offer a national-level estimate of climate policy effectiveness without requiring assumptions about the pattern and shape of emissions trajectories, using the synthetic control method (SCM) (see also parallel SCM analysis by Bayer and Aklin (2020) on the European Union's Emissions Trading Scheme). SCM was developed to provide an empirically calibrated way of selecting comparison groups for policy impact analyses. It has since become a staple technique in policy impact analysis in fields such as comparative politics, economics, and criminology (Billmeier and Nannicini, 2013; Costalli et al., 2017; Heersink and Peterson, 2016; Robbins et al., 2017; Sills et al., 2015). The method offers a transparent and principled means of choosing comparison units that is blind to post-intervention outcomes; this means that researchers develop counterfactual scenarios without knowing how comparison group choice will shape their results.

Here, we evaluate the 2001 UK Climate Change Programme, a complex reform that included a carbon tax on large-scale energy users, industry-negotiated exemptions from the tax for meeting reduction targets, and a voluntary emissions trading scheme. The CCP was established in November 2000 to meet the Kyoto Protocol's EU-wide target of reducing emissions 8% by 2008-2012 compared to 1990 levels, including the country's more ambitious unilateral target of a 20% reduction by 2010, again compared to a 1990 baseline. The UK's CCP was one of the first comprehensive climate reform packages passed globally, in advance of action by most other OECD countries.

We leverage the synthetic control method to compare British emissions post-CCP to what would have happened if the policy had not been passed, rather than a stylized Business-As-Usual (BAU) or other benchmark scenario. We find evidence of substantial emissions reductions as a result of the policy: the UK's CO_2 emissions per capita were 9.8% lower relative to what they would have been if the CCP had not been passed.

SCM's ability to measure the causal effect of a complex, national climate policy contributes to debates over the potential efficacy of the current climate regime. Conventional accounts of global climate policy-making emphasize countries' weak incentives to act on climate change alone. Yet, we find that an early unilateral climate policy in the United Kingdom meaningfully reduced carbon pollution. The CCP was also effective despite the policy's hybrid nature (a combination of carbon pricing with negotiated industry agreements) and its substantial concessions to domestic polluters. Our findings thus provide evidence that even imperfect policy instruments can result in consequential reductions in national emissions.

2.1.1 Approaches to climate policy evaluations

Efforts to identify the effect of specific climate reforms on carbon pollution levels are typically *ex ante* economic simulations (Böhringer et al., 2005; Burniaux et al., 1992; Bruvoll and Larsen, 2004; Svendsen et al., 2001; Agnolucci, 2009; Hu et al., 2015) or *ex post* sectoral impact analyses (Ang et al., 2016; Martin et al., 2009; Future Energy Solutions, 2003). *Ex ante* approaches use Computable General Equilibrium (CGE) or Integrated Assessment models (IAM) to simulate the impact of a policy on a country's economy and environment. These models contain complex systems of equations that are stylistic representations of the relationships between different factors of production and agents in an economy, and (in the case of IAMs) the physical climate system. They are calibrated using historical data to reproduce the equilibrium state of an economy for a benchmark year. In general, CGE and IAM models can then compare a policy intervention against alternative reference scenarios that are chosen by the modeler, which often include scenarios of the form "climate stalemate" or total inaction, "Business-As-Usual" (BAU), or "optimal" scenarios where policies are implemented with welfare maximization (see for example Nordhaus (2013)).

Such models are useful to understand how a policy instrument is expected to affect different sectors of the economy, to identify potentially important trade-offs, and to derive comparative statics. However, theoretical predictions on how a carbon policy is expected to perform cannot take into account institutional and political barriers that emerge during policy enactment and implementation. These models also reflect complex assumptions on functional forms and parameter values that lead to highly divergent predicted outcomes between different models (Pindyck, 2017). For example, to generate BAU scenarios, modelers need to make assumptions about the growth rate of GDP, population, energy consumption elasticities (Böhringer et al., 2003), and (in the case of IAMs) environmental responses to these factors. Consequently, these models impose (often hidden) parametric assumptions on the hypothesized future emissions trajectories, leading some to criticize these approaches as akin to a "black box" (Böhringer et al., 2003; Pindyck, 2017) where model runs are not always grounded in empirical or theoretical realities. Moreover, *ex ante* models are calibrated using historical benchmark data (Böhringer et al., 2003) which often rely on outdated economic snapshots. For example, the model used to generate a BAU scenario to compare the effectiveness of UK climate policy in the early 2000s was calibrated using input-output tables from 1995 (see Ekins and Etheridge (2006)).

While CGE or IAM models offer clear advantages when conducting *ex ante* simulations about the general equilibrium effects of an exogenous policy treatment in comparison to a stylized reference scenario, the Business-As-Usual (BAU) scenarios they produce are not always appropriate to conduct *ex post* policy impact evaluations because those benchmarks are not clear counterfactuals for the policy outcome. In particular, the BAU assumption of *no action whatsoever* on climate is rarely the appropriate counterfactual to causally evaluate the effect of a climate policy. Rather, the counterfactual should be the potential outcome of carbon emissions in the absence of *that specific climate policy.*

Recognizing the weakness of these assumption-intensive counterfactuals, other analyses focus on *ex post* sectoral-level impacts rather than a policy's net capacity to decrease overall CO_2 emissions. The BAU scenarios in these cases are often rudimentary forward projections. For example, the consulting firm tasked by the UK government's Department for the Environment, Food and Rural Affairs (DEFRA) to estimate the results of the UK's climate change policies presents performance results as energy savings compared to what energy would have been used if sectors had produced the same throughput but at the energy-efficiency of a reference year (Future Energy Solutions, 2003, p. 13). Other studies use micro-level data or case studies to estimate the impact of the CCP on businesses (Ang et al., 2016; Martin et al., 2009). However, the net national impact of a policy is the most important measurement with respect to climate change risk mitigation (Allen et al., 2009), and these methods don't allow for *ex post* assessment of this critical feature.

By contrast, synthetic control methods (SCM) allow for causal identification of the net national impact of a policy, offering a different form of *ex post* policy impact evaluation that supplements existing approaches. In general, the SCM has been referred to as "arguably the most important innovation in the policy evaluation literature in the last 15 years" (Athey and Imbens, 2017, p. 10). While it is not possible to enumerate all of the possible drivers of CO_2 emissions in a given country and to specify how they interact, the synthetic counterfactual approach uses a diverse sample of countries to capture all of these latent trends in a way that does not require out-of-sample extrapolation (Abadie and Gardeazabal, 2003; Abadie et al., 2010, 2011, 2015).

This approach to causal identification of policy impacts is grounded within the potential outcomes framework (Holland, 1986; Rubin, 1974). Synthetic control methods borrow some elements from matching and difference-in-difference strategies. Matching is often used as part of selection-on-observables strategies, and aims to identify causal treatment effects by making the distributions of covariates that may impact an outcome as similar as possible between the treated and the control units. If the goal is to estimate the causal impact of some treatment T on some outcome Y, matching on some covariates

X that also impact Y may help attenuate bias. However, in the presence of unobserved confounders Z, matching will not identify the causal effect of treatment. Difference-in-difference strategies exploit panel data to identify causal effects, and control for time-invariant confounders across treatment and control groups. In addition, they assume that time-varying confounders do not vary across treatment and control groups, often referred to as the "parallel trends" assumption. By contrast, SCM does not require us to make this assumption, and can accommodate time-varying unobserved confounders. The problem can also be restated as one of estimating a latent factor model, where a linear combination of time-varying trends (e.g. demand for energy) and time-fixed confounders drive a country's per capita emissions. The goal then becomes to capture the same combination of those confounders in the donor pool, in order to replicate the same factors driving the treated country's emissions. These confounders are then "differenced out" when we compare the emissions trajectories of the treated country and its synthetic control (Hazlett and Xu, 2018; Xu, 2017). We further explicate the synthetic control method in the Methods section.

2.1.2 The 2001 UK Climate Change Programme

Our empirical focus is an evaluation of the UK's 2001 Climate Change Programme (CCP), one of the first major reform packages passed by any OECD country. The CCP was established in November 2000 to meet the Kyoto Protocol's EU-wide target of reducing emissions 8% by 2008-2012 compared to 1990 levels, including the country's more ambitious unilateral target of a 20% reduction by 2010, again compared to a 1990 baseline.

The CCP included three interlocking policy instruments: first, a Climate Change Levy (CCL) on large-scale energy users (including the public sector); second, sector-wide Climate Change Agreements (CCA) negotiated between with industry and government that discounted CCL rates if sectors hit pre-negotiated emissions reduction targets; and third, a voluntary unilateral emissions trading scheme (ETS).

The first of these components was the Climate Change Levy which came into effect in April 2001. The CCL taxed the energy intensity of different fuel sources. It was passed alongside a 0.3% reduction in employer National Insurance Contributions (NICs) and new renewable energy-oriented R&D funds. The CCL was not a pure carbon tax. While it did exempt most forms of renewable energy, it still included carbon-free nuclear energy. The CCL was levied on non-domestic consumers only, including the business and the public sectors.

The policy offered substantial producer flexibility through its second interlocking policy instrument, industry-level Climate Change Agreements (CCAs). CCAs exempted businesses from up to 80% of the levy if they agreed on voluntary carbon pollution reduction benchmarks[1]. By 2002, 44 sectoral associations had signed CCAs, including aluminum and steel (Bailey and Rupp, 2005). Performance under these agreements was assessed at the sector level, but it was possible for individuals to continue under the program even if their broader sector failed to meets its target. Under the CCAs, industry could choose their own base years, which ranged from 1990 through 1999, and could set targets in different accounting "currencies" (i.e. relative energy: GJ primary energy per unit ton of production; relative carbon: tons of carbon per unit ton of production; absolute energy: GJ; absolute carbon: tons of carbon).

Additional producer-oriented flexibility was introduced with the April 2002 UK Emissions Trading Scheme (ETS), a voluntary program that allowed participants to trade

[1]If private sector actors failed to meet these negotiated benchmarks, they would be forced back into the CCL system for at least two years.

emissions reduction permits relative to an absolute target baseline (the average of a participant's 1998-2000 emissions); CCA signatories could then buy and trade these permit as insurance against failure to meet CCL carbon pollution reduction benchmarks. Conversely, sectors who over-complied with their CCA targets were able to sell their excess permits on the UK's Emissions Trading Scheme.

In the appendix, we describe each of these policy components in more detail as part of a narrative history of UK climate policy-making from the 1980s through to 2015. We also detail the political controversy that accompanied the introduction of the CCP.

The CCP is particularly well-suited to synthetic control analysis. The UK was one of the first European countries to implement a comprehensive national climate reform package, and was the first country to unilaterally enact a domestic emissions trading scheme. With the exception of Northern European countries that enacted modest carbon tax systems in the early 1990s, most OECD countries had only implemented voluntary climate reforms up until 2005, when the EU emissions trading scheme began. This creates a window from 2001 through 2005 where domestic UK action largely stands alone against its peers. This allows us to construct of a credible counterfactual for the United Kingdom while avoiding possible policy diffusion effects from other countries.

2.2 Materials and Methods

2.2.1 Causal identification using synthetic control methods

We use the Synthetic Control Method to generate a "synthetic UK" as a weighted average of other OECD, upper middle, and high income countries in our sample, or "donor pool". Countries in the donor pool are selected through an algorithm so that the pre-CCP emissions trajectories of the UK and of the synthetic UK match each other as closely as possible. We then evaluate the causal effect of the UK's Climate Change Programme by comparing the trajectory of emissions in the "synthetic UK" with the observed post-treatment emissions in the UK.

More formally, assume a sample of $J+1$ countries where $j=1$ corresponds to the treated United Kingdom, and $J=\{2,\ldots,J+1\}$ is our donor pool. The intervention (i.e. the passage of the CCP) occurs at T_0+1 and so the pre-invervention time periods are indexed by $t=1,2,\ldots,T_0$ and the post-intervention time periods are indexed by $t=T_0+1,T_0+2,\ldots,T$. Let Y_{1t}^C represent the potential outcome under control for the UK, where $j=1$ indexes the UK. These are the potential CO_2 emissions in the UK if the CCP had *not* been passed. Let Y_{1t}^T represent the potential outcome under treatment; which are the potential CO_2 emissions in the UK if the CCP *had* been passed. The causal impact of the CCP is the difference between the two, and so our estimand of interest is $\alpha_{1t}=Y_{1t}^T-Y_{1t}^C$. However, Y_{1t}^C is unobserved.

Consider the following $J\times 1$ vector $\mathbf{W}=(w_2,\ldots,w_{J+1})^{\mathrm{T}}$ which contains the weights that reflect how much the jth candidate in the donor pool contributes to the synthetic counterfactual for the UK emissions trajectory. These weights are restricted to be non-negative and sum to 1, that is, $w_j\geq 0$ for $j=2,\ldots,J+1$ and $\sum_{j=2}^{J+1}w_j=1$. This restriction on the weights is imposed in order to avoid extrapolating when constructing the synthetic counterfactual (Abadie et al., 2010, 2015).

Let $\mathbf{X_1}$ be a $K\times 1$ vector of the pre-treatment values of the K predictor variables of CO_2 emissions in the UK. The $K\times J$ matrix $\mathbf{X_0}$ contains the corresponding values of the pre-treatment values of explanatory variables for the J control countries. In our case, the $K=11$ attributes correspond to pre-treatment values of the outcome variable chopped up into discrete segments corresponding to CO_2 per capita emissions in each pre-treatment time period, respectively. Using a specification which includes all pre-treatment lags of the outcome variable has been recommended as the benchmark specification, unless researchers have strong theoretical priors on how other covariates affect the outcome (Ferman et al., 2020).

The pre-intervention characteristics of the synthetic UK will be given by $\mathbf{X_1^*} = \mathbf{X_0}\mathbf{W}^*$. The optimal $\mathbf{W}^*$ should thus be chosen so as to minimize the distance $||\mathbf{X_1} - \mathbf{X_0}\mathbf{W}||$, in order to construct a synthetic counterfactual that best approximates the treated unit with respect to pre-treatment outcome values. In practice, the SCM implementation seeks a $\mathbf{W}^*$ that solves $\underset{\mathbf{W}^*}{\arg\min} \sqrt{(\mathbf{X_1} - \mathbf{X_0}\mathbf{W})^{\mathrm{T}}\mathbf{V}(\mathbf{X_1} - \mathbf{X_0}\mathbf{W})}$. $\mathbf{V}$ is a $K \times K$ positive semi-definite, diagonal matrix of weights applied to the K variables that predict CO_2 emissions. Therefore, the loss function is a scalar. The implementation of the SCM by its authors (Abadie and Gardeazabal, 2003) allows for the choosing of a custom $\mathbf{V}$ weight matrix. This can be a fruitful approach if we possess *a priori* knowledge on the relative predictive power of different explanatory variables. However, in the absence of strong priors, we follow Abadie and Gardeazabal (2003) and Abadie et al. (2011) and adopt a data-driven approach whereby the matrix $\mathbf{V}$ is the one that minimizes the mean square prediction error (MSPE) of the pre-treatment outcome variable, i.e. such that the average squared discrepancies between the pre-treatment CO_2 emissions of the UK and of the synthetic UK are minimized. A numerical optimization algorithm is used to solve for these optimal weights[2].

Finally, the observed emissions (pre- and post-treatment) of the UK are collected in a $T \times 1$ matrix $\mathbf{Y_1}$. The CO_2 emissions of the countries in the donor pool are recorded in a $T \times J$ matrix $\mathbf{Y_0}$. The emissions of the synthetic UK are simulated as $\mathbf{Y_1}^* = \mathbf{Y_0}\mathbf{W}^*$.

[2]We use the *Synth* package in R with the default optimization methods of Nelder-Mead and BFGS.

The estimated treatment effect is thus given by $\hat{\alpha}_{1t} = Y_{1t} - \sum_{j=2}^{J+1} w_{jt}^{*} Y_{jt}$.

Causal identification is achieved using SCM under less restrictive conditions than difference-in-difference strategies. First, there can be no treatment spillover to other countries in the donor pool. Although the authors of the SCM approach do not explicitly refer to this assumption as such, this assumption is the stable unit treatment values assumption, or SUTVA, which states that "[t]he potential outcomes for any unit do not vary with the treatments assigned to other units, and, for each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes" (Imbens and Rubin, 2015, p. 10). Second, to avoid interpolation bias, variables used to form the weights must be within the same support of the data for the treated unit and countries in the donor pool (Abadie et al., 2010, 2015). In other words, the variables used to form the weights must have values for the donor pool countries that are similar to those for those of the UK. This is because interpolation biases may be severe if the procedure interpolates across different regions with very different characteristics (Abadie et al., 2010).

In general, the United Kingdom during the early CCP era satisfies these conditions. The United Kingdom is the only country to be treated by the CCP in 2001, and is the only country in the sample that passed major climate legislation until the European Union launched its emissions trading scheme (EU ETS) in 2005. Our dependent variable is operationalized as CO_2 emissions per capita, which ensures that the outcome variable across regions is broadly on the same order of magnitude and thus avoids interpolation bias. Moreover, alternative specifications provided in appendix section A.7 also achieve a restriction of the data to a common support for all countries in the sample by employing a rescaled dependent variable (e.g. relative to a 1990 and a 2000 baseline, respectively). Running the synthetic control estimator on absolute CO_2 emissions levels is not appropriate given the variance in emissions levels across countries.

2.2.2 Data sources and sample selection

To implement the synthetic control method, we use data on CO_2 emissions and CO_2 emissions per capita from the World Bank's World Development Indicator (WDI) database, extracting indicators "EN.ATM.CO2E.KT" (CO_2 emissions in kilotons) and "EN.ATM.CO2E.PC" (CO_2 emissions per capita in metric tons), respectively. The CO_2 emissions measured are those stemming from the burning of fossil fuels and the manufacture of cement. We impute some missing data for Germany, Kuwait, and Liechtenstein using alternate data sources. This procedure is described in appendix section A.1.

We define our donor pool as the 51 countries which were either OECD members or classified by the World Bank as upper middle income or high income countries at the time of treatment in 2001, that had a population greater than 250,000, and that did not have a carbon pricing policy in place. The Work Bank classifies countries into income categories according to GNI per capita in US$. In fiscal year 2001, the World Bank classified high income (HIC) countries as those with GNI per capita above 9,265 US$, and upper middle income (UMC) countries as those with GNI per capita in the 2,996 US$ to 9,265 US$ range. In 2001, there were 47 high income countries, 38 upper middle income countries, and 30 OECD countries. Our donor pool is the union of those sets, minus countries for which data is missing or countries that were deemed "treated" in 2001, and minus countries with a very small population.

We determine whether countries in the sample were "treated" by building on the World Bank's *State and Trends of Carbon Pricing 2019* report (World Bank, 2019), albeit with some modifications. Even though the World Bank report notes that Poland had passed a carbon tax in 1990, we do not consider it "treated" until 2005 (the start of the EU ETS) because the Polish tax was so small in scope and incidence that it cannot be considered a materially important carbon pricing policy. Indeed, the Polish carbon tax of 1990 was less than 1 US$ per ton CO_2e and covered only 4% of the jurisdiction's emissions (World Bank, 2019).

Moreover, we consider the Netherlands to be "treated" in 2001, even though the World Bank report does not report the Netherlands as having a carbon tax. However, the Netherlands introduced a tax on energy in 1996, which complemented a tax on fuel that came into force in 1992. Tax rates were set as a function of CO_2 per energy content, and were estimated to be around NLG 30 per metric ton of CO_2 (Hoerner and Bosquet, 2001, p. 20).

The countries that were "treated" in 2001 were thus: Denmark (carbon pricing policy first passed in 1992), Estonia (2000), Finland (1990), Netherlands (1992), Norway (1991), Slovenia (1996), and Sweden (1992). These countries are excluded from the donor pool.

2.2.3 Specifications

In the main specification we report below, we construct this synthetic UK from a donor pool of countries that were either OECD, upper middle, or high income countries in 2001. We exclude small countries with a population less than 250,000 in 2001 since these may have different fundamental drivers of CO_2 emissions than the UK. Not all countries in this donor pool contribute equally to this synthetic control. In our main specification, 8 countries make up the effective sample (see figure A.1 in the appendix) accounting for 88% of the weights, with the other countries having weights of less than 1%. In figure A.2 of the appendix, we also display the CO_2 per capita emissions of the donor countries in the effective sample. In this specification, which generates the strongest pre-treatment fit and performs best according to diagnostics reported in the Findings section and in appendix section A.7, the counterfactual trend is estimated using a blend of 19% Poland, 19% Libya, 18% Bahamas, 16% Belgium, 6% Trinidad and Tobago, 5% Uruguay, 4% Luxembourg, and 1% Brunei. Here, the pre-treatment MSPE achieved with that donor pool was 1.24×10^{-4}. Figure A.1 in the appendix displays the weights applied to each country in the donor pool.

The fact that surprising countries, such as the Bahamas and Libya are part of the top donors, while an intuitively similar country like France is at the bottom should not be cause for concern. Rather, it suggests that there were latent, unobserved forces driving British emissions, and that a weighted combination of these forces was found in the top donor countries. Specifically, the synthetic control approach estimates a latent factor model with a linear combination of time-varying and time-invariant confounds. Some combination of the unobserved factors responsible for driving British emissions was also present in donor countries, which are then re-weighted to create a credible control for the UK.

Instead, an advantage of this effective donor pool is that it rules out spatial spillover effects[3]. One of the assumptions required for causal identification is that the treatment affected the treated unit only and did not spillover to other control units (the SUTVA assumption). Since the UK's untreated neighbors such as France and Germany are not part of the effective sample of countries used to generate the synthetic control, our results are not at risk of over-estimating the treatment effect of the CCP due to a violation of the SUTVA assumption.

As a robustness check, we also evaluate specifications generated by progressively smaller donor pools, again applying population filters: (1) on countries that were either OECD

[3]We thank an anonymous reviewer for making this point to us.

members or high income countries in 2001; and (2) on countries that were OECD members in 2001. The pre-treatment MSPE increases (indicating a poorer fit between the UK and the synthetic UK) as the donor pool decreases: from 5.24×10^{-4} (donor pool consisting of 2001 OECD and HIC countries) to 2.13×10^{-3} (donor pool consisting of 2001 OECD members). However, despite these specifications being slightly weaker from a SCM perspective, they still generate similar estimates of the effect of the UK policy (see section A.7 in the appendix). In this way, while we choose our specification in a principled way based on synthetic control method best practices, our results hold even for a range of donor pools that rely only on countries with substantively similar political and economic systems.

Generally, there are a multitude of observed and unobserved factors, both dynamic and constant in time, that drive British emissions in ways that are hard to specify *a priori*. Attempting to specify a functional form that would accurately reproduce the emissions trajectory of the UK is a difficult task. The advantage of the SCM is that it enables us to sidestep the need to enumerate all of the structural drivers of CO_2 emissions. By contrast, we employ a non-parametric approach where we find the combination of (latent) drivers in donor countries that serve as an appropriate control by numerically minimizing the distance between the pre-treatment trends of the UK and the control.

The predictor variables used to construct a synthetic UK are the pre-treatment values of per capita CO_2 emissions from 1990 to 2000, with no other covariates. Other covariates might be useful to improve the match between the UK's pre-CPP emissions and its synthetic counterpart. In section A.7.2 of the appendix, we show this was not the case, and therefore we report our estimates using pre-intervention values of the dependent variable only. Kaul et al. (2018) show theoretically that using all pre-treatment values of the outcome variable as separate predictors in the SCM algorithm leads to an optimization procedure that renders all other covariates irrelevant. We verify empirically that this is the case: specification 2 in the appendix uses 4 covariates as predictors (GDP per capita, renewable energy consumption, fossil fuel energy consumption, and energy use per capita), in addition to the pre-treatment values of per capita CO_2 emissions. The weights on the 4 covariates when constructing the synthetic UK are all 0.

We construct our synthetic UK on the basis of the lagged values of CO_2 emissions per capita alone for three reasons. First, doing so leads to an optimal pre-treatment fit between the UK and its synthetic control. Since the goal of SCM is to create a credible counterfactual for the treated unit in the absence of treatment, a guiding heuristic is to choose the specification that minimizes the distance in potential outcomes pre-treatment. Second, this research design choice minimizes the risk of specification searching on the part of researchers. Ferman et al. (2017, 2020) suggest that despite the advantage of the transparency of the SCM, researchers have some latitude to engage in specification-searching. By restricting our choice set to specifications that only include pre-treatment values of the outcome variable, we tie our hands at the outset. Third, we do not have strong theoretical priors on the types of covariates that would capture most of the drivers of British CO_2 emissions. While we may account for observable characteristics that correlate with the outcome, such as income per capita, this is by no means a guarantee that we would account for the *unobservable* characteristics that determine the pattern of emissions. Ferman et al. (2020) address this problem and recommend that in the case where researchers do not have strong theoretical priors on the covariates to use, a specification which uses all pre-treatment lags of the outcome variable should be used and reported as the benchmark specification. Nevertheless, as a robustness check, we also estimate the treatment effect using alternative specifications, which we report below and in further detail in section A.7 of the appendix.

2.3 Findings

2.3.1 Treatment effect of the CCP

We first construct a synthetic UK as a weighted average of the pre-treatment characteristics of countries in the donor pool, where weights are chosen so as to minimize the distance between the UK and its synthetic counterpart. The solid line in figure 2.1 displays the observed CO_2 emissions per capita path of the UK: the emissions trajectory remained relatively flat post-treatment. The dashed line represents the UK's emissions trajectory had the country not passed its 2001 reform, as estimated by SCM.

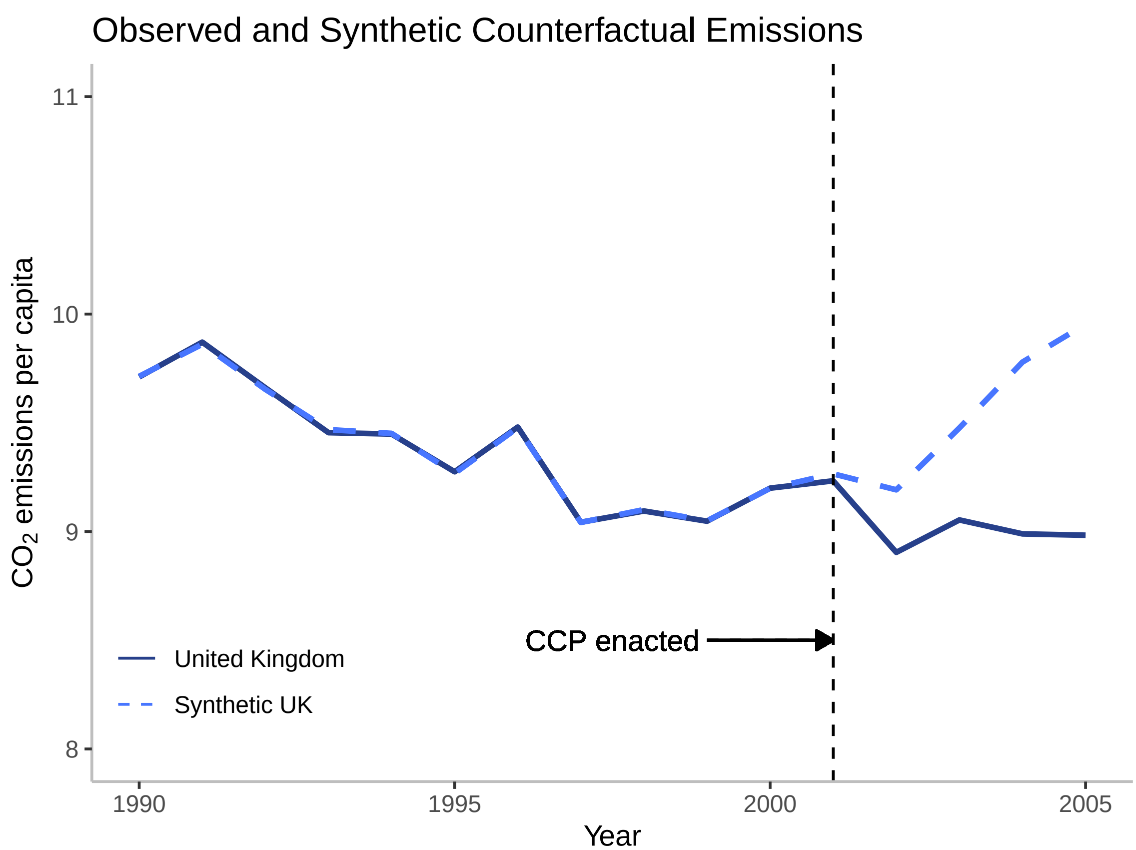


Figure 2.1: Observed and synthetic counterfactual per capita emissions for the UK. The solid line represents actual emissions trajectory. The dashed line represents the emissions trajectory of a synthetic UK, in the absence of the country's Climate Change Programme. Treatment occurred in 2001.

From 1990 to 2001, the difference in means between the pre-treatment CO_2 emissions of the UK and of the synthetic UK is statistically indistinguishable from 0 ($p = 0.981$)[4].

Figure 2.2 displays the difference between these pre-treatment CO_2 emissions in the UK and the weighted means and unweighted means, respectively. It indicates that the synthetic control achieves pre-treatment balance with the treated unit.

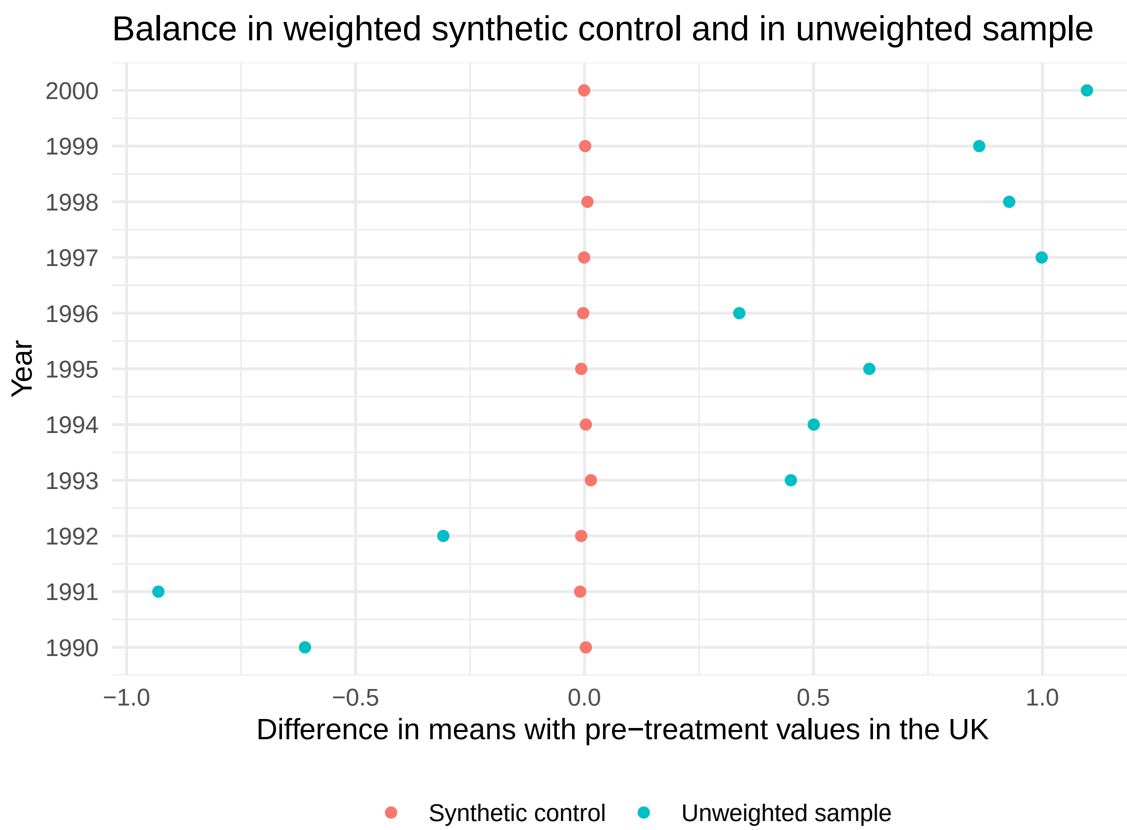


Figure 2.2: Difference in means in pre-treatment values observed in the UK and those estimated by the synthetic control (in orange) which is a weighted sample of the donor pool comprised of OECD, high, and upper middle income countries. Blue points represent the difference in means in pre-treatment values observed in the UK and those observed in the same donor pool sample, but unweighted.

[4]During the pre-treatment period, average emissions per capita in the entire sample were 6.3% lower than those of the UK's until 1992, and from 1993 onward they were 7.8% higher than those of the UK's. An unweighted sample is thus not an appropriate counterfactual for the UK.

However, after the 2001 passage of the Climate Change Programme, synthetic counterfactual emissions and observed CO_2 emissions start to diverge. The causal impact of the CCP can then be estimated as the difference in per capita emissions between the UK and the synthetic UK in the post-treatment period. By 2005, four years after the policy's passage, we estimate a treatment effect of -9.8% emissions per person in 2005. This is equivalent to a reduction of 148 Mt CO_2 during the period 2002-2005, an average annual reduction of 0.6 tons of CO_2 per capita. We do not estimate the causal impact of the CCP after 2005, since this corresponds to the launch of the EU-wide emissions trading scheme. After 2005, many countries in the donor sample are "treated" with comprehensive climate reform, and no longer act as appropriate donor countries.

We discuss the logic of our donor pool in the Methods section. However, it is important to (1) verify that our results are not dependent on the inclusion of certain countries in the donor pool, and (2) to re-run the synthetic control estimator on a donor pool of countries that have similar political and economic institutions as the UK. First, we run a "leave-one-out" robustness check that is detailed in a section below. We find that the findings are not dependent on the inclusion of any single country in the donor pool. Second, we also run the specification on a donor pool composed of 22 OECD countries that share institutional similarities with the UK. The top donors in this case are France (0.353), Japan (0.329), Belgium (0.123), Germany (0.099), Luxembourg (0.066) and Italy (0.018). The treatment effect attenuates slightly from -9.8% per capita emissions in 2005 to -5.3% per capita emissions, but retains statistical significance ($p < 0.05$). More details on this robustness check are provided in appendix section A.7.5.

2.3.2 Statistical inference

After estimating the treatment effect of the CCP on British emissions, we then ask whether our results are statistically significant, rather than the product of chance. Since SCM does not assume a data-generating process, nor do we estimate a specific functional form, we accomplish this through the use of falsification or placebo tests, rather than through parametric hypothesis testing. Placebo tests are commonly used in the literature to test whether an outcome or a unit that we know to be unaffected by treatment responds to a placebo treatment, in which case any positive treatment effect on the treated might be spurious (Bertrand et al., 2004; Abadie et al., 2010). To conduct our placebo analysis, we iteratively re-assign treatment to all countries in the donor pool. Since we know these countries were not treated, we should expect to see null treatment effects, other than by chance. The estimated treatment effect is given by the difference between the placebo unit and its synthetic control in post-treatment periods. This allows us to create a null distribution of gaps in post-treatment emissions trajectories for all countries in the sample. If the results in the UK are not driven by chance, we should expect the gaps in the post-CCP emissions trajectories in the UK to lie in the tails of that null distribution. This procedure is similar to testing Fisher's sharp null hypothesis, which tests a null hypothesis of no effect whatsoever (Imbens and Rubin, 2015).

However, it may be the case that the pre-treatment fit between a placebo unit and its synthetic control is poor. In this case, this particular placebo test is uninformative, since synthetic control estimators hinge on finding weights that minimize the distance in pre-treatment emissions trajectories. When the fit is poor, it is unlikely that the resulting synthetic counterfactual provides a credible control for the treated unit (placebo or otherwise). We thus exclude placebo countries with a pre-treatment MSPE greater than 30 times the pre-treatment MSPE of the UK in figure 2.3 below. However, the choice of cut-off for the treatment MSPE is rather arbitrary. We also provide figures in appendix section A.4 of the gaps between the treated unit and its synthetic control with cut-offs for excluding placebo runs that have a pre-treatment MSPE greater than 50 and 100 times that of the UK's for illustration.

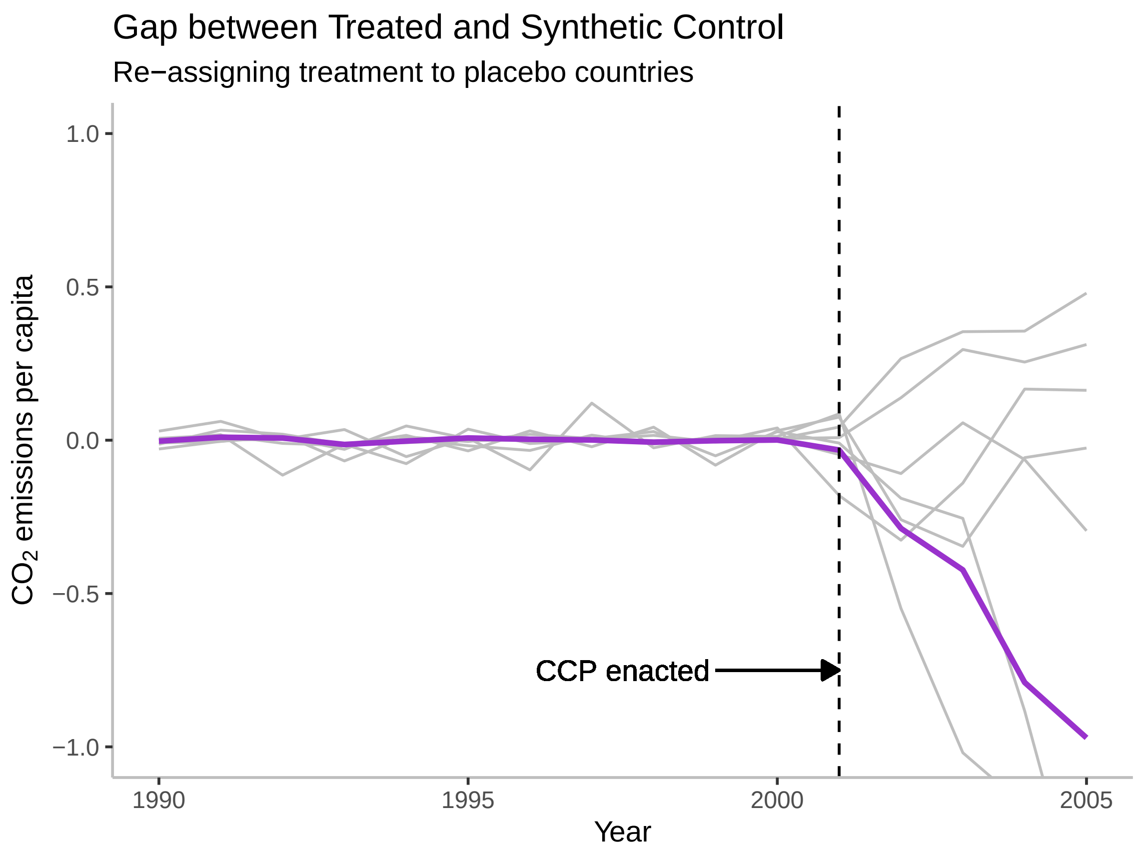


Figure 2.3: Gaps in emissions per capita between the treated unit and its synthetic counterpart. The thick purple line represents the gaps for the UK. The grey lines represent the distribution of placebo treatment effects. Countries with a pre-treatment MSPE greater than 30 times that of the UK have been excluded (see Methods for details).

Figure 2.3 displays the results of iteratively re-assigning treatment to countries in the donor pool (minus the UK). The purple line displays the gaps between the emissions in the UK and in the synthetic UK. The grey lines represent the gaps in emissions between each placebo unit and its synthetic counterpart. Only placebos with high-quality pre-treatment counterfactuals are informative to evaluate whether the treatment effect of the CCP is robust to a falsification test. Thus, figure 2.3 only includes placebos whose pre-treatment MSPE is not more than 30 times greater that of the UK's. The causal effect of the CCP in the UK lies at the edge of this null distribution. In other words, we would be unlikely to see a treatment effect as large as we see for the UK by chance alone.

Since we know that none of the placebo countries had a climate policy, we should expect null treatment effects on each of these placebo treatments, as only the United Kingdom was treated with the Climate Change Programme in 2001. The donor pool includes countries that were Annex I parties to the United Nations Framework Convention on Climate Change (UNFCCC) in 1992. To the extent that Annex I membership might constitute a shadow treatment on these countries, this will bias against finding an effect; and our estimates can thus be seen as a lower-bound on the treatment effect of the CCP. After iteratively assigning a placebo treatment to countries in the donor pool, we then calculate the gaps in emissions between the placebo units and their synthetic controls. We should expect to see little to no variation in these post-2001, other than by chance.

It may be the case that the synthetic control algorithm on a placebo unit failed to achieve a good pre-treatment fit, in which case this placebo run would be uninformative. We account for this by calculating the mean squared prediction error (MSPE), which is the average of the squared gaps between the per capita CO_2 emissions in the treated unit and its synthetic control. If the fit achieved by the synthetic control algorithm was good, then we should expect a low pre-treatment MSPE; and conversely, if the fit was poor, the pre-treatment MSPE for any given country would be larger. If a country (placebo or the UK) has a large MSPE post-treatment, this is suggestive of a large treatment effect. We compute the ratio of the post- to pre-treatment MSPE for the United Kingdom and each placebo country in the sample, as recommended by Abadie et al. (2010, 2011, 2015). By dividing the post-treatment gaps with the pre-treatment gaps, the statistic downweights the ill-fitting synthetic controls. This effectively penalizes the treatment effect when the fit achieved by the synthetic control algorithm was poor. The ratio of post- to pre-treatment MSPE for all countries in the donor pool is the statistic that we use to create a non-parametric null distribution.

We can then look at the empirical distribution of this statistic to ascertain whether the ratio of post- to pre-treatment MSPE in the UK falls in the tails of this distribution, which would indicate that the results in the UK are unlikely to be driven by chance. When we re-assign treatment to all countries in the sample, we find that the UK has the largest ratio statistic. If we were to pick a country at random under uniform sampling from the entire sample, the probability of obtaining a ratio statistic as large as the UK's is $1/51 \approx 0.02$. In other word, the probability of obtaining a treatment effect as large as the UK's would be 0.02, which is conventionally seen as statistically significant for parametric analyses. Figure 2.4 displays the empirical distribution of this ratio statistic: this is our null distribution. The UK's ratio statistic is approximately 3687, and it falls in the right tail of that distribution, which suggests that we can reject the null hypothesis that the CCP had no effect in favor of the alternative hypothesis that the CCP had an effect on emissions per capita.

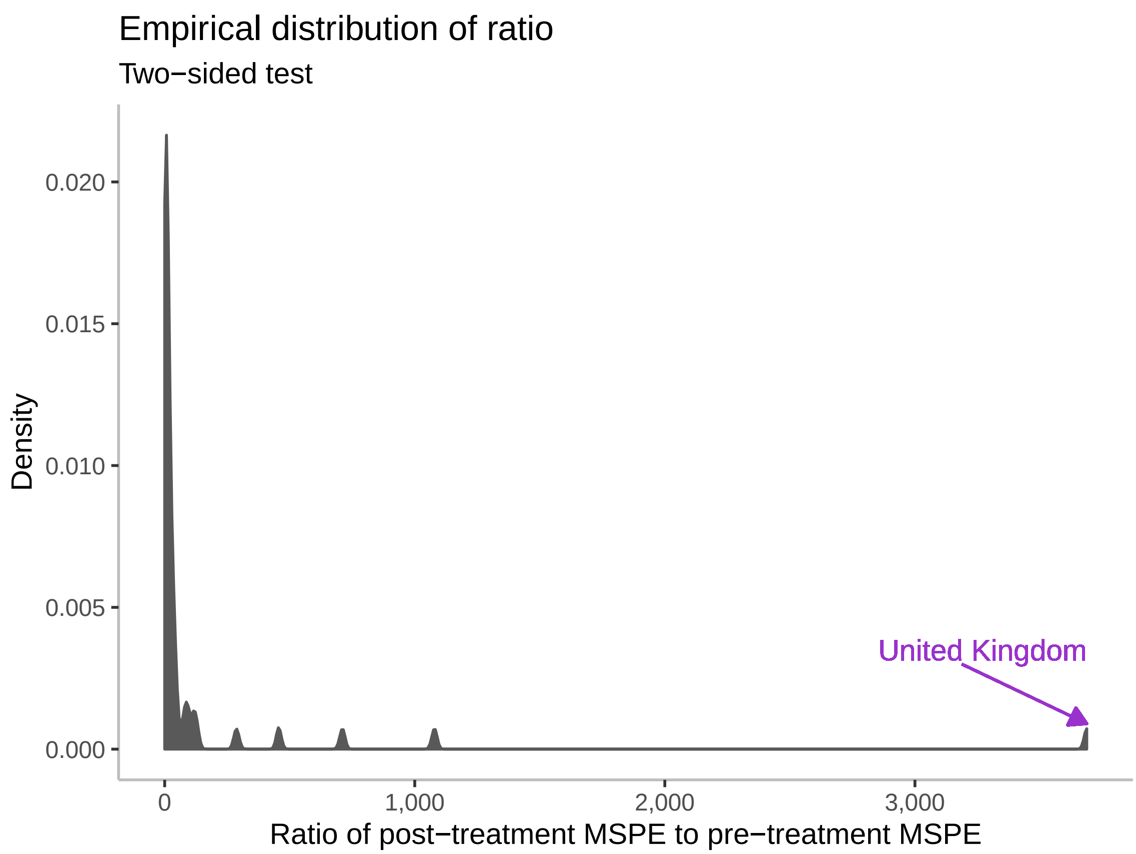


Figure 2.4: Null distribution for a two-sided test. The density represents the empirical distribution of the ratio statistic (computed as the ratio of post- to pre-treatment Mean Square Prediction Error) for all countries in the sample.

2.3.3 Robustness checks

Finally, we conduct additional checks to verify that are results are robust. These include "leave-one-out" robustness checks where we iteratively drop a single country from the donor pool to ensure that our results are not an artifact of individual donor countries, placebo "in time" tests where we re-assign treatment to earlier years, and a series of alternative specifications for synthetic control construction.

First, we might ask whether the weights in the synthetic UK are driven by certain countries in particular. To test this, we conduct a "leave-one-out" robustness check where we iteratively drop a single country at a time from the donor pool used to construct the synthetic UK. This allows us to check that the emissions trajectory of the synthetic UK is not driven by a single country, and that achieving balance between the pre-CCP emissions trajectories of the UK and its synthetic control does not depend on the inclusion of a single country. As shown by figure 2.5, our results remain robust to the omission of single countries from the donor pool.

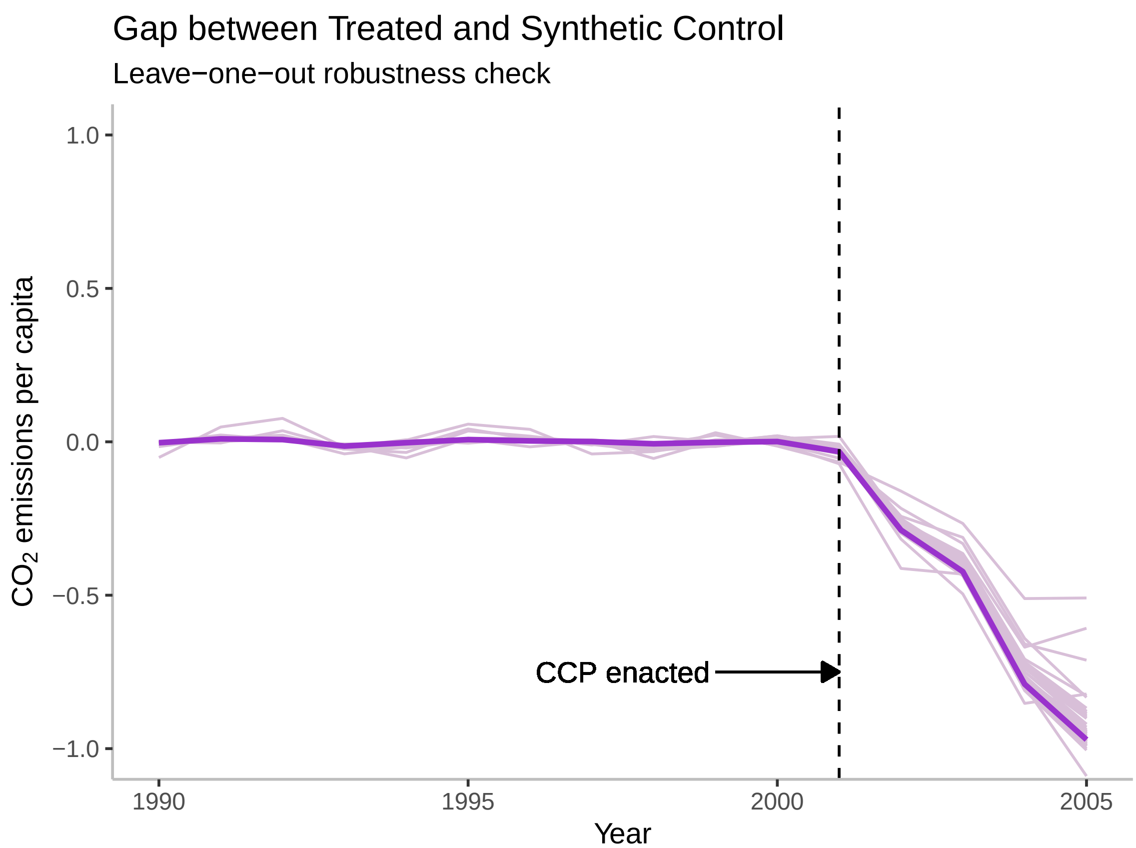


Figure 2.5: Gaps in per capita emissions between the UK and the synthetic UK. The thick purple line represents the gaps when the synthetic UK is constructed using all countries in the donor pool (51 countries). Each thin purple line represents the gaps when one country is dropped from the donor pool.

Second, we run placebo "in time" tests, where re-assign treatment to previous years.

Since we know that treatment occurred in 2001, and not earlier, we should not expect to find a large divergence between the UK and its synthetic control in those placebo years, other than by chance. Figure 2.6 below displays the results of this test for the year immediately preceding the passage of the CCP. The emissions trajectory of the synthetic control for the placebo year 2000 do not start diverging from those of the UK until after 2001 and not earlier, which further reinforces the impression that there indeed was a structural break in emissions after the treatment. Additional placebo tests for other years prior to treatment can be found in appendix section A.5.

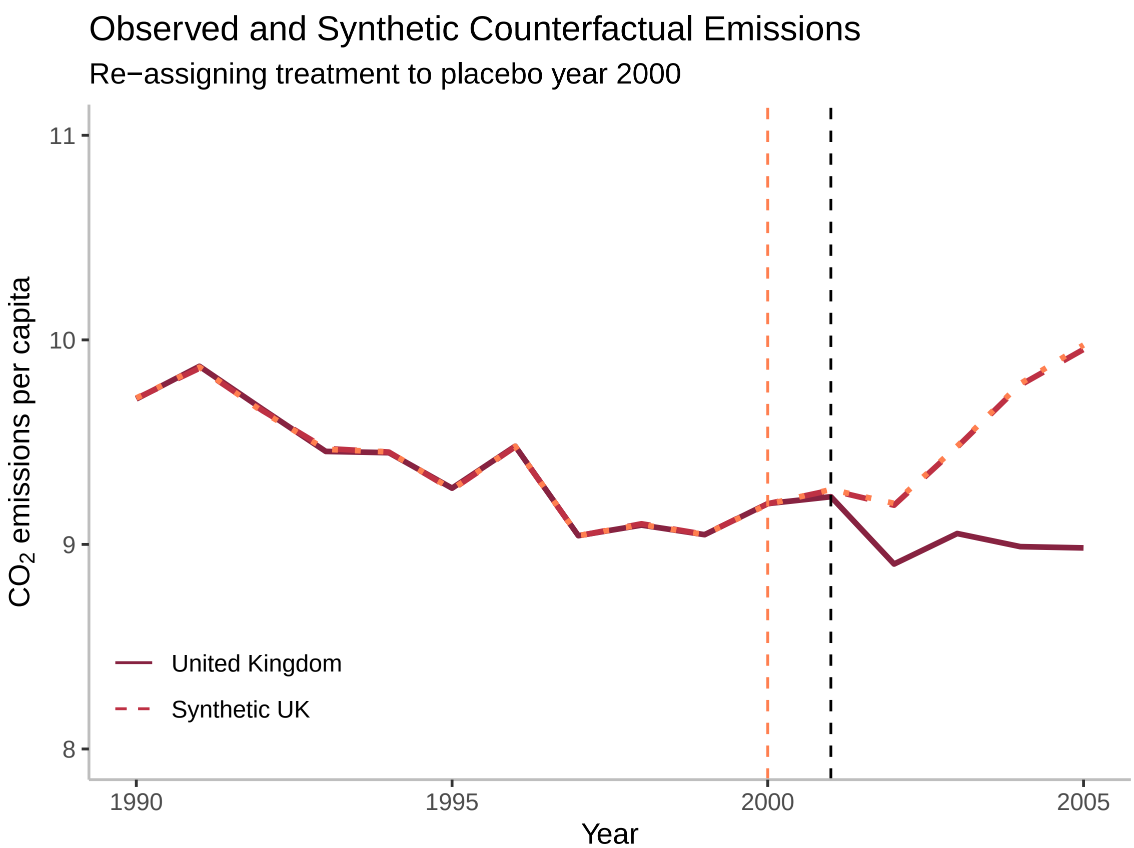


Figure 2.6: Observed and synthetic counterfactual emissions for a placebo run where treatment occurs in 2000. The synthetic control's emissions trajectory for the placebo year 2000 are in the dashed orange line.

Third, we run the synthetic control procedure on a variety of alternative specifications and samples. We considered two alternative ways to operationalize the outcome variable: CO_2 emissions rescaled to a 1990 baseline, and CO_2 emissions rescaled to a 2000 baseline. Both dependent variables are rescaled to ensure that they are within the common support of the data. The first outcome variable is rescaled to the baseline used in the formulation of the Kyoto targets, and can help us visualize at a glance the extent to which the UK met its targets. The second dependent variable can then help us understand the immediate impact of the CCP at $t + 1$.

We also consider three samples for the donor pool: (A) countries that were either OECD, high or upper middle income countries in 2001; (B) countries that were either OECD or high income countries in 2001; and (C) countries that were OECD members in 2001. In all of these samples we exclude Northern European countries that we consider to have been treated by 2001.

We report the results run on donor pool (A) as our preferred model, but the results run on donor pools (B) and (C) are also statistically significant. However, the smaller donor pool sample means that achieving a good pre-treatment fit between the UK and its synthetic counterpart is dependent on the inclusion of a single country, Luxembourg. This is not a problem when we use the larger donor pool (A): if we drop Luxembourg, the treatment effect is comparable (-8.5%) and is statistically significant ($p = 0.02$).

Finally, we also run the synthetic control method on the main donor pool sample (A) using a specification that includes covariates (specification 2), and one that increases the pre-treatment optimization period to 1980 (specification 3). The treatment effect of the CCP is substantively large and statistically significant in both of those cases too.

Table 2.1 below summarizes all the specifications that have been run as a robustness check on our results. The detailed results for our alternative specifications can be found in section A.7 of the appendix. As our main finding, we choose to report a specification where the outcome variable is CO_2 emissions per capita, rather than emissions rescaled to a baseline, since per capita emissions are a meaningful and readily interpretable measure of climate abatement. Within the specifications that have per capita CO_2 emissions as their outcome variable (specifications 1-5), we choose the specification that achieves the best pre-treatment fit (i.e. the lowest pre-treatment MSPE), which occurs when the donor pool comprises countries that were either OECD, high or upper middle income countries in 2001.

Best practice in SCM analysis is to report several specifications as a robustness check (Ferman et al., 2017, 2020). Ferman et al. (2020) discuss how to approach generating a valid hypothesis test that encompasses all the different specifications. On the one hand, a decision rule that rejects the null hypothesis of no effect only if all the specifications individually reject the null would be unduly conservative; though it should be noted that our results would pass that test (at a 10% significance level). On the other hand, a decision rule that rejects the null if at least one specification has rejected the null would inflate the rate of false positives. They thus suggest to generate a new test statistic, inspired by work by Imbens and Rubin (2015): for each unit j and across all specifications s, compute the ratio of post- to pre-treatment MSPE, and compute p-values using the same statistical inference procedure as before.

We compute such a test statistic across specifications that share the same donor pool. For all 3 donor pools, those omnibus p-values are highly statistical significant – pool (A): $p = 0.0385$; pool (B): $p = 0.0303$; pool (C): $p = 0.0435$. This indicates that our findings are not the result of a single spurious specification; we can thus reasonably conclude that the CCP had a significant and negative effect on British per capita CO_2 emissions.

Spec	Outcome variable	Donor pool	Obs	Covariates	Optimization period	Pre-treatment MSPE	p-value
1	CO_2 per capita	(A) OECD, high & upper middle income	$n = 51$	No	1990-2001	1.24×10^{-4}	0.020
2	CO_2 per capita	(A) OECD, high & upper middle income	$n = 37$	Yes	1990-2001	2.07×10^{-3}	0.053
3	CO_2 per capita	(A) OECD, high & upper middle income	$n = 51$	No	1980-2001	7.86×10^{-3}	0.058
4	CO_2 per capita	(B) OECD & high income	$n = 32$	No	1990-2001	5.24×10^{-4}	0.030
5	CO_2 per capita	(C) OECD	$n = 22$	No	1990-2001	2.13×10^{-3}	0.043
6	1990 baseline	(A) OECD, high & upper middle income	$n = 51$	No	1990-2001	6.42×10^{-6}	0.038
7	1990 baseline	(B) OECD & high income	$n = 32$	No	1990-2001	1.07×10^{-5}	0.030
8	1990 baseline	(C) OECD	$n = 22$	No	1990-2001	3.12×10^{-5}	0.043
9	2000 baseline	(A) OECD, high & upper middle income	$n = 51$	No	1990-2001	8.69×10^{-6}	0.038
10	2000 baseline	(B) OECD & high income	$n = 32$	No	1990-2001	1.42×10^{-5}	0.030
11	2000 baseline	(C) OECD	$n = 22$	No	1990-2001	3.75×10^{-5}	0.043

Table 2.1: Summary of alternative specifications. Specification 1 is reported as the main finding. Details on alternative specifications including robustness checks are reported in the appendix section A.7. The p-value reported here is two-sided.

2.4 Discussion

Collectively, national climate policies remain insufficient to mitigate the catastrophic risks of climate change (Peters et al., 2015). However, we show that a unilateral climate policy in the United Kingdom meaningfully reduced carbon pollution, even in the absence of a legally binding global climate treaty. Conventional accounts of global climate policy-making emphasize countries' weak incentives to act on climate change alone. Yet, we show that the United Kingdom reduced its per capita carbon pollution by 9.8% in the face of free-riding disincentives to act.

The CCP included a mix of several policy instruments: a type of carbon tax (the Climate Change Levy collected from industry and the public sector), negotiated industry agreements (the so-called Climate Change Agreements), and a domestic emissions trading scheme (ETS). These policies individually had several shortcomings which cast doubt on the CCP's ability to achieve substantial emissions reductions. In particular, empirical evidence suggests that the Climate Change Agreement (CCA) targets negotiated with industry were too lax at the outset (Ekins and Etheridge, 2006), which would have resulted in "hot air" on the emissions trading scheme (ETS) market. The CCL was not a pure carbon tax and carbon-free nuclear energy was not exempt from it. The Climate Change Agreements were negotiated with industrial polluters and made substantial concessions to producers. Sectors who overcomplied on their CCA targets could sell those surplus emissions as allowances on the UK's domestic ETS, and conversely sectors could meet their CCA targets by purchasing permits on the market. These provisions introduced additional flexibility for business managers who could decide on the least-cost way to meet their CCA targets.

While the CCA targets themselves were lax, the CCA sectors outperformed their 2002 targets. Ekins and Etheridge (2006) argue that this was due to an “awareness effect”: there were many cost-effective opportunities to improve energy efficiency that had previously not been recognized by industrial business managers. The excise rates of the CCL were high enough to be considered a credible threat and succeeded in bringing industrial actors to the table to negotiate the voluntary CCA targets, and it was this process which allowed the private sector to realize that there were low-hanging fruit energy efficiency gains to be made (Ekins and Etheridge, 2006). Many of those energy improvements were made on financial grounds alone, and the fact that the targets were not stringent was counterbalanced by the process of learning from industrial managers about how energy efficiency could improve their bottom line. These findings provide suggestive evidence that a combination of imperfect policy instruments can result in meaningful emissions mitigation.

Appendix section A.8 provides an additional narrative of the mobilization against the CCP by both labor and industry groups which succeeded in watering down the stringency of the policy and resulted in important concessions to polluters. Still, despite regulatory capture by industry, and even if it was voluntary and unilateral, the CCP was nevertheless able to abate 148 Mt of CO_2 over 4 years, or around 37 Mt of CO_2 per annum. The IPCC estimates that mitigation pathways that keep warming within 1.5°C would cap emissions in 2030 to 25-30 Gt CO_2e per year (Rogelj et al., 2018). Our results suggest that the UK was able to mitigate emissions on the order of magnitude of 0.5% of the *global* annual carbon budget remaining in 2030.

Finally, even though evaluating the overall impact of a given climate policy on national-level carbon emissions is crucial for the development of climate budgets, existing efforts are stifled by the reliance on unrealistic BAU scenarios. BAU scenarios used for causal impact evaluations need to be developed with the explicit aim of being counterfactual.

CGE and IAM models are useful for *ex ante* simulations of the general equilibrium effects of an exogenous policy on the economy and on the environment. However, the BAU scenarios that are used by these models as comparisons are not necessarily appropriate for an *ex post* policy impact evaluation. This is because the correct counterfactual to estimate the impact of a climate policy is a scenario where the policy had *not* been passed, and not a baseline of no action or other stylized vignette. However, it is difficult to enumerate all the possible drivers of that counterfactual emissions trajectory, and furthermore to specify how they interact with each other. We demonstrate the advantage of using a non-parametric approach which obviates the need to specify a functional form for all of the (observed and unobserved) drivers of emissions. The synthetic control estimator captures the specific combination of underlying dynamic and static structural drivers of British emissions in the control units and reweights them accordingly to create a credible synthetic control.

Alongside parallel work by Bayer and Aklin (2020), our findings show the promise of synthetic control methods as a tool for *ex post* climate policy impact analysis that can provide net national estimates of CO_2 abatement without relying on simplistic forward projections of emissions. More accurate climate policy evaluations can in turn inform the analysis of national and global carbon budgets, which form the basis of actionable goals for climate stabilization.

Bibliography

Abadie, A., Diamond, A., and Hainmueller, J. (2010). Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. *Journal of the American Statistical Association*, 105(490):493–505.

Abadie, A., Diamond, A., and Hainmueller, J. (2011). Synth: An R Package for Synthetic Control Methods in Comparative Case Studies. *Journal of Statistical Software*, 42(13):1–17.

Abadie, A., Diamond, A., and Hainmueller, J. (2015). Comparative Politics and the Synthetic Control Method. *American Journal of Political Science*, 59(2):495–510.

Abadie, A. and Gardeazabal, J. (2003). The Economic Costs of Conflict: A Case Study of the Basque Country. *American Economic Review*, 93(1):113–132.

Agnolucci, P. (2009). The effect of the German and British environmental taxation reforms: A simple assessment. *Energy Policy*, 37(8):3043–3051.

Allen, M. R., Frame, D. J., Huntingford, C., Jones, C. D., Lowe, J. A., Meinshausen, M., and Meinshausen, N. (2009). Warming caused by cumulative carbon emissions towards the trillionth tonne. *Nature*, 458:1163–1166.

Ang, C. P., Toper, B., and Gambhir, A. (2016). Financial impacts of UK's energy and climate change policies on commercial and industrial businesses. *Energy Policy*, 91:273–286.

Athey, S. and Imbens, G. W. (2017). The State of Applied Econometrics: Causality and Policy Evaluation. *Journal of Economic Perspectives*, 31(2):3–32.

Bailey, I. and Rupp, S. (2005). Geography and climate policy: a comparative assessment of new environmental policy instruments in the UK and Germany. *Geoforum*, 36(3):387–401.

Bayer, P. and Aklin, M. (2020). The European Union Emissions Trading System reduced CO_2 emissions despite low prices. *Proceedings of the National Academy of Sciences*, 117(16):8804–8812.

Bertrand, M., Duflo, E., and Mullainathan, S. (2004). How Much Should We Trust Differences-in-Differences Estimates? *The Quarterly Journal of Economics*, 119(1):249–275.

Billmeier, A. and Nannicini, T. (2013). Assessing Economic Liberalization Episodes: A Synthetic Control Approach. *The Review of Economics and Statistics*, 95(3):983–1001.

Böhringer, C., Hoffman, T., Lange, A., Löschel, A., and Moslener, U. (2005). Assessing Emission Regulation in Europe: An Interactive Simulation Approach. *The Energy Journal*, 26(4):1–21.

Böhringer, C., Rutherford, T. F., and Wiegard, W. (2003). Computable general equilibrium analysis: Opening a black box. *ZEW Discussion Papers*, 03-56:1–34.

Bruvoll, A. and Larsen, B. M. (2004). Greenhouse gas emissions in Norway: do carbon taxes work? *Energy Policy*, 32(4):493–505.

Burniaux, J.-M., Nicoletti, G., and Oliveira-Martins, J. (1992). GREEN: a global model for quantifying costs of policies to curb CO_2 emissions. *OECD Economic Studies*, 19:49–92.

Costalli, S., Moretti, L., and Pischedda, C. (2017). The economic costs of civil war: Synthetic counterfactual evidence and the effects of ethnic fractionalization. *Journal of Peace Research*, 54(1):80–98.

Downs, G. W., Rocke, D. M., and Barsoom, P. N. (1996). Is the good news about compliance good news about cooperation? *International Organization*, 50(3):379–406.

Ekins, P. and Etheridge, B. (2006). The environmental and economic impacts of the UK climate change agreements. *Energy Policy*, 34(15):2071–2086.

Ferman, B., Pinto, C., and Possebom, V. (2017). Cherry Picking with Synthetic Controls. *Munich Personal RePEc Archive*, 78213:1–55.

Ferman, B., Pinto, C., and Possebom, V. (2020). Cherry Picking with Synthetic Controls. *Journal of Policy Analysis and Management*, 39(2):510–532.

Future Energy Solutions (2003). *Climate change agreements: Results of the First Target Period Assessment (version 1.2, revised July 2004)*. AEA Technology, Harwell, UK.

Hazlett, C. and Xu, Y. (2018). Trajectory Balancing: A General Reweighting Approach to Causal Inference With Time-Series Cross-Sectional Data. *SSRN Electronic Journal*, pages 2–38.

Heersink, B. and Peterson, B. D. (2016). Measuring the Vice-Presidential Home State Advantage With Synthetic Controls. *American Politics Research*, 44(4):734–763.

Hoerner, J. A. and Bosquet, B. (2001). *Environmental Tax Reform: The European Experience*. Center for a Sustainable Economy, Washington, DC.

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, 81(396):945–960.

Hu, J., Crijns-Graus, W., Lam, L., and Gilbert, A. (2015). Ex-ante evaluation of EU ETS during 2013-2030: EU-internal abatement. *Energy Policy*, 77:152–163.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference for Statistics, Social and Biomedical Sciences: An Introduction*. Cambridge University Press, Cambridge, UK.

Kaul, A., Klößner, S., Pfeifer, G., and Schieler, M. (2018). Synthetic Control Methods: Never Use All Pre-Intervention Outcomes Together With Covariates.

Martin, R., de Preux, L. B., and Wagner, U. J. (2009). The impacts of the Climate Change Levy on business: evidence from microdata. *Centre for Climate Change Economics and Policy Working Paper*, 7:1–41.

Nordhaus, W. (2013). Integrated Economic and Climate Modeling. In Dixon, P. B. and Jorgenson, D. W., editors, *Handbook of Computable General Equilibrium Modeling*, volume 1, chapter 16, pages 1069–1131. Elsevier B.V.

Peters, G. P., Andrew, R. M., Solomon, S., and Friedlingstein, P. (2015). Measuring a fair and ambitious climate agreement using cumulative emissions. *Environmental Research Letters*, 10(105004):1–9.

Peters, G. P., Marland, G., Le Quéré, C., Boden, T., Canadell, J. G., and Raupach, M. R. (2012). Rapid growth in CO_2 emissions after the 2008–2009 global financial crisis. *Nature Climate Change*, 2(1):2–4.

Pindyck, R. S. (2017). The Use and Misuse of Models for Climate Policy. *Review of Environmental Economics and Policy*, 11(1):100–114.

Robbins, M. W., Saunders, J., and Kilmer, B. (2017). A Framework for Synthetic Control Methods With High-Dimensional, Micro-Level Data: Evaluating a Neighborhood-Specific Crime Intervention. *Journal of the American Statistical Association*, 112(517):109–126.

Rogelj, J., Shindell, D., Jiang, K., Fifita, S., Forster, P., Ginzburg, V., Handa, C., Kheshgi, H., Kobayashi, S., Kriegler, E., Mundaca, L., Séférian, R., and Vilariño, M. (2018). Mitigation Pathways Compatible with 1.5°C in the Context of Sustainable Development. In Masson-Delmotte, V., Zhai, P., Pörtner, H.-O., Roberts, D., Skea, J., Shukla, P., Pirani, A., Moufouma-Okia, W., Péan, C., Pidcock, R., Connors, S., Matthews, J., Chen, Y., Zhou, X., Gomis, M., Lonnoy, E., Maycock, T., Tignor, M., and Waterfield, T., editors, *Global Warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty*, chapter 2, pages 93–174.

Rubin, D. B. (1974). Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies. *Journal of Educational Psychology*, 66(5):688–701.

Schleich, J., Eichhammer, W., Boede, U., Gagelmann, F., Jochem, E., Schlomann, B., and Ziesing, H. J. (2001). Greenhouse gas reductions in germany - lucky strike or hard work? *Climate Policy*, 1(3):363–380.

Sills, E. O., Herrera, D., Kirkpatrick, A. J., Brandão Jr., A., Dickson, R., Hall, S., Pattanayak, S., Shoch, D., Vedoveto, M., Young, L., and Pfaff, A. (2015). Estimating the Impacts of Local Policy Innovation: The Synthetic Control Method Applied to Tropical Deforestation. *PLoS ONE*, 10(7):1–15.

Svendsen, G. T., Daugbjerg, C., Hjollund, L., and Pedersen, A. B. (2001). Consumers, industrialists and the political economy of green taxation: CO_2 taxation in OECD. *Energy Policy*, 29(6):489–497.

World Bank (2019). *State and Trends of Carbon Pricing 2019*. World Bank Group, Washington, DC.

Xu, Y. (2017). Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models. *Political Analysis*, 25(1):57–76.

Chapter 3

Presenting a new atlas of illicit financial flows from trade misinvoicing

3.1 Introduction

Tackling illicit financial flows (IFFs) has become a key international policy priority in recent years. The fight against illicit finance has been the subject of international co-operation efforts at the United Nations, the OECD, and various intergovernmental fora. There is a general recognition that illicit financial flows erode the ability of governments to generate resources and directly undermine the efforts of the global community to successfully achieve the Sustainable Development Goals (SDGs). The United Nations Conference on Trade and Development (UNCTAD) estimates that there is an annual financing gap of $2.5 trillion for developing countries to achieve the Sustainable Development Goals (Doumbia and Lauridsen, 2019). Illicit financial flows create an uneven playing field both domestically by increasing wealth disparities and internationally by threatening the prospects of development for poor countries. Combating IFFs is of primordial concern to developing countries if they are to mobilize domestic resources to finance their own development.

The term illicit financial flows was first coined by Baker (2005) who defines IFFs as the movement of money across borders that is illegally earned, transferred, or utilized. At some point in the origin, destination, or movement of the money, laws were broken and hence the corresponding financial flow is considered illicit (Kar, 2010). Trade misinvoicing is the main source of illicit financial flows (see, e.g., Spanjers and Salomon (2017); Salomon (2019)) and existing estimates have suggested that developing countries lose hundreds of billions of dollars each year through trade misinvoicing (Spanjers and Salomon, 2017), while other literature suggests that such practices are a key weakness in the fight against corruption, transnational organized crime, and the financing of terror (Findley et al., 2020; UNODC, 2011; FATF, 2019).

Measuring and tracking illicit flows is extremely challenging, since by their very nature illicit flows are not systematically recorded. The difficulties of quantification are a significant hindrance to understanding the extent of the problem and where it is most severe. This paper contributes a novel methodology to estimate trade misinvoicing at scale and with sufficient resolution, and offers an "atlas of trade misinvoicing" that contains bilateral estimates of misinvoiced trade for 167 countries during 2000-2018 at a disaggregated sectoral level for all commodities reported to UN Comtrade.

Existing trade misinvoicing estimates have faced intense scrutiny about the robustness of their methodologies; with some authors considering that the methodological flaws of the estimates render them devoid of any substantive meaning (Nitsch, 2016) and some lamenting that the debate on the scale of illicit outflows might be a distraction from the more pressing underlying issues (see, e.g., Reuter (2012)). The methodology in this paper provides improvements that seek to address long-standing concerns in the literature on estimating trade misinvoicing. One of the main criticisms holds that the estimates use discrepancies in mirror trade statistics as a proxy for trade misinvoicing, but that there are many potential "non-illicit" sources for such discrepancies, which may be the cause of most of this apparent trade misinvoicing. The approach presented here allows for a systematic way of adjusting for all important sources of "non-illicit" discrepancies.

The paper proceeds as follows. In order to understand the added value of the methodology, section 3.2 first presents the main concepts behind trade manipulations and discusses how different channels of misinvoicing are harmful. Section 3.3 of the paper makes the general case for how and why the "atlas" method offers improvements to mitigate the problems of existing methodologies. Six criteria by which to judge whether an estimate is a credible measure of trade misinvoicing are advanced, that consider both methodological cogency and practicality (section 3.3.1). The most revelant existing methodologies in the literature are critically appraised in terms of how much they fulfill those criteria (section 3.3.2). Then, the main innovations of the "atlas" methodology are presented in order to demonstrate that the measure exhibits all the characteristics of a good measure of trade misinvoicing (section 3.3.3).

Step-by-step details of the methodology are provided in section 3.4, in addition to further discussions of the assumptions and methodological choices involved. Section 3.5 presents the main findings and provides a practical application of how the "atlas" can be used to zoom in to different views of the problem. Section 3.6 discusses the limitations of the approach and section 3.7 concludes.

3.2 Channels of trade misinvoicing

Trade misinvoicing is the deliberate mis-statement of invoices presented to customs in order to clandestinely shift money abroad or repatriate money domestically. The stratagem is used for a variety of nefarious purposes including money laundering, tax evasion, and the financing of terrorism. Both imports and exports can be misinvoiced and can result in either an illicit outflow or an illicit inflow. The type of trade manipulation that is used depends on the underlying motives for concealing money transfers, and these in turn will harm the prospects for sustainable development and good governance in a variety of ways. In order to critically appraise a measure of trade misinvoicing, it is necessary to understand the directions of the illicit flows and how misinvoicing manifests in both import and export trade flows. This section presents the four main types of trade manipulations, explains how each channel is exploited for different purposes, and briefly discusses the development impacts of these manipulations.

Trade invoices can be faked by either the importer, the exporter, or both, which gives rise to four types of manipulations that are executed for varied reasons. 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 and Davis, 2020). When the value of imports is overstated, excess funds leave the country disguised as a form of trade payment (Schuster and 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 and Davis, 2020), which deprives countries of precious foreign exchange and erodes their tax base.

Import under-invoicing and export over-invoicing, on the other hand, will result in an inflow. The potential to evade tariffs by understating the value of imports has been pointed out since Bhagwati (1964). The conventional wisdom among economists, bolstered by empirical evidence (Sachs and Warner, 1995), is that tariffs usually depress economic growth. The existence of the World Trade Organization (WTO) is predicated on this view, and its stated mandate is to reduce tariffs and other barriers to trade. The WTO aims to prevent "beggar thy neighbor" policies where countries engage in zero-sum mercantilist policies which end up leaving every trading partner worse off. Nevertheless, tariffs can also be seen as protective instruments designed to shore up infant industries, promote import substitution industrialization, or even temper the unequal distribution of gains and losses resulting from trade liberalization (Chang, 2005; Rodrik, 2018). Therefore, tariffs are elements of both a country's trade policy and its foreign policy. Irrespective of their economic desirability, tariffs are tools at the disposal of a sovereign nation. Evading a tariff is illegal and thus weakens the rule of law. Moreover, even though tariff evasion will manifest as an inflow (i.e., import under-invoicing), it effectively robs governments of tax revenues.

In addition, misinvoicing occurs opportunistically to exploit subsidy regimes. Export over-invoicing is used to take advantage of incentives that the government puts in place to encourage exports, such as subsidies or tax credits (Gara et al., 2019). As part of their overall economic strategy, countries sometimes seek to subsidize certain industries. Industries can be subsidized in order to champion certain strategic sectors that are in the national interest, in order to sustain a long-run comparative advantage in international trade, or even to guide a national transition towards a different sectoral make-up of the economy. By opportunistically over-stating the true value of their goods, misinvoicers can take advantage of such subsidy regimes in order to capture rents (Baker et al., 2014). Similarly, taking advantage of export subsidy regimes will look like an inflow, but it is a form of market abuse that can make it more difficult for the state to finance other socially beneficial activities.

More generally, trade misinvoicing is used to hide transfers of capital. Motivations for disguising transfers of capital range from financing terrorism and laundering criminal proceeds to tax evasion by individuals and corporations. For example, organized crime syndicates may use trade misinvoicing to repatriate capital and incorporate the proceeds of crime into the domestic legal financial system (UNODC, 2011). Trade misinvoicing can also be used to conceal transfers of wealth that do not stem from criminal activity. For example, capital that is gainfully earned can be moved out of a country to low-tax jurisdictions in order to avoid tax, or to secrecy jurisdictions in order to escape the rules and regulations of the home country. Multinational corporations frequently use misinvoicing to reduce their domestic tax burden by shifting their profits to a lower-tax jurisdiction (Leblanc, ay 9; ECLAC, 2016; Vicard, 2015). Widespread tax avoidance by multinational corporations impacts developing countries more severely than developed countries (UNECA, 2019).

Trade misinvoicing impedes the prospects for sustainable development in developing countries. Illicit financial outflows through trade misinvoicing reduce the level of aggregate demand and result in a reduction of economic output (at least in the short term) (UNCTAD, 2016). Even if the funds end up being "round-tripped" to the country from which they departed, less will return than originally left, due to the portion of the funds that will inevitably be paid to various enablers through the process of round-tripping (UNECA, 2018a). This may be particularly damaging where natural resources owned by the state are being exported: the amount of under-invoicing in such cases represents direct diversion of wealth from the national treasury to whoever collects the benefits of the under-invoicing on the other end of the transaction (UNCTAD, 2016). In addition, by circumventing foreign exchange controls, trade misinvoicing may also undermine national strategies for managing the exchange rate, potentially causing the price of imports to rise or (conversely) lowering export competitiveness, which may have negative consequences depending on the circumstances of the affected country (e.g., Griffiths (2003)).

Trade misinvoicing reduces tax revenues and erodes the tax base (Kar, 2010; Jha and Truong, 2015), which undermines public spending and governance, in turn slowing economic growth and worsening poverty (Ibis Ghana and Africa Centre for Energy Policy, 2015; ACTSA, 2019; Baker et al., 2014; Moore, 2007). While the loss of capital is the most immediate consequence of illicit outflows, the indirect consequences of trade misinvoicing are the erosion of governance and weakening of state institutions. Illicit inflows are detrimental to development since they are untaxed and invisible to governments. 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 and Janský, 2020). Therefore, illicit inflows from trade misinvoicing have the potential to be just as corrosive to good governance and state institutions as illicit outflows (Blankenburg and Khan, 2012; Spanjers and Salomon, 2017; Salomon, 2019)

A (perhaps less obvious) impact of trade misinvoicing is on the quality of official statistics. Misinvoicing leads to incorrect recording of the market value of goods and services being traded, which may mislead countries as to the relative value or potential of different industries (ESCWA, 2018), leading to poorer economic policy-making (Jerven, 2013).

Therefore, preventing illicit financial flows from trade misinvoicing is an urgent policy priority, and difficulties in quantifying the phenomenon have slowed progress. This paper contributes a novel approach to estimating trade misinvoicing and offers an "atlas of misinvoicing" – a comprehensive collection of bilateral estimates for country pairs. Understanding the four types of trade manipulations presented above is a prerequisite for generating bilateral estimates. At this stage, it is necessary to distinguish between "reporter" and "partner" countries. Following the practice of "double entry accounting" in the compilation of international trade statistics, every trade transaction is reported twice to the United Nations Commodities Trade (Comtrade) database. A given country i (the "reporter") will report the value of its imports from a foreign country j (the "partner"), and that foreign country will in turn report the value of its exports to i. The exports reported by i's partner j are the "mirror exports". Likewise, country i will also report the value of its exports to its partner j to Comtrade, while the partner j will declare the corresponding "mirror imports" to Comtrade.

The "atlas of misinvoicing" approach always proceeds from the perspective of the reporter i (whether the trade flow reported to Comtrade is imports or exports): trade misinvoicing is estimated both in import and export transactions for *reporters*. In other words, this paper estimates the misinvoicing that is present in the import and export invoices that are presented at country i's customs, not j's. In turn, estimating the misinvoicing for all countries i in the set of reporters will yield the misinvoicing for partners too (since a partner j also reports to Comtrade). More specifically, reporters are the set of countries $\{i, \ldots, n\} \in \mathcal{I}$ that report to Comtrade a trade transaction with a partner j. Since not every country i trades with every other country in the world, the set of possible partner countries is a subset of the reporter set: $\{j, \ldots, k\} \in \mathcal{J} \subset \mathcal{I}$ with $k \leq n$. Therefore, to calculate illicit trade for every country that reports data to Comtrade, the methodology proceeds from the perspective of the reporting country i. Hence, the reporter i is the proverbial "atlas" (the topmost vertebra which supports the backbone) from whose vantage point trade misinvoicing is estimated.

As explained above, trade misinvoicing can result in an inflow or an outflow, and this can be achieved by misreporting the value of imports and/or exports. Figure 3.1 represents the direction of illicit flows from the perspective of the reporting country i and the associated mechanisms. Money can be moved out of country i by over-invoicing imports, where country i pays too much money to buy goods from its partner j; or by under-invoicing exports, where country i does not charge enough money for the goods that it sells to its partner j. Conversely, money can be illicitly routed from country j into country i by under-invoicing imports, where i pays too little money to buy goods from its partner j; or by over-invoicing exports, where i charges too much for the goods that it sells to its partner j. The direction of illicit flows from the perspective of the reporting country and the associated mechanisms is represented in the stylized figure 3.1 below.

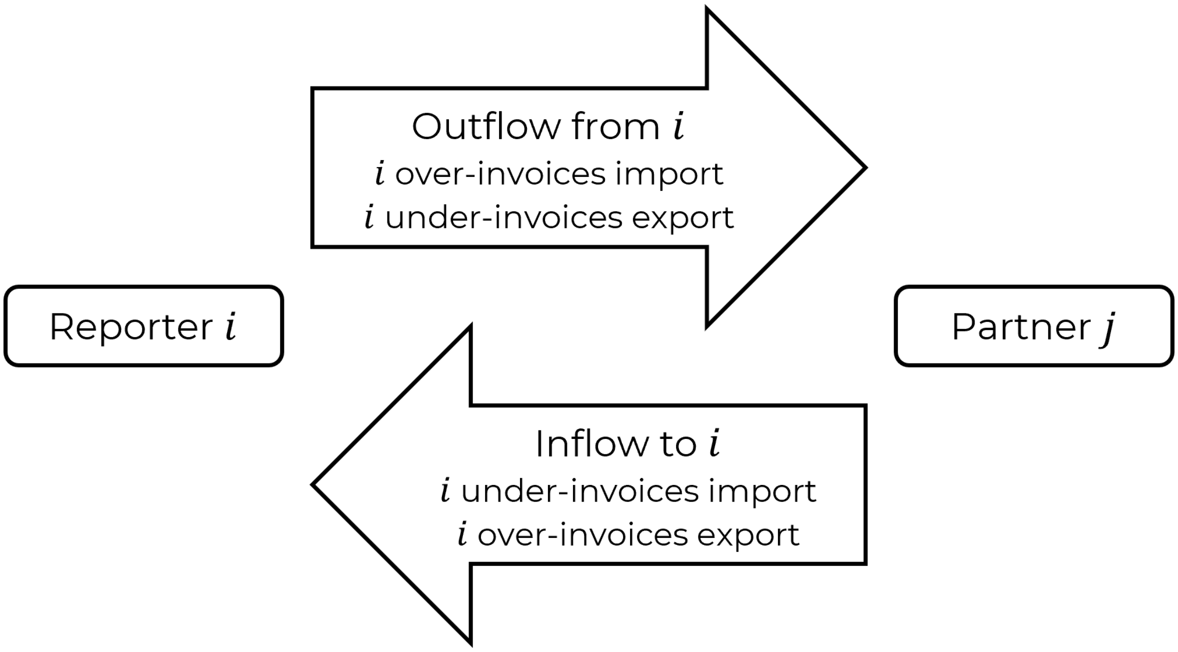


Figure 3.1: Mechanisms of trade misinvoicing from the perspective of the reporter.

With the requisite preliminaries out of the way, section 3.3 now sets out to explain how to measure trade misinvoicing by proposing a set of desirable features that a measure should possess in order to convincingly address the key methodological critiques in the literature.

3.3 Measuring trade misinvoicing

Policy-relevant estimates of illicit financial flows should serve two main functions: first, to highlight the extent of the problem so that countries can decide to what extent to prioritize policy action; and second, to indicate where the problem is worse and where attention should be focused to counter it, by indicating the main channels through which illicit finance is routed, and the main destinations at which it arrives.

The credibility of existing estimates of trade misinvoicing has been hotly contested, and the usefulness of existing estimates in informing policy interventions against IFFs is the subject of ongoing debates (see, e.g., Nitsch (2012, 2016); Cobham and Janský (2020); Picard (2003)). Some authors have highlighted their value in drawing attention to the scale of the problem and galvanizing much needed policy action to combat trade misinvoicing (UNECA, 2018a; Spanjers and Salomon, 2017; Salomon, 2019), while others have dismissed trade misinvoicing as an irrelevant sideshow whose importance has been vastly overstated as a result of the poor methodologies used in attempts to estimate it (Nitsch, 2016; Forstater, 2016). For this reason, developing a robust measure of trade misinvoicing is not only important to advance scholarship on IFFs, but it is also an urgent policy priority in order to justify reforms.

The definitional and methodological debates that have raged in the literature on IFFs are reflected politically by the lack of agreement by the United Nations member states on a comprehensive measure of illicit financial flows. Though the global community has recognized the importance of combating illicit finance by enshrining it as a Sustainable Development Goal (SDG), there is no concensus on how to evaluate progress towards that goal. Goal 16.4 of the SDGs aims to "by 2030, significantly reduce illicit financial flows and arms flows, strengthen the recovery and return of stolen assets and combat all forms of organized crime" (UN General Assembly, 2015) without specifying what constitutes a reduction of IFFs, let alone what the baseline measure is.

The development of common frameworks and indicators for measuring progress towards the SDGs has been the subject of international cooperation at the highest political levels. The consortium of governments and intergovernmental organizations tasked with developing a statistical framework for the SDGs initially ranked the indicator of IFFs at the lowest possible level, meaning that there was no internationally established methodology or standard for measurement, compared to other SDGs that have well-defined indicators for measuring progress.[1] Policy work is ongoing to clarify the subcategories of IFFs that will be included in the indicator and how they can be measured at a disaggregated level (see UNODC and UNCTAD (2020)). Therefore, the development and creation of policy-relevant proxy indicators of IFFs by researchers is a timely and valuable endeavor.

This paper contributes a novel indicator of trade misinvoicing (a subset of the IFF target) that offers broad country coverage *and* disaggregated estimates at the same time. To my knowledge, there are no existing estimates of misinvoicing that do so at a global scale. The "atlas of trade misinvoicing" provides measures of illicit trade for 167 countries during 2000-2018 for all commodities reported in Comtrade, disaggregated by 99 commodity sectors. This measurement of trade misinvoicing has already been used by international organizations,[2] notably to motivate a pilot initiative to strengthen customs system in selected African countries,[3] which bolsters the value of this database for providing tailored

[1] In 2015 the United Nations Statistical Commission created the Inter-agency and Expert Group on SDG Indicators (IAEG-SDGs) and tasked it with developing and implementing the global indicator framework for the targets of the 2030 Agenda. The IAEG-SDGs is composed of UN Member States and includes regional and international agencies as observers. The conceptual statistical framework for illicit financial flows measurement was initially classified under tier 3 of the SDGs global indicators. It was only in October 2019 that the IAEG-SDGs endorsed a reclassification of the indicator to tier 2, meaning that the indicator is conceptually clear, has an internationally established methodology and standards are available, but data are not regularly produced by countries.

[2] Results based on an earlier iteration of this methodology were used as the United Nations Economic Commission for Africa (UNECA) estimates of illicit financial flows from Africa. UNECA was represented in a series of expert meetings on statistical methodologies for measuring illicit financial flows, as part of the multilateral push to develop indicators for the IFF target of the SDGs. Further, results were also included in the *Financing for Sustainable Development Report 2019* of United Nations Inter-agency Task Force on Financing for Development (2019), available at `https://developmentfinance.un.org/sites/developmentfinance.un.org/files/FSDR2019.pdf`.

[3] UNECA works with African countries to scale up domestic resource mobilization and implement policy interventions against IFFs at the national government level. Early results of the "atlas" measure were used to inform a pilot project in six African countries that focused on building the capacity of national customs authorities and Financial Intelligence Units to detect and control trade misinvoicing. The measure was used to identify the key sectors, sources, and sinks of misinvoiced trade in Egypt, Nigeria, Senegal, South Africa, Tunisia and the United Republic of Tanzania, and to provide operational insights for these countries' customs administrations. For more information, see `https://repository.uneca.org/handle/10855/43054`.

intelligence to decision-makers working on IFFs.

In this section, I first develop the properties that a credible measure of trade misinvoicing should possess, then evaluate the extent to which current methodologies satisfy those criteria, and finally I show how the estimation strategy of the “atlas of misinvoicing” meets these criteria and mitigates long-standing problems in the literature and demonstrates both methodological rigor and applicability in practice.

3.3.1 Properties of a good measure of trade misinvoicing

There exist certain criteria that a measure of trade misinvoicing should meet in order to deliver estimates that are both theoretically cogent and practically meaningful. I submit the following set of six desirable properties for candidate measures of trade misinvoicing.

1. Avoid uncritically equating observed trade irregularities with misinvoicing
2. Partition the trade transaction into licit and illicit components in order to account for persistent non-illicit reasons for discrepancies
3. Account for the variance in countries’ statistical reporting
4. Scale across jurisdictions and over time
5. Provide enough granularity to support policy prioritization
6. Use open government data

The first three properties are concerned with the integrity of the methodological construct, while the final three characteristics are desirable in order to generate meaningful insights for researchers and practitioners.

Criterion 1 ***Avoid uncritically equating observed trade irregularities with misinvoicing.***

Irregularities in trade statistics do not necessarily imply foul play. Although irregularities might be indicative of misinvoicing in some cases, it would be incorrect to deduce that they are necessarily due to deliberate trade misinvoicing. Conversely, the *absence* of irregularities does not imply an absence of misinvoicing (World Customs Organization, 2018; Nitsch, 2012).

The following examples illustrate both types of logistal mistakes. There have been several cases of highly publicized estimates of lost revenues for African governments in the mineral sector that were later revealed to be "false positives", and as a result were publicly rebuffed and gave way to sweeping retractions. Instead of representing widespread theft of assets and rampant smuggling, the anomalies identified by these estimates could be attributed to readily explainable facts such as re-exporting and differences in reporting procedures. A prominent report by the United Nations Conference on Trade and Development (UNCTAD, 2016) suggested that up to 67% of gold exports from South Africa left the country unrecorded and that the country lost $78 billion dollars in IFFs during 2000-2014. The South African Revenue Service and the South African Chamber of Mines strongly objected to these findings, and argued that the mismatch between South Africa's records of gold exports and the import declarations of its trade partners was due to the peculiarities of South Africa's reporting practices, rather than egregious misappropriation of export revenues by mining companies.[4] In particular, South Africa

[4]See the press statement that was immediately issued by the South African Revenue Service (SARS) disputing the claims (South African Revenue Service, 2016), a report commissioned by the Chamber of Mines which lambasted the methodology of the UNCTAD report (Eunomix Research, 2017), and critical coverage in the South African media (Van Rensburg, st 1).

has a special trade regime for gold where (a) before 2011, gold exports were not recorded as a commodity to Comtrade but rather as a monetary flow in the IMF's Balance Of Payments, and (b) after 2011, even though gold exports were reported to Comtrade, they were not broken down by destination; both of which introduced spurious discrepancies in trade statistics (Schuster and Davis, 2020; Van Rensburg, st 1; Eunomix Research, 2017).

The other notable "false positive" case was that of Zambian copper and Switzerland. Zambia is a major copper producer and declares that more than 50% of its copper exports are destined for Switzerland (Schuster and Davis, 2020). By contrast, Switzerland reports no imports of copper from Zambia, but declares high export values of copper to third countries. The resulting trade gaps were used to make a – now retracted – claim that, if Zambia received the same export prices for copper as had been declared on Swiss exports, then Zambia's GDP in 2008 would have been 80 percent larger.[5] However, Switzerland is a major trading hub and the observed trade discrepancies are likely due to merchanting, whereby a Swiss company buys copper from a Zambian company, but stores the copper in bonded warehouses on the London Metal Exchange before reselling it to a final destination, without the copper ever entering Switzerland (Schuster and Davis, 2020). Therefore, usual practices in international commodity markets such as re-exporting can create illusions of IFFs due to asymmetric reporting.

While the two above examples are cautionary tales about the dangers of "false positives", there is also a risk of "false negatives". The absence of trade irregularities cannot be taken as evidence that there is no misinvoicing (Hong and Pak, 2017). One reason for this is if the importer and the exporter collude at both ends of the transaction to

[5] See the original claim by Cobham et al. (2014) and the subsequent retraction at https://cgdev.org/blog/how-much-are-developing-countries-losing-commodity-mispricing-really.

present inflated invoices to customs, a phenomenon called "same invoice faking", then the trade records will match even though they are falsified (Kar, 2010; World Customs Organization, 2018). Therefore, inferring that a particular transaction has not been misinvoiced from the absence of discrepancies in records is a logical fallacy that appeals to ignorance as the main premise for the argument. The silver lining is that, since strategies that exploit bilateral trade gaps to produce IFF estimates (an approach that this paper also adopts) cannot account for all instances of misinvoicing, they are conservative as a result. Therefore, estimates should be interpreted as a lower-bound of the true extent of the phenomenon.

In practice, however, all methods that estimate trade misinvoicing from reported data[6] exploit asymmetries and/or discrepancies in the data as an entry point to identifying illicit trade transactions. This leads to the second desirable property.

Criterion 2 ***Partition the trade transaction into licit and illicit components in order to account for persistent non-illicit reasons for discrepancies.***

In order to avoid equating all observed discrepancies with misinvoicing, it is necessary to account for persistent non-illicit reasons for discrepancies, such as honest reporting mistakes. In turn, this requires a strategy to plausibly partition a given trade transaction into its respective licit and illicit components.

There are legitimate reasons why imports and the corresponding mirror export values should differ. The most evident reason is that imports tend to be reported on a Cost

[6] As opposed to cases where misinvoicing is identified in a live setting during inspection of shipments by customs. Note that the measures that this paper is concerned with are not designed to be used for law enforcement purposes. Measures based on aggregate economic and financial data are used for *retrospective* studies rather than *prospective* applications.

of Insurance and Freight (CIF) basis, while exports tend to be recorded Free-On-Board (FOB), so reported import values are often inflated with transport and other transaction costs (World Customs Organization, 2018).

Other non-illicit reasons for discrepancies in records include: a delay between the recording of an import at time t and the recording of the corresponding export in the next time period $t+1$; asymmetric reporting of re-exports which will introduce artificial discrepancies in bilateral trade statistics; and idiosyncrasies in each country's quality of declaration.

Therefore, a good measure of misinvoicing should have a strategy to account for benign discrepancies in order to generate credible estimates of illicit trade.

Criterion 3 ***Account for the variance in countries' statistical reporting.***

The quality of official statistics varies with the level of economic development of countries (Jerven, 2013). The reliability of a country's declaration to UN Comtrade will be also be a function of its bureaucratic capacity and the robustness of its statistical reporting procedures (Devarajan, 2013; Jerven, 2009; Jerven and Johnston, 2015). The uncritical use of trade data should be avoided as estimates of misinvoicing might instead pick up statistical noise generated by shaky statistics rather than signals of deliberate trade falsification.

Likewise, though there are efforts at standardizing reporting practices, countries sometimes implement different rules for reporting, notably on rules of origin to determine the "economic nationality" of a tradeable product.

Yet, in order to create a measure of trade misinvoicing that is scaleable across jurisdictions, making manual adjustments to a country's reported trade in order to correct for declaration quality and country-specific idiosyncrasies is not practical. Therefore, a systematic approach to adjusting for the variance in bilateral trade declarations is needed.

Criterion 4 ***Scale across countries and over time.***

Relatedly, a desirable charasteric for a policy-relevant measure of trade misinvoicing is that it should scale across countries in order to provide the broadest country coverage possible. While micro-level measures can allow a customs official or auditor to conduct forensic investigations into whether a particular transaction is mispriced, the requisite particulars of the case will impede generalization. For example, cross-checking a trader's name from a blacklist of known financiers of terrorism can help in tracking and dismantling a particular plot, but it will not capture all other instances of misinvoicing. By contrast, macro-level measures of misinvoicing can help identify general trends and patterns and can provide analytical leverage to understand the dynamics of the phenomenon through time.

Note that the data requirements to provide a time series of estimates are particularly onerous and require that the trade statistics used as an input to the model are comparable through time. Moreover, estimating trade misinvoicing over time relies on the assumption that time-specific shocks do not affect IFFs; or at least on an empirical strategy to make this assumption plausible.

Criterion 5 ***Provide enough granularity to support policy prioritization.***

While a useful measure of trade misinvoicing will be scaleable across jurisdictions and over time, the possibility to zoom in with some degree of precision is also valuable. There is a trade-off between the coverage and the resolution of trade misinvoicing measures Cobham and Janský (2020). Measures that scale easily and have broad coverage (macro measures) will necessarily have lower resolution and offer less details on the particulars of a case. What is needed is a meso-level measure that provides the analytical traction of macro-level measures for understanding patterns with the flexibility afforded by micro-level measures for identifying heterogeneity. A meso-level measure can illuminate specific countries that act as conduits and sinks of illicit flows and how these vary across sectors.

Detailed case studies can be used to understand the specific purposes that trade misinvoicing is used for and the conditions that facilitate the shifting of illicit financial flows. However, these case studies rely on expert knowledge and presuppose knowledge by policy-makers of the existence of the problem. For example, the under-invoicing of exports from Uganda to the United Arab Emirates (UAE) has been attributed to Ugandan companies smuggling gold from conflict regions in the Democratic Republic of Congo (Schuster and Davis, 2020). A 2005 UN Security Council resolution imposed sanctions on gold trade with certain regions of the Democratic Republic of Congo (DRC), notably the Ituri region, to stem the financing of arms for militia and para-military groups. Yet it has been established that large gold trading companies in Uganda (Machanga Ltd and Uganda Commercial Impex) were buying gold from Ituri-based non-state armed groups (Schuster and Davis, 2020). The DRC has not reported export statistics since 1986, while in recent years exports of gold from Uganda have significantly increased despite the country's modest gold reserves. Likewise, the exports of gold that Uganda reports to the UAE are much smaller than what the UAE report to be importing from Uganda. Documented cases of gold that is smuggled from the DRC to Uganda and which is then exported to the UAE have allowed analysts to infer that gold exports from Uganda are under-invoiced in order to disguise illicit capital flight out of the country (Lewis et al., l 24; Schuster and Davis, 2020).

However, these case studies are highly specific and require *ex ante* knowledge of the potential risks of illicit trade. A more systematic approach to identifying which cases to investigate further would be valuable. Thus, meso-level measures can shed light on previously overlooked combinations of trading partners and commodity sectors that merit further investigation, and that might benefit from detailed case studies as a folow-up.

Criterion 6 ***Use open government data.***

A pre-condition for generating a measure that is practically useful is that it can be estimated with open data, to the extent that this is possible. By open data, I refer to data that adheres (as much as possible) to the principles of "open government data".[7] "Open" government data is an ideal-type that espouses a set of eight aspirational properties for data: complete, primary, timely, accessible, machine processable, non-discriminatory (i.e., available to anyone with no requirement of registration), non-proprietary (i.e., available in a format over which no entity has exclusive control), and license free (Tauberer, 2014).

Of course, rare are the datasets that evince all these qualities, but this standard provides a useful benchmark that can be used to compare how far away data used to estimate trade misinvoicing is from this ideal standard. For example, a measure that relies on detailed commodity pricing data compiled by an industry organization and that can only be accessed under restrictive conditions would be, according to this criterion, a

[7] The growing movement of "open government data" aims to increase the accountability of governments to their citizens through greater transparency. See, e.g., the Open Government Partnership: `https://www.opengovpartnership.org/`.

relatively worse measure than a measure that uses government statistics compiled by National Statistical Offices and that can be exploited, with some transaction costs, by researchers.[8]

To summarize, a credible and policy-relevant measure of trade misinvoicing should meet standards of methodological consistency and of practical validity. The first three criteria proposed above are necessary to ensure that a measure of misinvoicing is approximately unbiased and consistent. The last three criteria are pre-requisites for generating a practical measure of trade misinvoicing that has sufficient reach and can be robustly and transparently replicated. This paper now proceeds to first evaluate how extant measures of misinvoicing score on these criteria, and then demonstrates how the "atlas of misinvoicing" measure is a methodological improvement that meets these criteria.

3.3.2 Existing approaches to measuring trade misinvoicing

Several methods attempt to estimate the scale of illicit financial flows, including the proceeds from illegal markets, international corporate tax avoidance, and the amount of capital and wealth held offshore (Cobham and Janský, 2020). Here I focus on reviewing approaches to measuring IFFs that occur in the international trade system only. Existing strategies to estimate misinvoicing in trade can be categorized as looking for anomalies in either transactions, prices, or country-level trade statistics (Cobham and Janský, 2020). This section critically evaluates the extent to which these methods generate estimates that meet the six criteria of a credible measure of misinvoicing. Table 3.1 synthesizes the salient features of the three approaches that are the closest relatives of the method-

[8] More precisely, the example measure that I give will be a worse measure according to this criterion while *holding the other criteria constant*. I do not attempt to solve the optimization problem of maximizing the performance of a measure across all six dimensions, nor do I propose relative weights that should be placed on these characteristics. The paper does not suggest an index-like scoring of the validity of trade misinvoicing measures. Instead, these criteria should be interpreted as heuristics.

ology introduced by this paper, and appraises how well they perform along the requisite analytical dimensions.

First, there exists a category of misinvoicing measures that operate on the transaction-level, contrary to the country-level estimates that are the focus of this paper. Measures that use transaction-level trade data provide evidence of misinvoicing by looking for systematic differences in the reported prices for goods traded between related parties and those traded between unrelated parties (see, e.g., Vicard (2015); Davies et al. (2018)). These approaches are powerful for estimating transfer mispricing within multinational groups but they are less useful for the other types of trade misinvoicing discussed in section 3.2. In addition, it would be highly challenging to obtain the data needed to apply this approach to a broad range of countries and so these measures fare poorly on criteria 4 and 6. These approaches are not discussed further since the nature of the data they use (viz., micro-level data on individual transactions) is different from the measures that leverage country-level trade data.[9] Since the ambition of these measures is conceptually different, i.e., they have different estimands, these approaches are not included in the synthesis table 3.1.

The next category of misinvoicing measures are price-based approaches that look for irregularities in the pattern of prices to detect evidence of illicit financial flows (see, e.g., Hong et al. (2014); Hong and Pak (2017)). The price-filter method calculates per-unit prices for internationally-traded goods and assumes that prices outside a certain range are anomalous, and hence labels the corresponding transaction as an illicit flow, e.g., prices that deviate from the inter-quartile range of the distribution of prices (Zdanowicz, 2004), or prices that are 50% above or below the average price in that country (Zdanowicz, 2009). Building on an example from Zdanowicz (2009), a terrorist wishing to launder

[9] For a critical review, see Cobham and Janský (2020).

$1 million dollars to a foreign country might purchase 10,000 razor blades domestically for 10 cents a piece, export these to a colluding importer in a different country at $100 per razor blade, and thus succeed in moving $1 million to the foreign country less the $1,000 transaction cost of the razor blade. To detect that this is an illicit flow, the estimation technique rests on recognizing that $100 per razor blade is an anomalous price by comparing it to the distribution of prices in that product category. Implicitly, this requires a counterfactual of what the normal price of an arms' length transaction should be (Cobham and Janský, 2020); this information is often unknown.

Consequently, the price-filter method has been criticized for the use of arbitrary thresholds to identify outlier prices and the lack of robustness of the estimates (Nitsch, 2012; Cobham and Janský, 2020; Collin, 2019). While this method has the advantage of meeting criterion 5 of a good measure of misinvoicing because it uses micro-level data that is disaggregated by product category, it struggles to meet the remaining criteria. Observing an aberrant price in the price data could be explained by non-illicit reasons (failing criterion 1), such as the export of a high-quality good in a product category that usually trades in cheap low-value merchandise (Nitsch, 2012). This method provides no systematic way to deal with "benign outliers" and so does not meet criterion 2. Compounding this difficulty, estimates derived from this method have been shown to be sensitive to the inclusion of new data (Nitsch, 2012) and increasing with the price variance of the product category (Collin, 2019), and so this method contradicts the principle of criterion 3.

The practical applications of this method also have some limitations. When the price-filter method is applied to advanced economies, researchers are often able to access detailed trade data that is compiled by a government agency, such as data from the United States Merchandise Trade database from the Department of Commerce's census bureau (see, e.g., De Boyrie et al. (2005)). However, countries with less bureaucratic power might not compile such data or make them easily accessible, and so performance on criterion 6 will be mixed. As a result of the data requirements of the price-filter method, these estimates do not scale easily, and are often provided for a single country's illicit trade with one or more partners, thus scoring poorly on criterion 4.

The next category is the class of estimates that leverage country-level statistics of international trade data (which may be aggregated at the commodity-level or not)[10] – the "atlas" measure falls under this category. Extant country-level estimates can be traced back to two historical approaches: the IMF Direction of Trade Statistics (DOTS)-based method and the UN Comtrade method. Both use bilateral or multilateral mismatches in recorded trade flows to measure trade misinvoicing, but differ in the data source that they employ. The approaches are distinguished by the data they use because this coincides with an inflection point in the literature on estimating trade-based IFFs. The data sources broadly represent a first generation (DOTS-based) and a second generation (Comtrade-based) of estimates.[11] Both methods look for "trade gaps" in the data to detect illicit activity, but with varying degrees of sophistication.

The first generation of this type of misinvoicing estimates were pioneered by the think tank Global Financial Integrity (GFI, see, e.g., Kar and Cartwright-Smith (2008) and Spanjers and Salomon (2017)) and were based on the IMF's DOTS database. The DOTS-based approach leverages asymmetries in the bilateral DOTS data to provide evidence of misinvoicing. As discussed in section 3.2, country-level trade statistics should be recorded

[10] Other approaches to estimate IFFs using country-level statistics use errors and omissions in official Balance Of Payment statistics as a proxy for illicit flows. However, these estimates are concerned with capital flight rather than trade-based misinvoicing. Thus, they cannot provide disaggregated estimates of trade misinvoicing by commodity sector. They are not discussed further.

[11] This categorization is imperfect, as the methods sometimes overlap. Moreover, the estimates by Salomon (2019) use both DOTS and Comtrade data.

twice: once by the reporter i and once by the partner j.[12] Thus, the method compares mirror trade statistics, that is, it compares country i's records of exports to country j with j's reported imports from i in the same year (and *vice versa*), to look for irregularities.

Criterion 2 requires a measure to account for persistent non-illicit reasons for discrepancies, since there are predictable reasons for why an import value is expected to differ from its corresponding mirror export value. The most obvious reason is that records of import values usually include the Cost of Insurance and Freight (hereafter called "CIF" rate or cost) while recorded export values do not. The DOTS-based method adjusts trade gaps for CIF costs but otherwise uncritically equates the CIF-adjusted trade gaps with misinvoicing and risks flagging false positives. As a result, the method has been widely criticized for estimating instances of phantom illicit financial flows and producing results that have no substantive meaning (Nitsch, 2016; Forstater, 2016); it thus fares poorly on 1.

Moreover, the CIF rate has been assumed to be 1.1 by convention which sets the cost of insurance and freight at the constant value of 10%.[13] Treating the CIF costs as constant is a strong assumption that is often not realistic in practice. There are other benign reasons for which mirror trade statistics may not match aside from the cost of insurance and freight, such as asymmetric reporting of re-exports. When goods are re-exported, it is often the case that the re-exporting country will report a time lag in the arrival of shipments (those that are exported in year t and arrive in year $t + 1$).[14] The first generation of DOTS-based misinvoicing estimates adjust for this but through a manual

[12]In practice, data will sometimes contain "orphaned" transactions.

[13]Salomon (2019) revises the assumption from 10% to 6%, but this does not change the tenor of critique: CIF costs are still assumed to be constant.

[14]However, as noted in UNCTAD (2016), discrepancies due to asymmetric reporting of re-exports are eliminated when trade misinvoicing is aggregated at national level on a net basis (i.e., illicit outflows net of inflows).

coding procedure that tries to account for all *known* country data idiosyncracies, such as not counting re-exports from known trading hubs such as Hong Kong as misinvoicing (Spanjers and Salomon, 2017). Therefore, the manual and unsystematic strategies employed to adjust for known non-illicit factors in trade gaps imply that the DOTS-based method only partly satisfies criterion 2.

Another major shortcoming of the first generation DOTS-based method is that it often implicitly treats trade declarations by advanced economies as relatively accurate and consequently assumes that the misinvoicing must have happened in the declarations of developing countries (e.g., Ndikumana and Boyce (2010)). The method has been faulted for its uncritical use of developed countries' trade statistics, without pausing to consider whether those statistics are accurately collected (Mevel et al., 2013). Similarly, the calculation of the total amount of misinvoicing in developing countries with the rest of the world is done through simple extrapolation, which does not account for the possibility of varying levels of misinvoicing and declaration quality across countries. Therefore, criterion 3 is not met.

Finally, in terms of practical applications, the IMF's DOTS database meets all of the criteria of open data standards (criterion 6). The DOTS database is also valued because it has superior country coverage than UN Comtrade (criterion 4) according to Cobham and Janský (2020), though the country coverage of Comtrade is by no means negligible. The limitation of DOTS compared to Comtrade, however, is that it does not provide disaggregated statistics for commodities, thereby limiting its usefulness for sectoral targeting of IFF interventions (criterion 5).

Recognizing the limitations of the DOTS-based approach, the second generation of misinvoicing methods turns to UN Comtrade for more granular estimates, and employs more sophisticated adjustment techniques. This method ("the UN Comtrade method") is used by several United Nations bodies[15] and is used in updated GFI estimates (e.g., Salomon (2019)). Similarly to the DOTS-based method, it employs trade gaps analysis but on trade data that is disaggregated by commodity. Moreover, studies in this category recognize that the cost of insurance and freight will vary by commodity and country pair, and so CIF costs are estimated using data rather than assumed to be constant.[16] The estimates of the CIF margin are based on a gravity-type model of trade costs that takes into account distance between countries and barriers to trade.

Some versions of the UN Comtrade method also adjust for differences in the quality of statistical reporting by using the variance of different partners' reporting to attribute how much misinvoicing occurs at each end (export or import) of the transaction (Mevel et al., 2013). The goal of this econometric adjustment is to eliminate phantom discrepancies that are in fact the result of poor statistical practices in countries' customs (Kravchenko, 2018). Moreover, some studies seek to account for non-illicit reasons for trade gaps by using data on the quantities (rather than the prices) of the commodities being traded. ECLAC (2016); Kravchenko (2018); Salomon (2019) downweigh observations where there is a discrepancy in the reported weight being traded, in order to reduce the impact of instances where the discrepancy is due to either statistical errors, asymmetric reporting of re-exports, or delays in the arrival of shipments. Therefore, the UN Comtrade studies aim to provide an empirical and data-driven way to account for the benign components of the transaction and to adjust for idiosyncracies in the trade declarations, and broadly meet criteria 2 and 3, respectively.

[15] See, e.g., ECLAC (2016); ESCWA (2018); High Level Panel on Illicit Financial Flows from Africa (2015); Kravchenko (2018); Mevel et al. (2013); Schuster and Davis (2020).

[16] One exception is UNCTAD (2016) that uses Comtrade data but still assumes the CIF rate to be constant.

These studies offer improvements from the DOTS-based analyses, but suffer from limitations that preclude their ability to meet criterion 1 whereby trade irregularities (even if netted of CIF) should not necessarily be attributed to misinvoicing. Since the models used to estimate costs of insurance and freight do not factor in the possibility of trade misinvoicing, their estimates of the CIF margin may be picking up trade misinvoicing (Gaulier and Zignago, 2010). Where studies use reported data on the costs of insurance and freight instead of estimates, these data also face challenges because the reported data themselves may be distorted by misinvoicing to avoid detection (moreover, the data are only available for a few countries – see Miao and Fortanier (2017) – and for the single year 2016).[17] One notable exception is the "residual approach" of Gara et al. (2019), who seek to address this issue by estimating a model of trade discrepancies that controls for the main legal determinants of gaps, and then use the residuals from this regression as proxies for the illicit component of such discrepancies. Therefore, Gara et al. (2019) explicitly aim to control for the licit components of a transaction (criterion 2) and employ an estimation strategy geared towards addressing the requirements of criterion 1.

The use of UN Comtrade has shown promise in terms of practical applications. The UN Comtrade database broadly accords with the principles of open government data (criterion 6). The coverage of the database starts from 1961 to the present, though not all countries report trade values in every year; overall, the coverage of Comtrade is good (Cobham and Janský, 2020). The widespread availability of the data and the standardized estimation techniques of the method would make it easier to provide trade misinvoicing estimates on a large scale (criterion 4). In practice, however, many of the studies tend to concentrate on specific geographical regions and do not provide global estimates (see, e.g., ECLAC (2016); ESCWA (2018)). Similarly, the disaggregated com-

[17]See https://unctad.org/news/why-and-how-measure-international-transport-costs.

modity data provided by Comtrade have the potential to generate detailed estimates of trade misinvoicing disaggregated by sector, which provide high value added for policy-makers and can help target interventions. Though there are studies that zoom in to specific sectors (e.g., extractives in Africa (UNCTAD, 2016) or cultural property in the US (Fisman and Wei, 2009)), the full potential of the Comtrade database has yet to be realized, and the performance on criterion 5 stands to be improved.

Method to estimate misinvoicing from trade data			
	Price-filter	**DOTS-based (1st generation)**	**UN Comtrade (2nd generation)**
1	*Poor*. Equates transactions where the price is outside an arbitrary distributional threshold with misinvoicing. No counterfactual for a "normal" price.	*Poor*. Equates CIF-adjusted trade gaps to misinvoicing.	*Poor*. Some equate CIF-adjusted trade gaps to misinvoicing.[18] *Good*. Some use regression residuals to identify misinvoicing.[19]
2	*Poor*. No way to distinguish between outlier prices that are benign or illicit.	*Mixed*. Adjusts for CIF costs but assumes they are constant. Adjusts for re-exporting but manually.	*Mixed*. Some adjust for CIF costs but assume they are constant.[20] Some adjust for re-exporting but manually. *Good*. Some estimate rather than assume CIF costs.[21] Some econometrically isolate legal determinants.[22]
3	*Poor*. Measure is sensitive to sample size and variance of the product-category data.	*Poor*. Assumes that advanced economies report accurately and that the misinvoiced declaration are by developing countries.	*Good*. Some adjust for variance of quality in statistical reporting.[23] Some adjust for discrepancies in reported quantities.[24]

Table continued on next page

[18] UNCTAD (2016); Schuster and Davis (2020).
[19] Gara et al. (2019).
[20] UNCTAD (2016); Salomon (2019).
[21] ECLAC (2016).
[22] Gara et al. (2019).
[23] Mevel et al. (2013); Kravchenko (2018)
[24] ECLAC (2016); Kravchenko (2018); Salomon (2019).

	Method to estimate misinvoicing from trade data		
	Price-filter	**DOTS-based (1st generation)**	**UN Comtrade (2nd generation)**
4	*Poor.* Coverage often limited to a single or a group of countries.	*Good.* Better country coverage than Comtrade.	*Good.* Broad country coverage (>100).
5	*Good.* Provides detailed estimates within disaggregated product categories. Useful for audit purposes.	*Poor.* No disaggregation by commodity. Bilateral estimates only.	*Mixed.* Some provide detailed commodity results but only for certain regions or countries.[25]
6	*Mixed.* Some governments compile detailed transaction-level statistics (e.g., US Census Bureau) but this does not apply generally.	*Good.* IMF DOTS database is open data.	*Good.* UN Comtrade database is open data.
E.g.	De Boyrie et al. (2005); Hong et al. (2014); Hong and Pak (2017)	Ndikumana and Boyce (2010, 2018); Early GFI estimates (Kar and Cartwright-Smith, 2008; Spanjers and Salomon, 2017)	Fisman and Wei (2009); Later GFI estimates use DOTS and Comtrade (Salomon, 2019); UN regional commissions (ECLAC, 2016; ESCWA, 2018; High Level Panel on Illicit Financial Flows from Africa, 2015; Kravchenko, 2018; Mevel et al., 2013; Schuster and Davis, 2020)

Table 3.1: Appraisal of existing trade misinvoicing measures with respect to the 6 desired properties.

3.3.3 Features of the "atlas" measure

This section presents the main features of the "atlas" measure and demonstrates how they meet the criteria of a credible measure of trade misinvoicing. First, an abridged summary of the methodology is provided, and then the major methodological improvements of the approach are highlighted. The detailed steps to reproduce the measure are given in section 3.4.

The strategy exploits the principle of double-entry accounting in international trade statistics to identify illicit trade gaps, an approach that has an extensive history in development economics (see Morgenstern (1950); Bhagwati (1964); Morgenstern (1974)). The methodology is most similar to the UN Comtrade method described above, but offers several refinements. A bilateral trade transaction is recorded twice in UN Comtrade: once from the perspective of reporter i who declares the value of imports (exports) from its partner j, and once from the perspective of the corresponding partner j who reports the mirror exports (mirror imports). In theory, these mirrored values should be equal to one another, plus or minus unobserved latent factors, and statistical noise. Moreover, the quality of countries' declarations to UN Comtrade will vary according to country, commodity, and year-specific idiosyncracies. The true unknown value of the trade is assumed to lie somewhere in between: it is a convex combination of declarations made by i and j. The "atlas" method adopts both a *residual* and a *reconciliation* approach to estimating misinvoiced trade. First, reported imports are "cleaned" from predictors of trade discrepancies and converted to a FOB basis. Second, the harmonization procedure suggested by Gaulier and Zignago (2010) is applied to produce a "reconciled value" of the trade, which is a weighted average of reporter and partner declarations according to the quality of the declaration of each country. The weights corresponding to declaration quality are calculated according to a regression of trade gaps on reporter, partner, commodity, and year fixed effects to isolate the relative quality of declarations by i and j (Gaulier and Zignago, 2010). The procedure is applied twice to generate a reconciled value for imports and one for exports. Finally, misinvoiced imports are calculated as the difference between the reconciled import value (which has been stripped of the licit predictors of trade gaps, hence the "residual" approach) and reported imports; while misinvoiced exports are equal to the difference between the reconciled export value and reported exports.

The methodology offers some refinements that are designed to ameliorate long-standing problems in the estimation of trade misinvoicing that have been highlighted in the literature. These innovations are designed to improve the validity of trade misinvoicing estimates according to the criteria established in section 3.3.1.

Criterion 1 ***Avoid uncritically equating observed trade irregularities with misinvoicing.***

The methodology does not directly use (adjusted) trade gaps as proxies for misinvoicing. With the exception of Gara et al. (2019) who use the residuals of an econometric regression of trade gaps on legal determinants as the proxy, the other studies presented earlier that use either the DOTS-based or the UN Comtrade method have in common that trade misinvoicing is taken to be some measure of trade gaps between reported and mirror trade values, that may or may not have been adjusted for transport costs and/or re-exporting distortions. However, the existing econometric models that have been used to estimate the CIF margin do not factor in the possibility of misinvoicing, and so run the risk that the adjustment factor used to net import values from the cost of insurance and freight is actually picking up misreporting rather than transaction costs.

By contrast, the methodology presented here takes additional precautions to avoid uncritically equating trade irregularities with illicit activity. Misinvoiced trade is calculated indirectly using a "residual" approach that takes the difference between a harmonized value that represents the best quality estimate of the transaction, and import values that have been cleaned of the licit predictors of discrepancies. Since import values are systematically cleaned from most licit predictors, the remaining discrepancies must be due to illicit factors and statistical noise. Moreover, this value is not directly compared to the mirror trade value, but rather to a harmonized value that takes into account the quality of declarations. Details on this calculation are provided in section 3.4.3. Therefore, the strategy to avoid indiscriminately deducing IFFs from observed trade irregularities rests on both a residual and a reconciliation/harmonization (which are used interchangeably here) strategy.

Note that this "residual" approach (indirectly) assumes that remaining trade discrepancies that *cannot* be accounted for due to benign reasons are the result of *either* deliberate misinvoicing or statistical noise. In an alternative approach, one could assume that only the portion of trade discrepancies that *are* explained by predictors of illicitness are related to trade misinvoicing. However, this approach would suffer from a major limitation: predictors of illicit activity for which there is good data cover only a small share of the motivations for trade misinvoicing, and estimating trade misinvoicing as the share of trade discrepancies attributable to these factors would likely miss the majority of trade misinvoicing. For this reason, the indirect approach of the "atlas" method is preferred.

Criterion 2 ***Partition the trade transaction into licit and illicit components in order to account for persistent non-illicit reasons for discrepancies.***

The "atlas" methodology remedies one of the main criticisms levelled against extant misinvoicing measures – that trade gaps could in fact be due to persistent non-illicit reasons rather than foul play – by explicitly partitioning the trade transaction into its respective licit and illicit components (plus statistical noise). In related work, Fisman and Wei (2009) and Kellenberg and Levinson (2019) use econometric models to estimate the share of trade discrepancies due to predictors of the level of illicit activity in an economy, e.g., corruption. Though they do not do so, Fisman and Wei (2009) point out that one could estimate trade misinvoicing based on such a model, that is, by assuming that the portion of trade discrepancies that is not explained by predictors of licit discrepancies (e.g., distance between countries, reporting mistakes, etc.) is due to trade misinvoicing. This is the "residual" approach that this paper adopts (though the "atlas" model uses different predictors and also conducts an additional "harmonization" step). By explicitly including predictors of both licit and illicit discrepancies in the regression, the "atlas" measure seeks to estimate more accurately both a) what portion of trade discrepancies is actually explained by trade costs and other benign factors and b) what portion is illicit.

Moreover, the method supplements the traditional predictors of CIF costs (such as distance or barriers to trade) with a new approach to econometrically adjust for asymmetric reporting of re-exports and delays in the arrival of shipments. Full details on how the estimated trade gaps are partitioned are given in section 3.4.3.

Criterion 3 ***Account for the variance in countries' statistical reporting.***

The third main innovation of the "atlas" measure is that it does not take country declarations as given. The first generation of estimates implictly assumed that reporting from developed countries could be better trusted than declarations from poorer countries (see, e.g., Ndikumana and Boyce (2010)). While it may be the case that economic development correlates with the robustness of a country's statistical reporting procedures (Jerven, 2013), this is not necessarily always the case, and hence this imposed a strong assumption on the problem. Likewise, making *no* adjustment between the reporter declarations and the partner declarations makes the implicit assumption that the declarations on either end of the transaction are equally precise, which is not likely to hold in practice.

The approach presented here addresses this problem by empirically determining the relative quality of reporter and partner declarations. In addition, the quality of reporting may differ not only due to country idiosyncracies, but also due to the particularities of the reporting regime for a certain commodity (see the example of gold described under criterion 1) and year-specific shocks. Therefore, the "atlas" measure presented in this paper follows a reconciliation procedure proposed by Gaulier and Zignago (2010) to improve the quality of bilateral trade statistics. A reconciled valued of the trade is calculated using weights that minimize variance and adjust for country, commodity, and time-specific idiosyncracies. Reporting distances are estimated using an econometric model that contains reporter, partner, commodity, and year fixed effects. This has the effect of estimating the quality of a given country's customs declaration independent of its product specialization (Gaulier and Zignago, 2010). Finally, the harmonization procedure computes a variance-minimizing weighted average of country declarations to ascertain with greater precision the value of the trade on an FOB basis. See 3.4.3 for further details of this procedure.

The methodological refinements offered above strive to increase the theoretical cogency of the measure. Next, the features of the "atlas" measure described below pertain to its practical usability by academics and practitioners.

Criterion 4 ***Scale across countries and over time.***

There is broad academic and policy interest in obtaining a measure of trade misinvoicing that has a wide coverage (Cobham and Janský, 2020; UNECA, 2018a) to obtain a global picture of the extent of illicit finance. The "atlas of misinvoicing" provides comprehensive bilateral estimates of misinvoicing for 167 jurisdictions over 2000-2018.

This is possible thanks to the nature of the data source that is used (UN Comtrade) and to the relatively undemanding data requirements of the methodology. As mentioned earlier, some studies use reported quantities of the traded goods to adjust for the quality of country declarations. While this method has its merits, it may ignore misinvoicing where reported quantities or weights are different, e.g., where shipments are smuggled at either export or import but not both or where weights are mis-stated (Forstater, 2018). Moreover, the method is not applicable for countries that do not report weight or quantity data, which is the case for most African countries. The "atlas" measure only relies on observations of the price of the traded good, which has much better coverage than data on quantities, which permits the scaling of this measure across many countries and over time. The nature of the data and a more detailed description of the methodological choices regarding the data are described in section 3.4.1.

Criterion 5 ***Provide enough granularity to support policy prioritization.***

The "atlas" method generates estimates that are disaggregated by commodity sector at a level of resolution that allows sectoral analysis, but that is not so disaggregated that the results are less robust (for more details, see section 3.4.1).

In order to support evidence-based policy-making in the fight against illicit financial flows, the estimates are disaggregated by trading partner and by commodity. The initiatives needed to combat illicit flows will sometimes necessitate a sectoral approach or regional cooperation, and the disaggregated estimates will be useful to indicate where those initiatives might bear fruit.

The "atlas" database also offers summary datasets that present aggregate results and are designed to facilitate further analysis by researchers and to support targeted policy interventions. These datasets demonstrate the different lenses that can be applied to the "atlas" measure (e.g., by country, sector, etc.); they variously provide gross outflows, gross inflows, and net flows by income group, geographical region, development status, and commodity sectors.

Criterion 6 ***Use open government data.***

Finally, the "atlas" method makes use of the UN Comtrade database which, as mentioned above, broadly meets the criteria of open government data. Moreover, the method does not require any additional data, such as a separate database of transport costs (see, e.g., Schuster and Davis (2020)). The less onerous data requirements of this method further facilitate its accessibility and reproducibility by interested researchers and other stakeholders. The results of the "atlas" method are available online in a publicly available database.[26]

While this section has highlighted the salient features of the "atlas of misinvoicing" measure that seek to offer various refinements, the following section provides a full account of the methodology and detailed steps to replicate the measure. Finally, the "atlas" has surfaced key insights about the global, regional, and sectoral patterns of illicit trade;

[26] Available at https://doi.org/10.5281/zenodo.3610557.

those findings are presented in section 3.5.

3.4 Methodology

3.4.1 Data

The data used by the "atlas" measure come from the United Nations Commodities Trade Database (Comtrade), which provides disaggregated commodities trade data using the Harmonized System (HS), the international nomenclature for trade classification, which assigns commodities to a certain product category that can be hierarchically mapped to a less detailed product category, and so on. The entire Comtrade database was scraped for all participating jurisdictions and all commodities over a panel of 20 years. At the lowest level of commodity aggregation, the raw data contains approximately 490 million observations. The "atlas" measure uses data at the 2-digit level of aggregation, which is made up of 99 "chapters" (groupings of commodities). The raw data panel contains trade flows from 1999 to 2019 for 236 distinct jurisdictions. Prior to implementing data cleaning procedures, the sample size at the 2-digit level of aggregation is 23,266,944. One unit of observation consists of a reporter-partner-commodity-year quadruple, where the commodity belongs to one of the 99 HS chapters.

The 2-digit level is chosen to avoid the risk that accidental misreporting of the customs code by customs officers, or differences in national nomenclature (see Van Rensburg (st 1)), result in "false positive" identification of trade misinvoicing. A plausible assumption is that, while the 6-digit or 4-digit code may be incorrectly reported due to the number of detailed product categories that could be assigned to a shipment, this is less likely with the 2-digit code since it represents a higher level of aggregation. This means that the estimates are conservative, since they will leave out instances where the customs code is deliberately falsified to benefit from lower taxes or subsidies, but where the false customs code still falls within the same 2-digit chapter as the correct code (see Kravchenko (2018)). Moreover, this will also result in "within-sector netting", i.e., inflows and outflows between the same country pair for the same 2-digit commodity code will be netted against one another. Therefore, researchers can interpret estimates as a lower bound.

3.4.2 Notation and conceptual model

It is instructive to define the notation that will be used throughout the rest of the paper. Let i index reporters, j index partners, c denote the commodity, and t denote the year. The value of the trade is denoted by V, and superscripts denote whether the trade flow corresponds to an import V^M or an export V^X. For ease of exposition, it is usually possible to remove the commodity and year subscripts without loss of generality.

Exports are considered net of re-exports, that is, $V_{ij}^X = V_{ij}^{exports} - V_{ij}^{re-exports}$. Unless otherwise stated, when the paper refers to exports, it designates net exports.

The declarations in Comtrade are thus:

V_{ij}^M Imports reported by country i from country j

V_{ij}^X Net exports reported by country i to country j

V_{ji}^X Net exports reported by i's partner, which is the mirror value of V_{ij}^M

V_{ji}^M Imports reported by i's partner, which is the mirror value of V_{ij}^X

As explained in section 3.2, the estimand of interest is the amount of trade misinvoicing both in the imports and the exports of the *reporters*. In turn, estimating the misinvoicing for reporters will yield the misinvoicing for partners. Therefore, "import values" and "export values" will refer to the import and export declarations, respectively, made by i. References to "mirror values" denote the corresponding trade flow recorded by the partner j. Since illicit flows are estimated from the perspective of the reporter, an illicit outflow will be considered to flow out of reporter i to partner j, and an illicit inflow will be considered to flow into reporter i from partner j.

The "atlas" method models the import transaction that is declared by reporter i as:

$$V_{ijct}^{M} = V_{jict}^{X} + licit + illicit + u_{ijct} \tag{3.1}$$

According to the model, imports reported by country i from partner j are equal to what the partner declared that it exported to country i, some amount of licit discrepancies (which can be positive or negative) due to benign or non-illicit reasons, trade misinvoicing (which can be positive or negative), and statistical noise.

Likewise, the export transaction occurring at i's customs is conceptualized as:

$$V_{jict}^{M} = V_{ijct}^{X} + licit + illicit + v_{ijct} \tag{3.2}$$

These two models underpin the "atlas" method of estimating the illicit financial flows that occur at a given country i's customs in both imports and exports, respectively.

3.4.3 Step-by-step procedure to calculate misinvoicing in imports and exports

The illicit flow in each transaction is estimated following a strategy that proceeds in three broad steps:

1. Estimate the discrepancies between mirror trades as a function of both licit and illicit predictors.

2. Perform a harmonization procedure in order to generate a reconciled value that represents the best estimate of the FOB value of the trade taking into account the relative quality of the declaration by the countries.

3. Calculate the IFF embedded in each transaction as the difference between the observed value (adjusted to remove the contribution of licit predictors) and the reconciled value.

The specific steps are detailed below.

Data cleaning

First, data cleaning procedures are implemented to remove unmatched or orphaned transactions (i.e., transactions that do not have a corresponding mirror value), and to remove observations that do not correspond to countries.[27] The sample size decreases from $n =$ 23,266,944 to $n =$ 2,559,456. The large drop partly reflects the fact that there exist many orphaned transactions where, for any given country, commodity, and year, the import declaration V^M_{ijct} is in the data but the corresponding mirror export value V^X_{jict} does not

[27] Comtrade also provides declarations where the partner is not an individual jurisdiction, but an aggregate, e.g., "World", "Other Europe, not reported elsewhere", etc.

exist, or where the export declaration V^X_{ijct} is observed but the mirror import value V^M_{jict} is not. Missing mirror values in the data could either be due to illicit activity or could be explained by other factors such as shaky statistical reporting practices of certain customs authorities (Jerven, 2009) – though it is not easy to disentangle those reasons. Since the estimation strategy of the "atlas" measure relies on bilateral trade asymmetries to calculate misinvoicing (though with adjustments, as discussed), it will not capture all types of illicit activity that can occur with merchandise trade – but this is a feature, not a bug, of trade misinvoicing measures where the estimand of interest is "technical smuggling", as opposed to "pure smuggling" where goods (e.g., illicit drugs) are exported clandestinely from a country and imported clandestinely into another and which as result will not be reflected in trade gaps (Schuster and Davis, 2020).

A further 15,264 observations are removed where the observed trade gap is greater than 100, to throw out cases that might be due to genuine and egregious statistical mistakes in reporting (e.g., reporting values in dollars versus thousands of dollars). Various thresholds were experimented with and the results remain robust. Following this, a statistical cleaning procedure is performed which removes observations that have a Cook's Distance greater than 2 (no cases), and iteratively drops statistically significant outliers with Bonferonni correction (not exhaustive).

After the data cleaning procedures are completed, the resulting panel covers 167 distinct reporting and partner jurisdictions, and has a sample size of $n = 2{,}446{,}679$.

Fitting gravity models

For any given country in the sample, the goal is to estimate all the trade misinvoicing that occurs at its customs, both when a country reports imports and when it reports exports to Comtrade. Therefore, two gravity models are econometrically fitted that represent the gap between the trade flow (import or export) reported by country i and the mirror trade flow reported by partner j (mirror export or mirror import, respectively) as explained by legitimate factors (e.g., reporting mistakes), discrepancies due to trade misinvoicing, and statistical noise. Since the "atlas" method operates from the perspective of the reporting country i, there are two models of the gaps between, on the one hand, reported imports and mirror exports, and on the other, reported exports and mirror imports.

As discussed previously, the methodology proceeds in this way in order to estimate illicit trade for the entire set of countries that report to Comtrade, and where the reporter i is the proverbial "atlas" from whose perspective illicit trade is systematically estimated, for both imports and exports.

Therefore, two gravity models of the form below are fitted:

$$\ln\left(\frac{V_{ijct}^{M}}{V_{jict}^{X}}\right) = \alpha_0 + \boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{Z}\boldsymbol{\gamma} + \epsilon_{ijct} \tag{3.3}$$

and

$$\ln\left(\frac{V_{jict}^{M}}{V_{ijct}^{X}}\right) = \beta_0 + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\lambda} + \xi_{ijct} \tag{3.4}$$

where the dependent variable is the gap between the transaction reported by i and the mirror transaction declared by partner j, $\boldsymbol{X}$ is a vector of licit explanations for discrepancies, and $\boldsymbol{Z}$ is a vector of illicit determinants of discrepancies. In both cases, the import values are the numerator of the trade gap, because the trade literature conventionally estimates trade gaps as a CIF-FOB margin between imports and exports (Yotov et al., 2016). Note that the import value in the numerator of the outcome variable in equation (3.4) corresponds to the *mirror* import value that partner j declares to Comtrade of its imports from i, while the export value in the denominator is the declaration from reporter i.

The objective is to partition the trade transaction into its respective licit and illicit components, as exhorted by criterion 2 of a credible misinvoicing measure. A transaction reported by i should be equal to the mirror value declared by i's partner j, plus factors explaining observed discrepancies, plus statistical noise. Import declarations include CIF and so will need to be converted to a FOB basis to be comparable to exports; this is accomplished by dividing the import declaration with the estimated coefficients on the factors that can explain observed discrepancies, thus "stripping" import values of the margin (which in existing methods is assumed to reflect transaction costs only) that is econometrically estimated. Moving the mirror declarations (i.e., what the partner j declares to Comtrade) to the left-hand side in equations (3.1) and (3.2) and taking logs will yield the gravity models noted in equations (3.3) and (3.4), respectively. Therefore, two models are fitted where the dependent variable is the gap between i's imports and the mirror net exports, and where the dependent variable is the gap between i's net exports and the mirror imports, respectively.

The innovation of the "atlas" methodology is that it explicitly partitions the factors that can explain trade discrepancies into those that can be attributed to benign reasons (captured in $\boldsymbol{X}$), and those that can be ascribed to underlying illicit activity (captured in $\boldsymbol{Z}$).

Thus, the vector $\boldsymbol{X}$ contains the predictors associated with non-illicit reasons of observed gaps between between mirror trade values. First, it includes traditional "gravity" variables representing various geographical factors that can be responsible for transportation and other transaction costs (Anderson, 1979; McCallum, 1995) from CEPII's *Gravity* database (Conte et al., 2021):

- $\texttt{dist}_{ij}$ and $\texttt{dist}^2_{ij}$, which are the distance between a country pair and the squared distance between a country pair, respectively;

- $\texttt{contiguous}_{ij}$, a dummy variable indicating whether the countries share a border;

- $\texttt{landlocked}_i$ and $\texttt{landlocked}_j$, which are a dummy variable indicating whether the reporter is landlocked, and a dummy variable indicating whether the partner is landlocked, respectively.

Moreover, year fixed effects are added to the models in order to control for period-specific idiosyncrasies in reporting, because a period-specific shock that affects each country's trade equally (e.g., a trade shock like a global pandemic) might partly explain the observed trade gap, for entirely non-nefarious reasons. Therefore, the vector $\boldsymbol{X}$ includes a series of year-specific indicator variables τ_t for the years $t = 2001, \ldots, 2018$ that are equal to 1 if τ_t corresponds to the year of the transaction, and 0 otherwise (omitting the first year since the models include an intercept). Implicitly, including the estimated year-specific intercepts in the vector of parameters on $\boldsymbol{X}$ and not in the vector of parameters on $\boldsymbol{Z}$ assumes that any factor leading to discrepancies that varies over time but is constant across countries is not due to illicit factors. This assumption is relatively plausible as it is difficult to think that there would be a sudden increase or decrease of criminal activity across countries globally, for instance.

The models also econometrically adjust for the other legitimate reasons that might readily explain discrepancies in bilateral trade statistics, such as when shipments arrive at their destination in a different calendar year from when they departed the country of origin, or when the asymmetric reporting of re-exports to third countries creates the illusion of discrepancies between dyads (as illustrated by the "false positive" example of Zambian copper). Thus, the "atlas" method avoids uncritically equating observed trade irregularities with misinvoicing that could be due to artifices of the recording process; and meets criterion 1 of a rigorous measure of misinvoicing.

Moreover, it is expected that the dependent variable is autocorrelated and that present values of trade gaps will depend on past values of trade gaps; and the models therefore include a lag of the dependent variable. Again, all of the factors described so far are assumed to represent persistent non-illicit reasons for discrepancies, and so they are included in the vector $\boldsymbol{X}$.

The operationalization of these explanatory variables will differ according to whether the reported transaction by i is imports or (net) exports:

- $V^M_{ijc,t+1}/V^M_{ijct}$ to capture the misreporting of imports at $t+1$ in model (3.3), and $V^M_{jic,t+1}/V^M_{jict}$ in model (3.4);
- $V^{re-exports}_{jict}/V^M_{ijct}$ to capture the misreporting of re-exports in model (3.3), and $V^{re-exports}_{ijct}/V^M_{jict}$ in model (3.4);
- $\ln\left(V^M_{ijc,t-1}/V^X_{jic,t-1}\right)$ to capture the persistence across periods in model (3.3), and $\ln\left(V^M_{jic,t-1}/V^X_{ijc,t-1}\right)$ in model (3.4).

Next, the models include, $\boldsymbol{Z}$, a vector of illicit determinants of discrepancies, composed of:

- $\texttt{corruption}_{it}$ and $\texttt{corruption}_{jt}$ of the reporter and partner, respectively, in any given year in the sample;

- $\texttt{PoorRegulation}_{it}$ and $\texttt{PoorRegulation}_{jt}$ to capture poor regulatory quality in the reporter and the partner country, respectively, in any given year in the sample;

- $\texttt{tariff}_{ijct}$ which is the average tariff imposed by reporter i on imports of commodity c from partner j in year t in equation (3.3); and $\texttt{tariff}_{jict}$ which is the average tariff imposed by j on imports from i, i.e., the tariff imposed on mirror imports used in equation (3.4).

The variables `corruption` and `PoorRegulation` are obtained from the *Worldwide Governance Indicators* (WGI) database, and capture perceptions of the extent to which public power is exercized for private gain, and perceptions of the government's ability to formulate and implement sound policies that permit private sector development, respectively (Kaufmann et al., 2010). The inverse of the variables from the WGI database is taken so that the high end of the variables (measured by percentile rank) corresponds to high amounts of corruption and poor regulatory quality.

The tariff measure is from the UNCTAD TRAINS database (UNCTAD, 2018) and captures the incentives to misinvoice imports in order to evade tariffs.

Therefore, the estimates of the coefficients on known licit reasons for discrepancies will be contained in the parameter vectors $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$, depending on whether the import or the export transaction, respectively, is modeled. Likewise, coefficient estimates for illicit factors will be contained in the parameter vectors $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\lambda}}$, respectively. These coefficients estimate the portion of the trade gap that is explained by licit (illicit) factors conditional on the illicit (licit) factors. In other words, they represent how the CIF-FOB margin varies as a result of changes in one group of factors (licit or illicit), while holding the other group of factors constant. Hence, any estimates of legitimate transport and other trade costs will be stripped of the effect of any illicit factors.

To improve the normality of the data, highly skewed predictor variables are transformed prior to fitting the gravity models. The lagged dependent variable and the variable capturing the misreporting in different calendar years are logged. The inverse hyperbolic sine transformation is applied to the variable capturing the misreporting of re-exports, since it cannot be logged due to the presence of zeroes.[28]

The vectors of parameters associated with licit and illicit factors (plus a constant) are estimated by fitting the gravity models in (3.3) and (3.4) using Ordinary Least Squares (OLS) on pooled data for the period 2000-2018. The advantage of using linear regression rather than a more flexible non-parametric model such as a Generalized Additive Model (GAM) is that it provides estimates of licit predictors that hold illicit predictors constant and vice versa. This is useful to calculate import values that are "cleaned" from benign predictors which allows interpreting the remaining discrepancies as the marginal effects due to illicit activity and statistical noise.

The estimated regression coefficients in the model where the reporter i's declaration is imports are displayed in the left-hand side column of Table 3.2 below, and the estimated coefficients in the model where the reporter i's declaration is exports are provided in the right-hand side column. Next, the estimated coefficients are briefly discussed.

[28]The inverse hyperbolic sine function is defined as $ihs(x) = \ln(x + \sqrt{x^2 + 1})$. It can be used to reduce the skew in data where the natural log cannot otherwise be taken (since $\ln(0)$ is undefined).

	Dependent variable	
	ln.ratio_CIF	ln.ratio_CIF_mirror
	(1)	(2)
dist_t	-0.007***	-0.007***
dist_t.sq	0.000***	0.000***
contiguous	-0.156***	-0.154***
landlocked_i	0.124***	-0.098***
landlocked_j	-0.091***	0.123***
ln.FutImport_misp	-0.266***	
ihs.ReExport_misrep	0.028***	
ln.ratio_CIF_lag	0.452***	
tariff	-0.001***	
ln.FutImport_misrep_mirror		-0.282***
ihs.ReExport_misrep_mirror		0.010***
ln.ratio_CIF_lag_mirror		0.443***
tariff_mirror		-0.001***
corruption_i	-0.002***	0.001***
corruption_j	0.001***	-0.002***
PoorRegulation_i	0.000**	-0.000**
PoorRegulation_j	-0.000	0.000**
Constant	0.133***	0.138***
Year fixed effects	Yes	Yes
Observations	2,446,679	2,446,679
Adjusted R^2	0.336	0.320
Residual Std. Error (df = 2446647)	1.136	1.180
F Statistic (df = 31; 2446647)	39,925.120***	37,126.330***

Notes: *p<0.1; **p<0.05; ***p<0.01

The dependent variable ln.ratio_CIF corresponds to $\ln\left(V^M_{ijct}/V^X_{jict}\right)$.

The dependent variable ln.ratio_CIF_mirror corresponds to $\ln\left(V^M_{jict}/V^X_{ijct}\right)$.

Table 3.2: Regression results.

The coefficient on the distance between transacting partners (`dist`) is negative and statistically significant in both models, which runs counter to the intuition that shorter distances should be associated with smaller transport costs. Yet, the inverse relationship between distance and the trade gap is a persistent empirical result in international economics, and has been dubbed the "distance puzzle" (see, e.g., McCallum (1995); Anderson and Van Wincoop (2003); Disdier and Head (2008); Yotov (2012)). Moreover, a non-linear relationship between distance and the discrepancy in mirror statistics is expected to the extent that, for greater distances, the price discrepancy is likely to be even larger (Gaulier and Zignago, 2010; McCallum, 1995; Yotov et al., 2016). This hypothesis is supported by the fact that the coefficient on the squared distance term (`dist.sq`) is positive and statistically significant.

The coefficient on the dummy variable indicating whether the trading countries are geographically contiguous (`contiguous`) is negative and statistically significant in both models, which is to be expected.

While the coefficient in model (1) on the dummy indicating whether the reporting country (i.e., the importer) is landlocked is positive, the corresponding coefficient for the partner country (i.e., the exporter) is signed contrary to expectation. If part of the price discrepancy is due to access and transport costs, the price discrepancy would be expected to rise if a country is landlocked, everything else constant. Nevertheless, Gaulier and Zignago (2010) also find a negative sign on the coefficient on landlocked exporters. Model (2) reports similar findings, where the coefficient is negative for landlocked exporters and positive for landlocked importers (the reporter and partner, respectively, in this model).

The coefficient on the misreporting of imports in the next calendar year is negative and statistically significant for both models. That is, when the ratio between a given country's imports at time $t+1$ and at time t increases (for the same partner and commodity), indicating that shipment arrivals were higher in the next calendar year, then the price discrepancy between imports at time t and corresponding mirror exports tends to decrease (holding other factors constant). This suggests that part of the observed discrepancy in bilateral trade statistics is simply due to calendar differences in the recording of shipments.

Both models also control for the misreporting of re-exports by including the share of re-exports in the other country's imports as an independent variable. Re-exports are the exports of foreign goods in the same state as previously imported and are recorded by the re-exporting country as exports.[29] However, the country of final destination (i.e., the importer) will tend to see the goods as coming from the country where value was last added, that is, earlier on in the value chain. This introduces artificial discrepancies in bilateral trade statistics. To counter this, the dependent variable uses exports net of re-exports, which thus represents exports of domestic goods only. Moreover, the share of a partner's re-exports in the corresponding reporter's imports for a particular commodity-year is included as an explanatory variable. Results for both models show that an increase in that ratio is associated with an increase in the observed discrepancy in mirror trade statistics, *ceteris paribus*; which supports the hypothesis that part of the gaps in mirror trade statistics can be explained by the misreporting of re-exports.

Finally, as expected, the positive and statistically significant coefficient on the lagged value of the dependent variable in both models suggests that observed discrepancies are persistent over time.

All the coefficients discussed so far can be treated as non-illicit predictors of observed

[29] See Comtrade database description: https://unstats.un.org/unsd/tradekb/Knowledgebase/Reexports-and-Reimports.

discrepancies in bilateral data. They capture either legitimate factors that would increase the price of imports, such as the cost of freight, or reflect artifices that occur during the recording of the data. Next, the coefficients that capture drivers of the discrepancies that may have an illicit motivation or nature, such as escaping barriers to trade or poor governance, are discussed.

One of the more surprising results in the models is that import tariffs are associated with a decrease in the observed price discrepancy (everything else held constant): the coefficient on the average tariff line imposed by a country on a specific commodity-year-exporter is negative and statistically significant (in both models). This finding is robust to different model specifications. One possible explanation is as follows. Customs officials are trained to protect revenues rather than to look for misinvoicing that may occur for other reasons (Mikuriya, 2018). As such, they are likely to concentrate their audit efforts on shipments with high *ad valorem* tariffs attached. If misinvoicers are aware of this, they are more likely to direct the bulk of their faking efforts on items at lower tariff lines to evade detection. This phenomenon would explain the negative sign on the `tariff` coefficient. Jean et al. (2018); Kellenberg and Levinson (2019) also find that higher tariffs may result in lower customs duty evasion. In addition, Patnaik et al. (2008) also find that higher tariffs result in lower over-invoicing of imports, which is a key source of illicit financial outflows. They note that higher tariffs would reduce the incentive to over-invoice for imports as doing so would result in firms having to pay higher tariffs.

To capture poor governance in the transacting countries, the models include variables measuring corruption and poor regulatory quality with respect to private sector development. Most of these coefficients are signed according to expectation. However, the coefficients on corruption in importers are negative and statistically significant (`corruption_i` in the first model and `corruption_j` in the second model correspond to the importer).

Likewise, the coefficients on poor regulation in exporting countries are negative. This may be because poor governance reduces trade misinvoicing to the extent that it makes other channels of illicit financial flows (e.g., use of the formal financial system, cash smuggling, etc.) easier to use (Ferwerda et al., 2013), or to the extent that those involved in illicit finance have less of a need to hide their illicitly-obtained funds abroad, reducing the extent of illicit outflows for a given level of proceeds of corruption (Walker, 1999).

Finally, potential issues of multicollinearity are examined by looking at the Variance Inflation Factor (VIF) scores for the coefficients in each model, reported in Table 3.3 below. The high VIF for the coefficients on distance and distance squared are not a cause for concern and are to be expected given that the models include a quadratic term and its lower-order term. Multicollinearity does not bias OLS coefficients, but it does inflate standard errors, making it harder to detect statistically significant relationships. The high VIF values for the variables capturing corruption and poor regulation occurs because they are highly correlated with each other – indeed, a highly corrupt country is likely to have a poorly governed regulatory system. The high VIF value for poor regulatory quality in the partner country ($\texttt{PoorRegulation}_j$) might explain why this coefficient is not statistically significant in model (1). Despite this, the variable is still included in the model since poor governance, as a driver of misinvoicing, is likely to operate on both sides of the transaction. Moreover, the estimates of interest here are the implied CIF rates due to legitimate and illegitimate predictors, which are found by accounting for the marginal effect of coefficients, which will still be unbiased despite the multicollinearity. Thus, it is important to retain theoretically important predictors in the model.

	Model	
	(1)	(2)
dist	15.22	15.22
dist.sq	15.319	15.319
contiguous	1.091	1.09
landlocked_i	1.026	1.026
landlocked_j	1.026	1.026
ln.FutImport_misrep	1.011	
ln.FutImport_misrep_mirror		1.011
ihs.ReExport_misrep	1.011	
ihs.ReExport_misrep_mirror		1.011
ln.ratio_CIF_lag	1.015	
ln.ratio_CIF_lag_mirror		1.015
tariff	1.046	
tariff_mirror		1.046
corruption_i	6.864	6.865
corruption_j	6.869	6.867
PoorRegulation_i	6.955	6.902
PoorRegulation_j	6.905	6.958
factor(year)	1.046	1.046

Table 3.3: Variance Inflation Factors of the models.

FOBization of imports

After estimating the gravity models, the third step is to "FOBize" imports by deflating them from transport and other costs, so that they are on the same basis as export declarations, in order to be able to compare them.

Subscripts for commodities c and years t are henceforth omitted for simplicity.

FOBized imports for reporter i are given by:

$$V_{ij}^{M;FOB} = \frac{V_{ij}^{M}}{\exp(\hat{\alpha}_0 + \boldsymbol{X}\hat{\boldsymbol{\alpha}} + \boldsymbol{Z}\hat{\boldsymbol{\gamma}})} = V_{ji}^{X} \cdot \exp(\hat{\epsilon}_{ij}) \tag{3.5}$$

The reported import value is stripped of the implied CIF margin given by the estimated coefficients in the gravity model represented in equation (3.3). Note that this formulation implies that FOBized imports are equal to mirror exports plus statistical noise.

As a robustness check, the residual $\hat{\epsilon}_{ij}$ was also stripped from import values, to investigate the consequences of a differing assumption which would hold that the CIF margin includes statistical noise. The findings remain similar.

Equivalently, FOBized imports for partner j are calculated as:

$$V_{ji}^{M;FOB} = \frac{V_{ji}^{M}}{\exp(\hat{\beta}_0 + \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{Z}\hat{\boldsymbol{\lambda}})} = V_{ij}^{X} \cdot \exp(\hat{\xi}_{ij}) \tag{3.6}$$

where mirror imports are stripped of the coefficients estimated in equation (3.4).

The estimated CIF margin between reporter imports and mirror exports is 1.73 and the estimated CIF margin between reporter exports and mirror imports is 1.72. Conceptually, the true (unobserved) CIF margin should be 1 plus CIF plus statistical noise. The results show that the estimated margins are much larger than is commonly assumed in the literature (see Gaulier and Zignago (2010)), and suggests that the commonly assumed CIF margin of 1.1 used in some trade misinvoicing estimates is inadequate (see, e.g., UNCTAD (2016); Spanjers and Salomon (2017)).

Further, imports are not FOBized for the countries that do not report their imports to

Comtrade on the recommended CIF basis. For these countries, the FOBization procedure is not performed, and instead the reported import values are simply used.[30]

In a separate step, FOB imports are also calculated by stripping out the estimated *licit* components of the CIF margin. Recalling that in the models defined by equations (3.1) and (3.2), the true (unobserved) bilateral trade considers reporter imports as equivalent to partner exports, plus discrepancies and statistical noise. Thus, the trade discrepancies are partitioned into those originating from licit sources (e.g., reporting mistakes) and those that can be explained by illicit motivations and thus are likely to represent trade misinvoicing, e.g.: $V_{ijct}^{M} = V_{jict}^{X} + licit + illicit + u_{ijct}$. Licit predictors are included in vector $\boldsymbol{X}$ while illicit predictors are contained in $\boldsymbol{Z}$.

Reporter imports that are stripped out of the licit components of the CIF margin are called $V_{ij}^{M;FOB,nonIFF}$, where the superscript refers to the component of the trade gap that has been stripped out. Likewise, partner FOB imports cleaned from the legitimate components of the CIF rate are denoted $V_{ji}^{M;FOB,nonIFF}$.

In this calculation, import values are divided by the estimated coefficients of the variables in the vector $\boldsymbol{X}$ that contains legitimate sources of discrepancies, such as misreporting due to different calendar years or due to re-exports. This implies that the other side of the trade is exports plus illicit discrepancies plus statistical noise.

$$V_{ij}^{M;FOB,nonIFF} = \frac{V_{ij}^{M}}{\exp(\hat{\alpha}_0 + \boldsymbol{X}\hat{\boldsymbol{\alpha}})} = V_{ji}^{X} \cdot \exp(\boldsymbol{Z}\hat{\boldsymbol{\gamma}}) \cdot \exp(\hat{\epsilon}_{ij}) \tag{3.7}$$

[30]Those countries are Brazil, Cambodia, Canada, Guinea, Mali, Paraguay, South Africa, Tajikistan, Ukraine, and the USA. For more information, see https://unstats.un.org/unsd/tradereport/questform_MM.asp?qid=7.02.

$$V_{ji}^{M;FOB,nonIFF} = \frac{V_{ji}^{M}}{\exp(\hat{\beta}_0 + \boldsymbol{X}\hat{\boldsymbol{\beta}})} = V_{ij}^{X} \cdot \exp(\boldsymbol{Z}\hat{\boldsymbol{\lambda}}) \cdot \exp(\hat{\xi}_{ij}) \tag{3.8}$$

The illegitimate components of the CIF margin are estimated to be $\exp(\boldsymbol{Z}\hat{\boldsymbol{\gamma}}) = 0.98$ and $\exp(\boldsymbol{Z}\hat{\boldsymbol{\lambda}}) = 0.97$. This implies that, holding legitimate reasons for trade gaps constant, the part of the trade gap that is explained by illicit factors alone would result in import over-invoicing or export under-invoicing (an illicit outflow).

Harmonization procedure

The following step seeks to remedy the one of the main problems identified in the literature, where most existing estimates make no attempt to account for the variance in the trade declarations of countries. Therefore, the next step is designed to fulfill criterion 3 presented above. A harmonization procedure is performed to generate the best estimate of the FOB value of the trade, following the reconciliation technique developed by Gaulier and Zignago (2010). The harmonization procedure rests on the view that different countries' declarations to customs will vary in reporting quality due to country-specific idiosyncrasies (e.g., robustness of national statistical procedures, etc.). For any given trade value, there are two declarations: one from the reporting country and one from the partner country. Thus, the goal is to generate a reconciled value as a weighted average of both declarations, where weights are proportional to a country's *relative* quality of declaration.

To implement the harmonization procedure, the two regression models below are used, where the outcome variable is the reporting distance between: in (3.9), reporter imports (previously FOBized using the procedure described above) and mirror exports; and in (3.10), reporter exports and mirror imports (on a FOB basis). The models employ reporter, partner, commodity, and time fixed effects to control for country-specific, commodity-specific, and year-specific idiosyncrasies in the trade gaps.

$$\left| \ln \frac{V_{ijct}^{M;FOB}}{V_{jict}^{X}} \right| = \phi_i + \psi_j + \kappa_c + \tau_t + \varepsilon_{ijct} \tag{3.9}$$

$$\left| \ln \frac{V_{jict}^{M;FOB}}{V_{ijct}^{X}} \right| = \phi_i + \psi_j + \kappa_c + \tau_t + \varepsilon_{ijct} \tag{3.10}$$

where

- ϕ_i are reporter fixed effects;
- ψ_j are partner fixed effects;
- κ_c are commodity fixed effects;
- τ_t are year fixed effects;
- ε_{ijct} is random noise;
- and with a sum-to-zero constraint for identifiability: $\sum_{i=1}^{I} \phi_i + \sum_{j=1}^{J} \psi_j + \sum_{c=1}^{C} \kappa_c + \sum_{t=1}^{T} \tau_t = 0$.

The fixed effects of interest are ϕ and ψ which reflect the accuracy of each transacting country's reports to Comtrade. The commodity and year fixed effects isolate the source of discrepancies that are independent of the quality of country declarations, e.g., a product code that is more prone to reporting mistakes because the merchandise is homogeneous and hard to distinguish (Gaulier and Zignago, 2010). Therefore, this means that the report and partner fixed effects are "cleaned" from the effects of any trade specialization in certain sectors (Gaulier and Zignago, 2010). Therefore, the estimated fixed effects $\hat{\phi}$ and $\hat{\psi}$ represent the marginal effect of a country's specific reporting practices on the trade gap, holding the quality of their partner's declaration constant and independent of any commodity or year-specific reasons for the gap between the mirror declarations.

Weights are computed in order to minimize the variance of the reconciled value, following the procedure originated by Gaulier and Zignago (2010).

As in Gaulier and Zignago (2010), the variance in reporter quality of declaration is computed as:

$$\sigma_i = \frac{\pi}{2} \cdot \left(\hat{\phi}_i - \min(\hat{\boldsymbol{\phi}}) - 2 \cdot SE(\hat{\phi}_i) \right) \tag{3.11}$$

and for the partner quality of declaration as:

$$\sigma_j = \frac{\pi}{2} \cdot \left(\hat{\psi}_j - \min(\hat{\boldsymbol{\psi}}) - 2 \cdot SE(\hat{\psi}_j) \right) \tag{3.12}$$

where $\hat{\phi}_i$ and $\hat{\psi}_j$ are the estimated least-square means of country-specific discrepancies for the ith reporter and the jth partner, respectively; and $SE(\hat{\phi}_i)$ and $SE(\hat{\psi}_j)$ are the corresponding standard errors of those fixed effect coefficients.

Next, the weight to give to the reporter i's declaration as opposed to the partner j's declaration is computed as:

$$\delta = \frac{e^{\sigma_j^2} \cdot \left(e^{\sigma_j^2} - 1 \right)}{e^{\sigma_i^2} \cdot \left(e^{\sigma_i^2} - 1 \right) + e^{\sigma_j^2} \cdot \left(e^{\sigma_j^2} - 1 \right)} \tag{3.13}$$

The next step is to compute the reconciled value, which represents the most precise estimate of the value of the trade by taking into account the quality and accuracy of each country's declaration. The reconciled value RV^M represents the best estimate of the import declaration.

$$RV^M = \delta \cdot V_{ij}^{M;FOB} + (1 - \delta) \cdot V_{ji}^X \tag{3.14}$$

The reconciled value RV^X represents the best estimate of the export declaration.

$$RV^X = \delta \cdot V_{ij}^X + (1 - \delta) \cdot V_{ji}^{M;FOB} \tag{3.15}$$

Note that in equation (3.14) the reporter i declares import transactions while in equation (3.15) it declares the value of its exports, and that the weight δ represents the relative precision of i's declaration compared to its partner j's. Therefore, instead of assuming that, e.g., declarations by developed countries are more trustworthy than declarations by poor countries, the relative accuracy of each country's declaration is determined empirically.

Computing the illicit flow embedded in each transaction

The final step is to compute the dollar value of trade misinvoicing contained in both imports and exports for the "atlas" reporter i.

The import discrepancy for country i is the difference between FOB imports stripped of licit trade discrepancies (so all that remains is the illicit gap plus statistical noise) as calculated in (3.7) and the reconciled value that represents the best estimate of reporter FOB imports controlling for the reporting quality of countries calculated in (3.14). This strategy represents the combination of a "residual" and "reconciliation" approach, and is one of the main innovations of the "atlas" method, as discussed in section 3.3.3.

Specifically, recall that $V_{ij}^{M;FOB,nonIFF}$ denotes imports that have been stripped of the estimated margin that can be explained by non-illicit factors alone, and is equal to $V_{ji}^{X} \cdot \exp(\boldsymbol{Z}\hat{\boldsymbol{\gamma}}) \cdot \exp(\hat{\epsilon}_{ij})$ as shown in equation (3.7). This follows from the data-generating model for a trade transaction discussed in section 3.4.2 – which is based on the macroeconomic identity that the true value of imports by i from j is equal to the true value of exports by j to i. Of course, the true value of the trade is unknown, and so the "atlas" method provides a model of the trade *declarations* where declarations by i are on one side of the equality, and the corresponding mirror declarations by j plus discrepancies and statistical noise are on the other side of the equality. By stripping import declarations of the discrepancies that can be explained by licit or benign factors, what remains on the other side of the equality are mirror exports, discrepancies that can be explained by determinants of illicitness, and the unexplained discrepancies (the residual). In other words, what remains on the other side of the transaction is the misinvoiced mirror export declaration (plus noise) – this is the nature of the "residual" approach.

Then, the reconciled value RV^M is the one that harmonizes declarations from both reporter and partner according to relative precision, and thus represents the best estimate of the "true" declaration (which lies somewhere between what the reporter declared and what the partner declared) – this is the nature of the "harmonization" or ("reconciliation") strategy of the "atlas".

Therefore, by subtracting the best guess of the true import declaration from the misinvoiced mirror export declaration (plus noise), what remains is the dollar amount of misinvoicing in i's imports from j.

$$IFF_{ij}^{M} = V_{ij}^{M;FOB,nonIFF} - RV^{M} \tag{3.16}$$

Positive values of IFF_{ij}^{M} correspond to import over-invoicing, i.e., an illicit outflow from i to j.

Using the same reasoning, the export discrepancy for country i is the difference between the reconciled value that represents the best estimate of reporter exports as calculated in (3.15) and the observed exports actually reported by i.

$$IFF_{ij}^{X} = RV^{X} - V_{ij}^{X} \tag{3.17}$$

Positive values of IFF_{ij}^{X} correspond to export under-invoicing by i and represent an illicit outflow from i to j.

Total trade misinvoicing for a reporter i trading with partner j for commodity c at time t is the sum of the import discrepancy and of the export discrepancy.

A summary of the step-by-step procedures is provided in Figure 3.2 below.

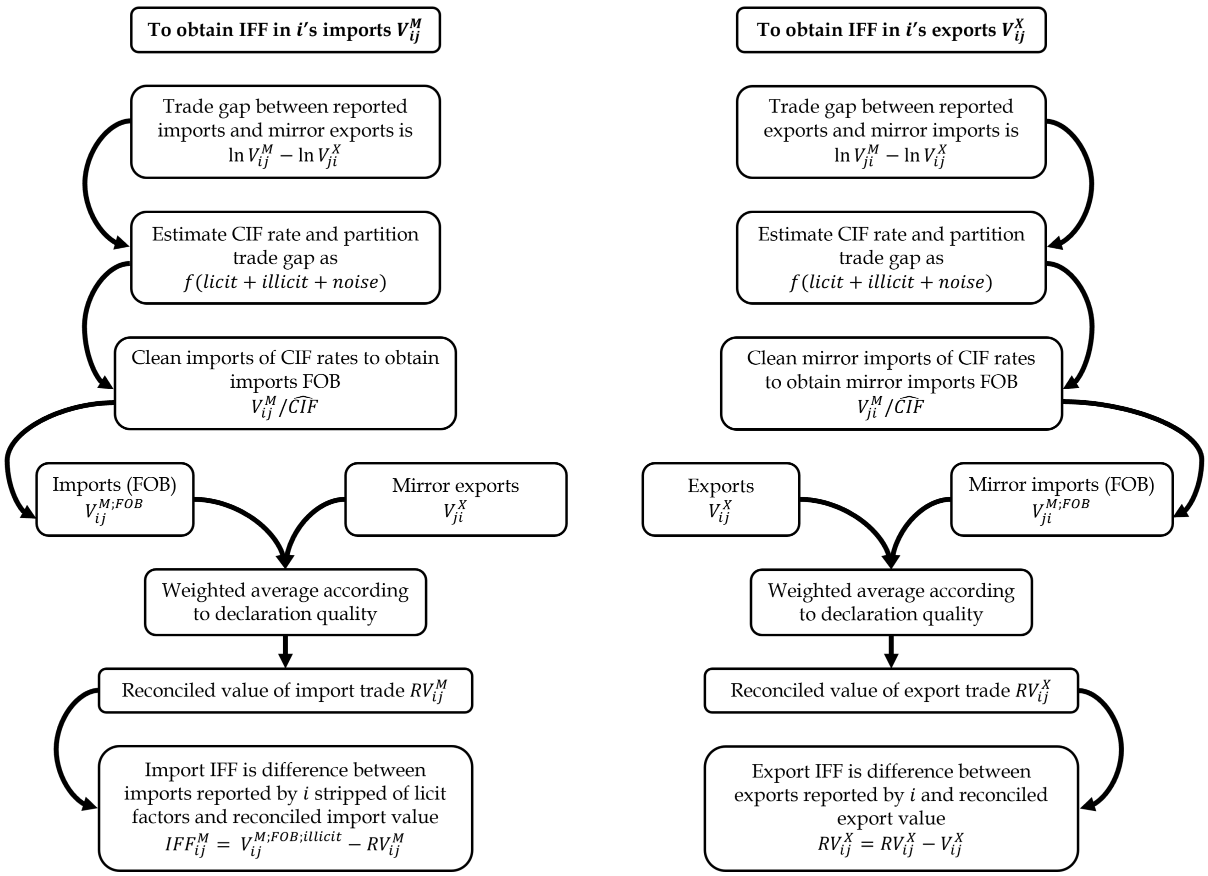


Figure 3.2: Main steps of the methodology to generate the "atlas of misinvoicing" estimates.

3.4.4 Aggregation strategy

The prior section provided the detailed steps to arrive at an estimate of the illicit flow embedded in a particular transaction between a reporter i and a partner j for a commodity c in year t. Illicit flows are then aggregated up to have an estimate of illicit flows for a particular country i. Broadly, aggregate IFFs can be presented on a "net" or on a "gross excluding reversals" (GER) basis (Salomon, 2019). The question of what technique to use to aggregate IFFs is more difficult than it seems, and has been the subject of vigorous disagreements by authors (see, e.g., (Nitsch, 2016; Spanjers and Salomon, 2017)).

Illicit flows presented on a net basis simply add up inflows (a negative value) and outflows (a positive value). Thus, positive and negative values will cancel out to yield a smaller number of aggregate IFFs for country i.

However, as argued by GFI, there is no such thing as "net crime" (Cobham and Janský, 2020), and so it makes sense to consider gross flows. Illicit outflows presented on a GER basis ignore all inflows (i.e., negative values) and simply add up all the positive outflows across trading partners. Analogously, illicit inflows on a GER basis are calculated by summing only negative values across partners (i.e., ignoring outflows). As this paper has argued, illicit inflows are also prejudicial to development since they are untaxed and invisible to governments. Illicit inflows can exacerbate resource curse issues and can be used to finance illegal activities such as drug trafficking and terrorism. Therefore, estimates of illicit inflows from trade misinvoicing should also be a quantity of interest.

It is important to note that for a given country pair i and j in a given year t, the same trade flow can be associated with either an inflow or an outflow according to what commodity is traded. While it might seem unlikely that illicit funds might be traveling in both directions for the same trade flow, there could be a variety of different actors doing this for different reasons. For example, country i might have export taxes on raw materials and export subsidies for manufacturing output, which would give an incentive to under-invoice exports of raw materials (resulting in an illicit outflow) and to over-invoice exports of manufactured goods (resulting in an illicit inflow). Alternatively, a criminal syndicate that has a legitimate front company may use re-invoicing to send money to an affiliate in another country to make an investment (e.g., hiring "muscle" to fight off a competitor) and then bring funds back using exports to the same country when the investment bears fruit.

Therefore, an aggregation strategy that nets out the illicit inflows and outflows might risk under-estimating the extent to which illicit activity occurs within a trade flow for the same country pair. Conversely, if illicit flows are presented on a GER basis, this should not be equated to funds departing a country, since inflows would not be included in the calculation. In addition, contrary to the GFI estimates (Spanjers and Salomon, 2017; Salomon, 2019), GER inflows and GER outflows are not summed, recognizing the critique by Nitsch (2012) that such an aggregate figure is so hard to interpret that it is devoid of any substantive meaning.

The object of analytical inquiry should guide the choice of aggregation strategy. For example, stakeholders interested in getting a picture of the total amount of funds departing a country on balance should favor a net aggregation basis. By contrast, stakeholders interested in better understanding the drivers and mechanisms of IFFs should favor aggregation using GER to identify where money is flowing in or out. In that way, IFFs presented on a GER basis can aid in tailoring policy responses across jurisdictions and sectors.

The "atlas" database provides aggregated results using both aggregation strategies. Since positive values represent illicit outflows and negative values represent illicit inflows, to calculate gross outflows on a GER basis, the positive values across j are summed for each reporter i.

$$IFF_{it}^{gross;out} = \sum_{j;IFF>0} IFF_{ijt}^{M} + \sum_{j;IFF>0} IFF_{ijt}^{X} \tag{3.18}$$

To calculate gross inflows using GER, negative IFF values are added up over partners j:

$$IFF_{it}^{gross;in} = \sum_{j;IFF<0} IFF_{ijt}^{M} + \sum_{j;IFF<0} IFF_{ijt}^{X} \tag{3.19}$$

Net aggregation is a simple sum of all IFF values for i over j:

$$IFF_{it}^{net} = \sum_{j} IFF_{ijt}^{M} + \sum_{j} IFF_{ijt}^{X} \tag{3.20}$$

Prior to summing across partners for each reporter i, for both methods of aggregation, the IFF value is summed across commodities c first.

3.5 Findings

This section synthesizes key insights from the "atlas of misinvoicing" and provides examples of how the dataset can be used by interested stakeholders. Policy-makers can use these results to understand the scale of the problem in their jurisdiction, in addition to the major destinations and sectors where misinvoiced trade flows to. Results are reported as a dollar value of trade misinvoicing, as a percentage of GDP, and as a percentage of trade. The research question at hand should guide the choice of variable to represent trade misinvoicing as an explanatory variable. In many cases, a scaled value of trade misinvoicing (e.g., as a percentage of GDP) will be more appropriate than a dollar value.

This section proceeds as follows. First, global results are presented in order to glean a high-level understanding of the problem. Subsequently, the analysis zooms in to various country groups in order to demonstrate the potential of this dataset in understanding the sources, sinks, and sectors that are responsible for most trade misinvoicing.

3.5.1 Global results

The "atlas of trade misinvoicing" provides results for most countries in the world (167). To my knowledge, there is no publicly available dataset of misinvoicing estimates that has such broad country coverage.

Globally, the top 3 countries with the highest average annual gross outflows during the period 2000-2018 were the United States ($221 billion), Canada ($65 billion), and China ($59 billion). The magnitude of trade misinvoicing in the USA is much larger than trade misinvoicing in other countries. However, the USA had a GDP of $21 trillion in 2018 and its total trade (calculated as the sum of reported imports and exports) in 2018 amounted to $4.28 trillion. Reporting trade misinvoicing on a dollar basis may yield results that emphasize open economies with large volumes of trade, since in those countries there is more trade that can be misinvoiced.[31]

Therefore, trade misinvoicing estimates are presented as a percentage of countries' trade (the sum of their reported imports and exports), as displayed in Figure 3.3. Africa and Latin America tend to have higher trade misinvoicing as a percentage of trade, compared to Europe which has the least. The figure also highlights the extent to which trade is misinvoiced in Africa. Further analysis on Africa is undertaken in the following section.

The top 10 countries with the highest average gross outflows as a percentage of trade during 2000-2018 were Yemen (58%), Congo (55%), Tanzania (41%), Cambodia (23%), Côte d'Ivoire (20%), Trinidad and Tobago (20%), South Africa (18%), Angola (16%), Costa Rica (16%), and Azerbaijan (16%). The fact that the majority of those countries are developing should be a cause for alarm for policy-makers focused on poverty allevia-

[31] This assumes that there is a limit to the extent that a given shipment can be misinvoiced. This may be correct – truly outrageous degrees of misinvoicing for a given shipment may risk detection by customs authorities.

tion and sustainable development. It should be noted that though it may be a factor, the methodology adjusts for the poor quality of data reporting practices in countries through a variety of methods (e.g., censoring the dataset to observations where the observed trade gap is less than 100, removing statistical outliers, performing the reconciliation procedure that downweighs poor quality reports, etc.). Robustness checks were performed to verify that the threshold used for removing outliers did not significantly change the main results.

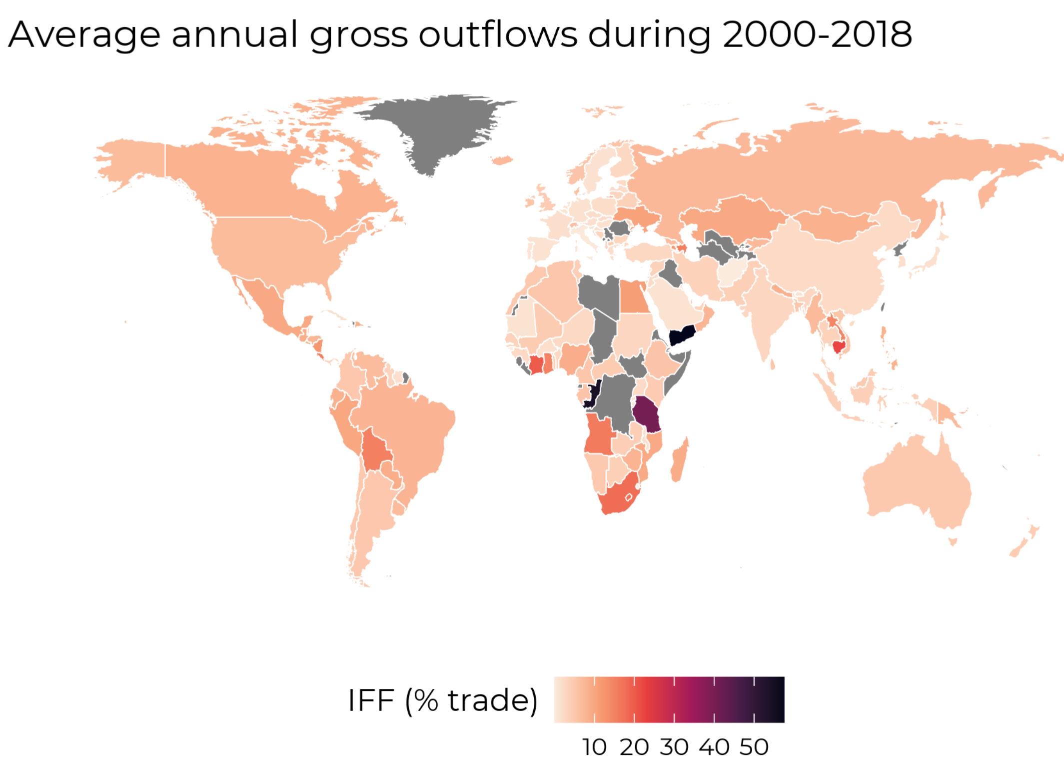


Figure 3.3: Average yearly gross outflows as a percentage of trade.

The deleterious impact of trade misinvoicing on domestic resource mobilization can best be understood by examining results as a percentage of countries' GDPs. Africa, eastern Europe, and central Asia experienced the highest average gross outflows as a percentage of GDP during 2000-2018. The top 10 countries on that basis were Congo (54%), Yemen (37%), Cambodia (21%), Trinidad and Tobago (19%), Tanzania (17%), Hong Kong (15%), Angola (15%), Côte d'Ivoire (14%), Singapore (13%) and Costa Rica (11%).

The dataset also permits identification of the greatest "sinks" for illicit flows, that is, countries which have the highest gross inflows (either through import under-invoicing or export over-invoicing). There is a negative and statistically significant (Spearman's $\rho = -0.58$; p-value < 0.01) correlation between a country's rank on the Financial Secrecy Index (FSI, Tax Justice Network (2018)) – where the top rank corresponds to the most financial secrecy – and the amount of illicit inflows that it receives. The FSI ranks countries on various dimensions of financial secrecy and according to the scale of their offshore activities. Indicators of financial secrecy used in the index include the degree of information around the beneficial owner of an asset, the degree of transparency on legal entities and the extent to which it is available to the public, the integrity of tax and financial regulation, and finally how cooperative countries are with regards to international standards for financial disclosure. The top 5 countries on the 2018 edition of the FSI are, in descending order: Switzerland, the United States, Cayman Islands, Hong Kong, and Singapore. Figure 3.4 shows that the top 3 countries with the highest average gross inflows in the period 2000-2018 are among the highest ranked on the FSI. The Netherlands and Russia are number 14 and 29, respectively, on the FSI.

This reaffirms that financial secrecy is a scourge and that efforts to increase global financial integrity are a vital component of achieving the SDGs and building a global architecture that is supportive of sustainable development. Recognizing this priority, in 2019 the UN General Assembly assembled the panel on Financial Accountability Transparency and Integrity (FACTI).

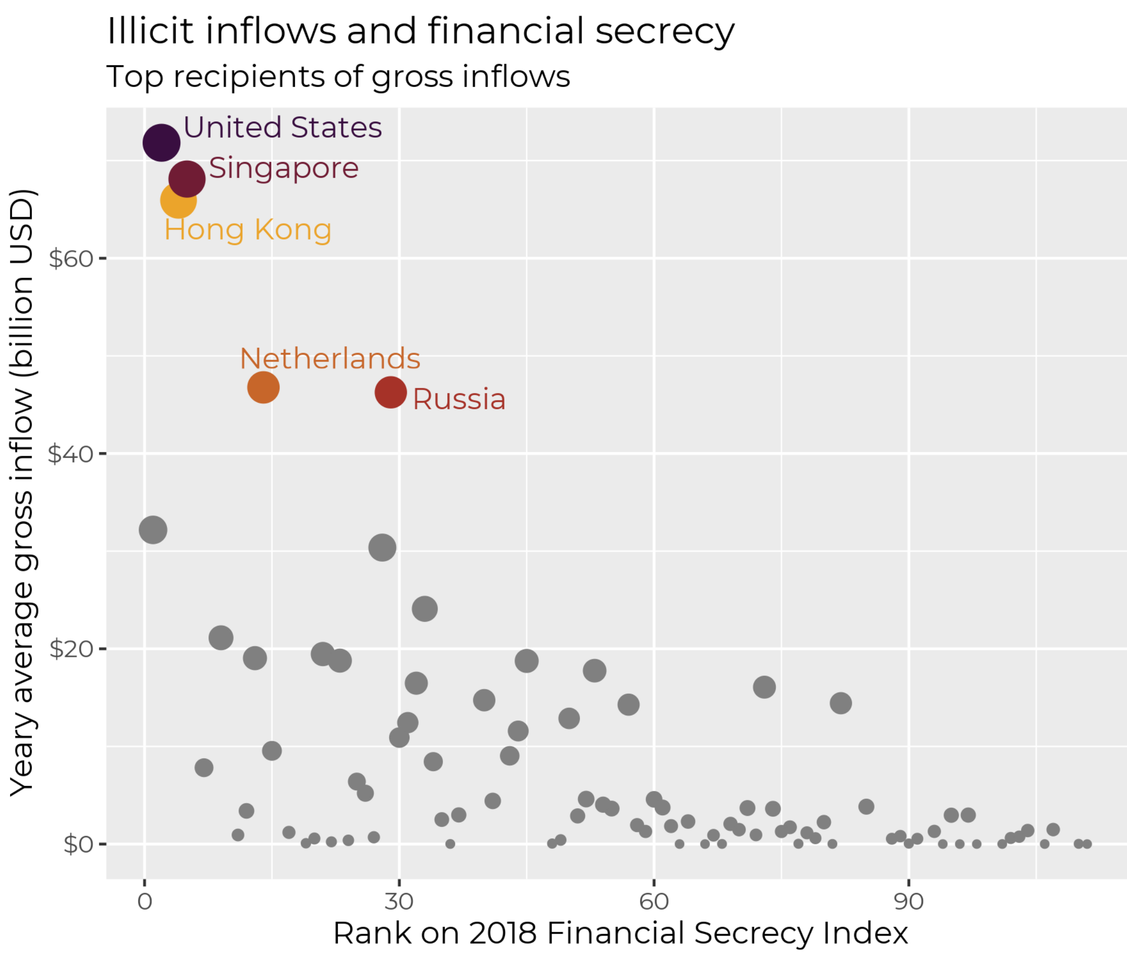


Figure 3.4: Association between financial secrecy and receipt of illicit inflows.

The underlying reasons for trade misinvoicing will vary by sector. Some sectors, notably natural resources, are more susceptible to misinvoicing that is used to finance conflict and to embezzle money from the state (Vézina, 2015; UNECA, 2017; Andreas, 2015). In other sectors, misinvoicing will primarily be explained by abuses of transfer pricing by multinational companies in order to book profits in lower-tax jurisdictions (UNECA, 2018a; Davies et al., 2018; UNECA, 2019; Tørsløv et al., 2018; UNECA, 2018b). This is likely to be the case in oligopolistic markets that are dominated by a few large multinational conglomerates, such as pharmaceutical products for example.

The Sankey diagrams in figures 3.5 and 3.6 provide an example of the sectoral breakdown of illicit financial flows. In each sector, the top 5 countries (by % of GDP) with the highest average gross yearly outflows during 2000-2018 are displayed on the left axis. The respective destinations of those illicit outflows are depicted on the right axis, with the width of segments proportional to the dollar value of the illicit flow.

This is an example of how this atlas of illicit financial flows can be used to study the sinks and sources for each of the 99 sectors in the Harmonized System. The potential for discovery of additional insights is large and will be a matter for future research.

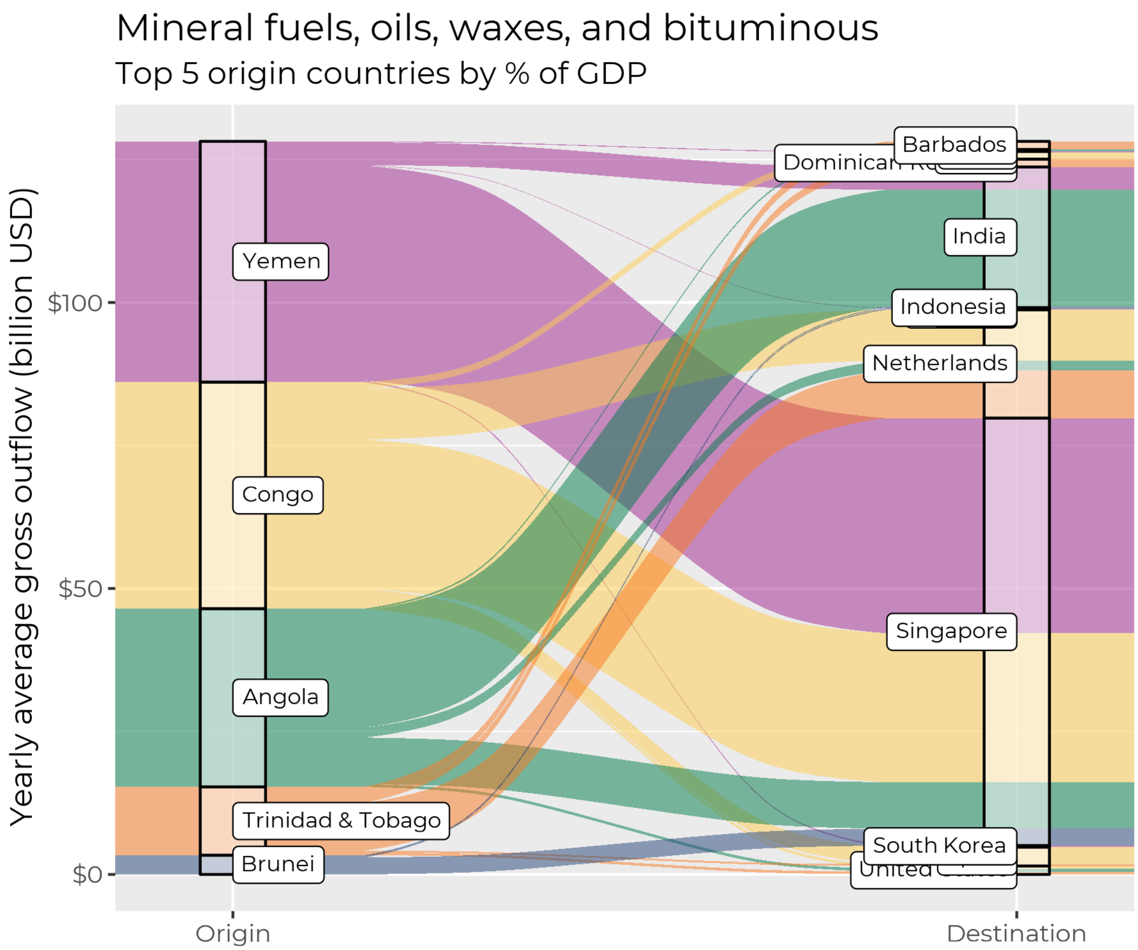


Figure 3.5: Destination and magnitude of flows originating from the top 5 countries in mineral products.

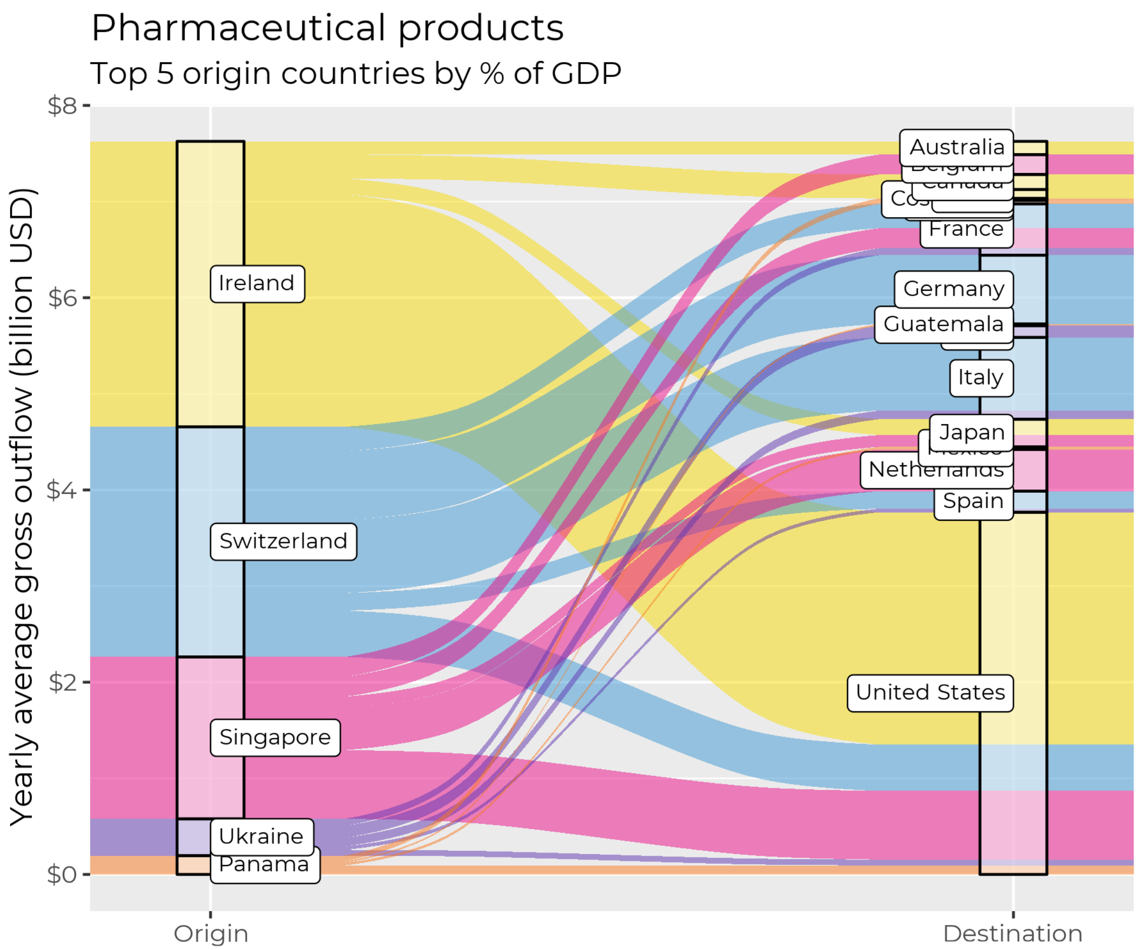


Figure 3.6: Destination and magnitude of flows originating from the top 5 countries in pharmaceutical products.

3.5.2 Results for Africa

Given that African countries feature prominently in the top conduit countries for illicit outflows (both as a percentage of GDP and of trade), this section turns to analyzing the extent and patterns of misinvoicing in Africa.

Figures 3.7 and 3.8 display the yearly evolution of gross and net financial outflows from the continent, both as a percentage of GDP and as a percentage of trade. Africa had net illicit inflows in the early 2000s but has experienced illicit outflows in the latter half of the 2010s. This suggests that gross illicit inflows are a large component of trade misinvoicing. As discussed earlier, those inflows are untaxed, invisible to governments, and can be used to strengthen corrupt elites and finance organized crime and terrorism. The magnitude of misinvoiced trade in the continent is around 10% which is broadly consistent with, though more conservative than, findings from Global Financial Integrity who estimate that the percentage of Sub-Saharan Africa's trade with advanced economies that was misinvoiced during 2006-2015 was on average 17.4% for gross outflows and 15.2% for gross inflows (Salomon, 2019, p. 2).

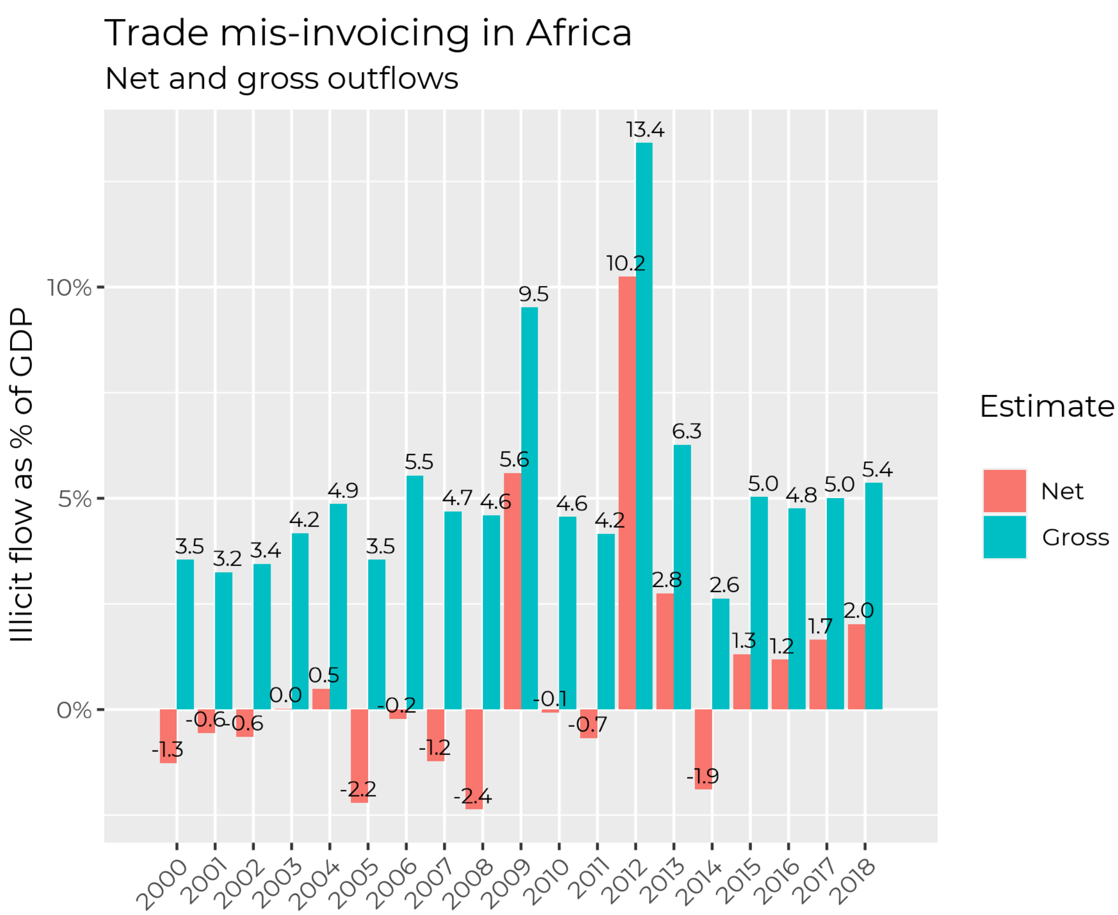


Figure 3.7: Net and gross outflows in Africa as a percentage of GDP.

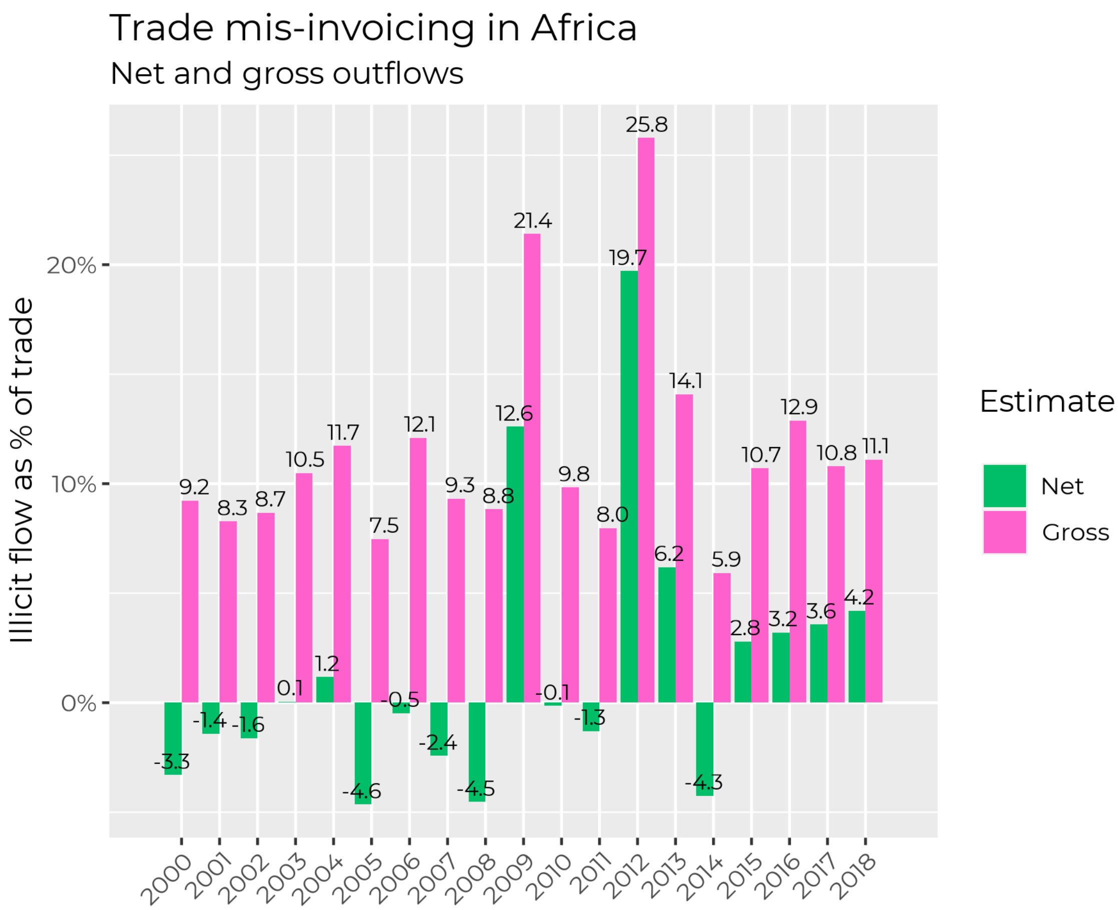


Figure 3.8: Net and gross outflows in Africa as a percentage of trade.

Next, the "atlas" dataset provides a sectoral breakdown of illicit flows on the continent, as illustrated in Figure 3.9. The overwhelming amount of gross outflows occurs in the natural resource sector. The extent to which natural resources can contribute to a resource curse, enable conflict, and hamper development has been well-documented and debated (see, e.g., Dunning (2008); Ross (2015)). Conventional accounts of the resource curse hold that windfall profits from natural resources can cause Dutch disease through an appreciation of the real exchange rate and can entrench the power of unaccountable elites (Ross, 1999; Oliver et al., 2017). These results provide additional insights on the resource curse by suggesting that windfall profits are not the only mechanism of harm, and that illicit outflows through trade misinvoicing will exacerbate capital flight and deplete governments' fiscal reserves.

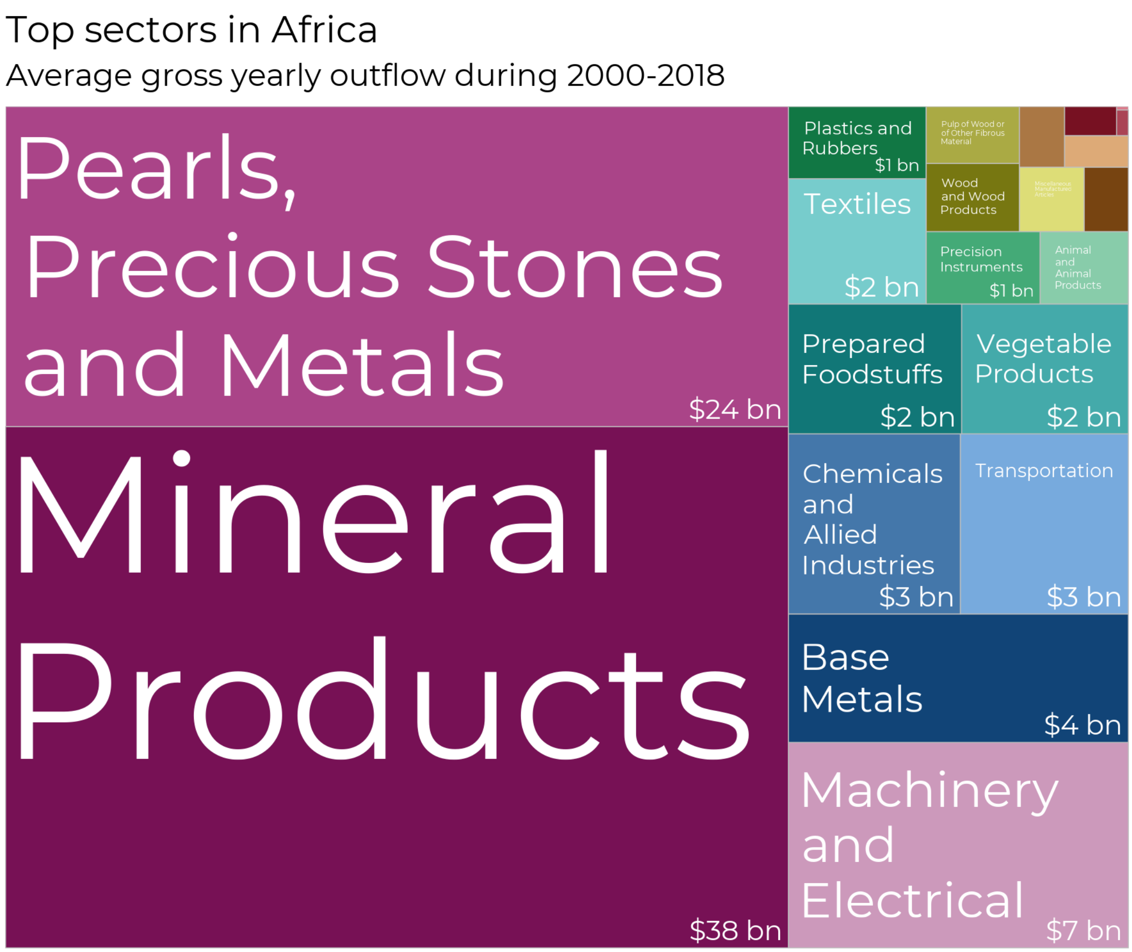


Figure 3.9: Top sectors in Africa for outflows during 2000-2018.

The data also reveal that mineral products are the main sources of misinvoicing. This is in line with the High Level Panel on Illicit Financial Flows from Africa (2015) which found that oil, precious metals, and minerals were the leading source of trade misinvoicing (via re-invoicing) from Africa from 2000 to 2010, followed by ores and electrical machinery and equipment. ESCWA (2018) excludes the main sector used here (Harmonized System classification 27, which includes mineral products) from its sectoral disaggregation but finds that machinery and electrical machinery are the main sources of illicit financial flows in the Arab region.

The "atlas" estimates that the natural resources sector is by far the most misinvoiced across the continent, yielding gross outflows of $62 billion annually on average. This is particularly consequential given that 46 out of the 54 countries on the continent are classified as highly dependent on the export of primary commodities (UNCTAD, 2020). Moreover, the extractives industry is characterized by a high degree of market concentration due to the capital-intensive activities involved in the large-scale extraction of minerals and other natural resources, and as such the market is dominated by Multinational Enterprises (MNE) who yield a considerable amount of influence over African governments. MNEs have the technical expertise to circumvent domestic laws, have the leverage to negotiate tax regimes that are advantageous to them but erode the tax base of national governments, and possess the market power to manipulate prices and other costs along the commodity value chain (UNCTAD, 2016, 2020; UNECA, 2017, 2019).

It is useful to examine the sources and sinks of illicit outflows in the top two sectors: mineral products[32] in Figure 3.10 and pearls, precious stones and metals[33] in Figure 3.11. The figures display the top 5 destinations of illicit outflows for the top 5 African countries in each sector (as a percentage of GDP). The top origin countries in the sector of pearls, precious stones and metals include large diamond producers such as Botswana and South Africa. Though Botswana is often heralded as a country that has managed to avoid the

[32] This sector includes HS chapters 25, 26, and 27 which correspond to "Salt; sulphur; earths and stone; plastering materials, lime and cement", "Ores, slag and ash" and "Mineral fuels, mineral oils and products of their distillation; bituminous substances; mineral waxes", respectively.

[33] This corresponds to HS chapter 71. The full description is "Natural or cultured pearls, precious or semi-precious stones, precious metals, metals clad with precious metal and articles thereof; imitation, jewellery; coin".

resource curse by entrusting the revenues to a sovereign wealth fund (Iimi, 2007; Sarraf and Jiwanji, 2001), the results suggest that revenues still escape the government through trade misinvoicing. These data can thus contribute to the evidence base for initiatives that aim to strengthen governance in the natural resource sector (UNECA, 2017).

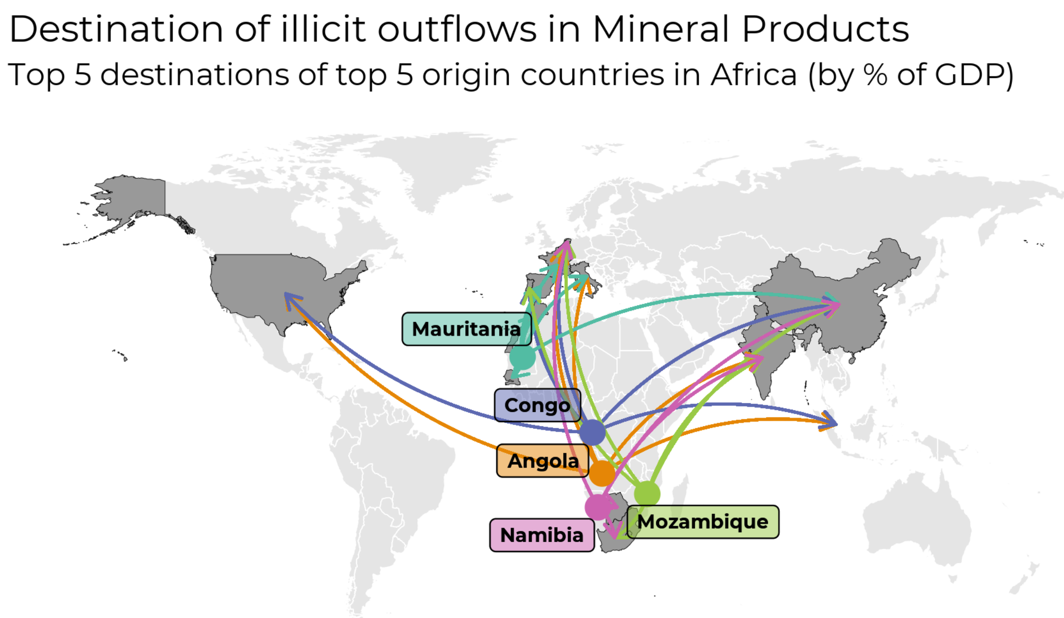


Figure 3.10: Destination of outflows in mineral products (highest sector).

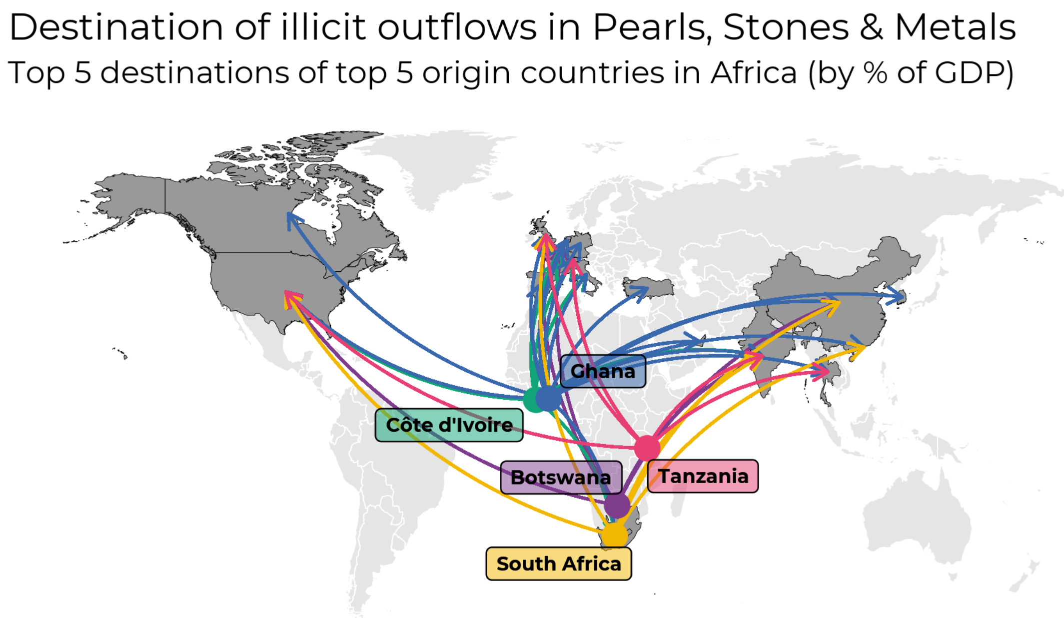


Figure 3.11: Destination of outflows in pearls, stones and precious metals (second highest sector).

3.5.3 Results for low and lower-middle income countries

Since IFFs pose significant challenges to the financing of development in poor countries, this section presents results for the 19 low income and 44 lower-middle income countries in the "atlas" dataset (classified according to the latest World Bank classification in July 2020). Low income countries are defined as those with a GNI per capita of $1,035 or less in 2019, and lower-middle income countries are those with a GNI per capita between $1,036 and $4,045.

Figure 3.12 presents yearly misinvoicing for low and lower-middle income countries in terms of gross outflows, gross inflows, and net flows, and further breaks down gross flows by transaction type. Negative values represent illicit inflows. Illicit outflows (inflows) occur through import over-invoicing (under-invoicing) and export under-invoicing (over-invoicing). The fact that net flows are much smaller can be explained by the fact that the LMIC group represents a large set of countries and that these include large sinks such as India, the Philippines and Nigeria. Net flows tend to be negative (indicating net illicit inflows to the group as a whole) in most years, except for large spikes in net outflows in 2009 and 2012.

The amount of misinvoicing in imports is slightly larger than the misinvoicing in exports. This might be due to the fact that misinvoicers have greater control in falsifying import invoices than export invoices.

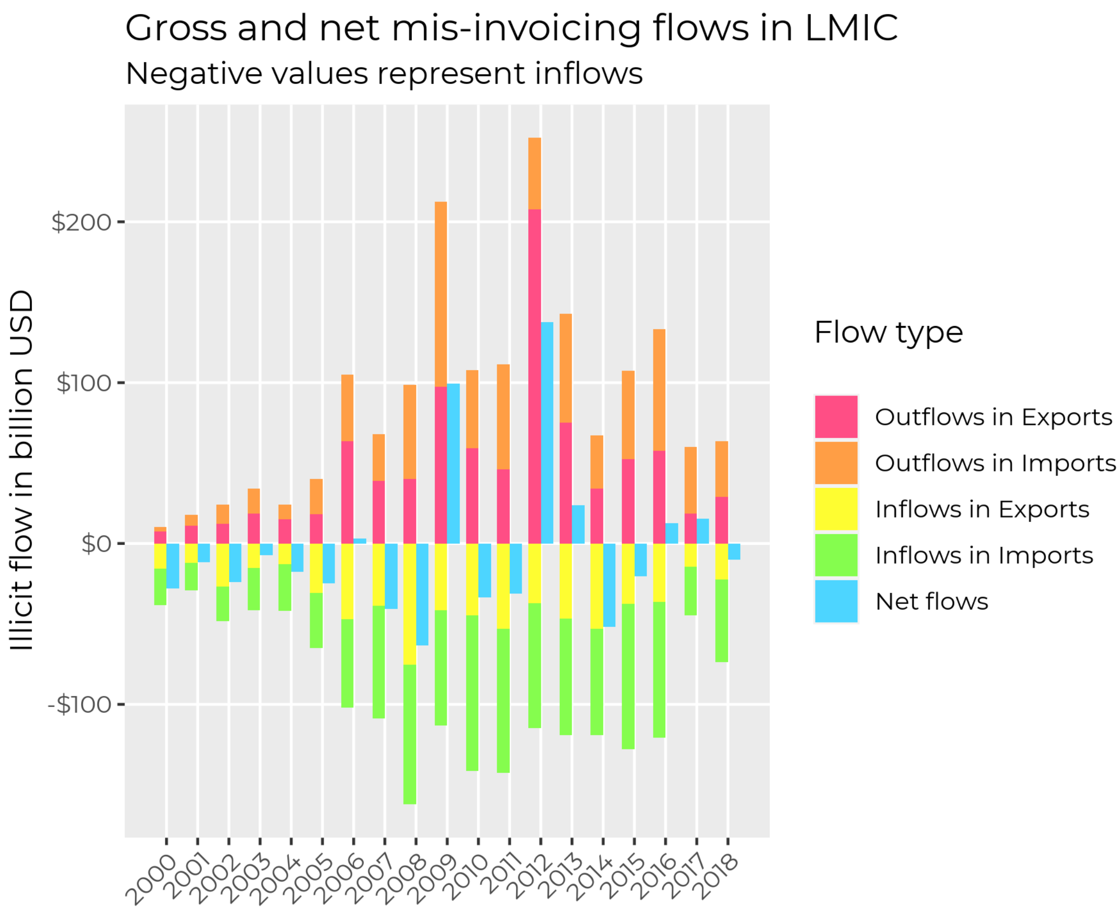


Figure 3.12: Breakdown of illicit outflows and inflows by transaction type.

The large discrepancy between gross and net flows is an interesting finding. This suggests that there are significant flows between low and lower-middle income countries (which would tend to increase gross outflows, but not net outflows), or that certain countries experience both substantial inflows and outflows, or that there is substantial misreporting of the name of the partner country or commodity, which would tend to increase gross outflows but not net outflows (as long as a shipment is recorded in trade data, incorrect reporting of partner country or commodity would lead to an apparent illicit outflow towards the true partner country and an inflow from the incorrect partner country of equal size, which could cancel out in aggregate net national statistics).[34]

Figures 3.13 and 3.14 display the top sources and sinks for illicit flows (as a percentage of GDP) in low and lower-middle income countries.

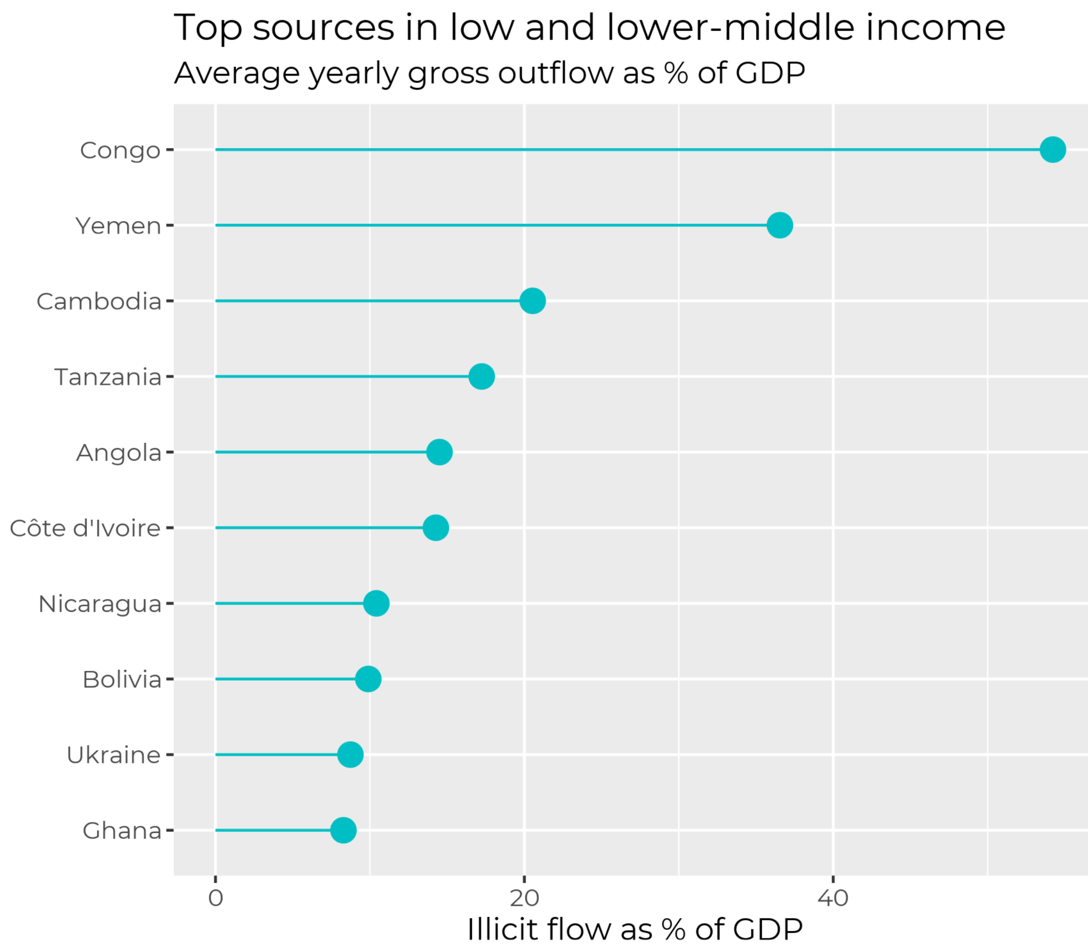


Figure 3.13: Top 10 sources of illicit outflows by percentage of GDP.

[34]As noted above, it is unlikely that misreporting of commodity codes would have a significant impact on the estimates, since the "atlas" uses data at the 2-digit level, and while customs officers may be confused about the specific commodity code that a product falls under, this would seem unlikely to occur with the broad categories used at 2-digit level.

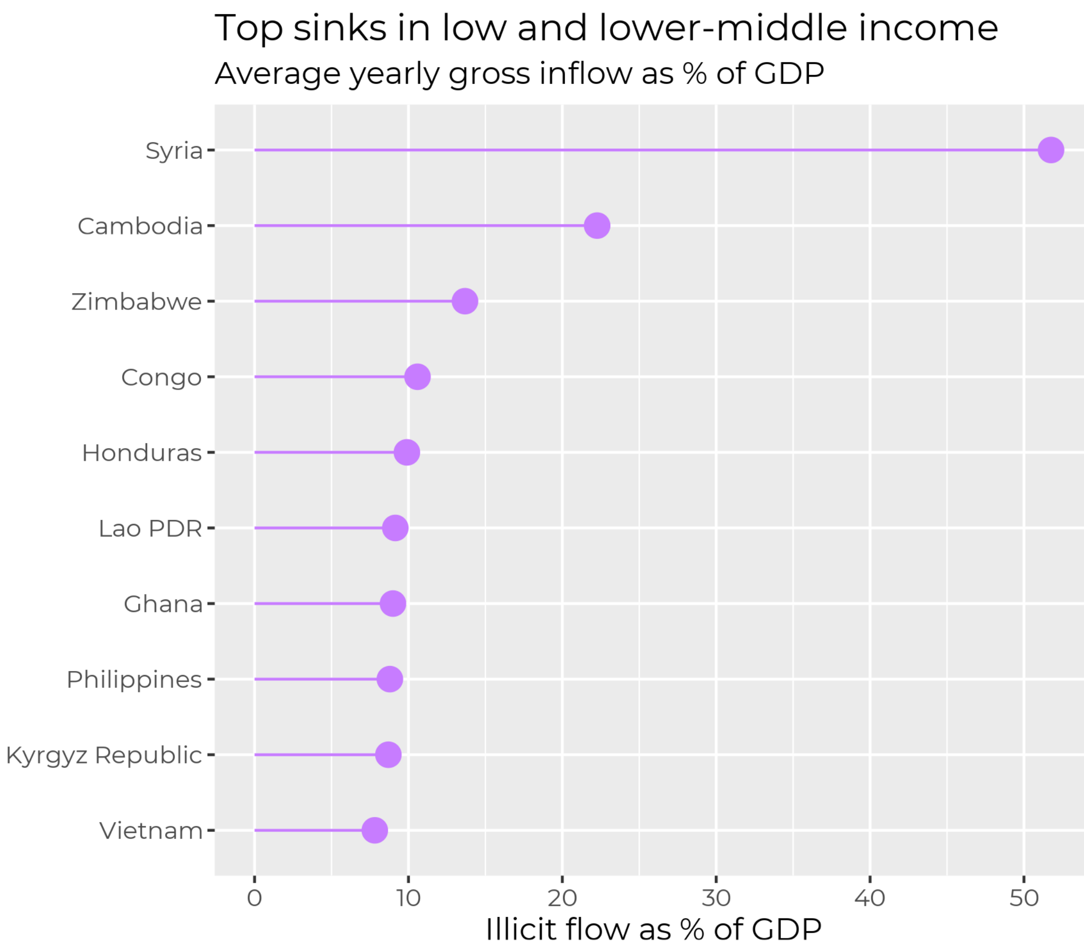


Figure 3.14: Top 10 sinks of illicit inflows by percentage of GDP.

Figure 3.15 provides the sectoral breakdown for top source countries using the Standard International Trade Classification (SITC) sector. While mineral products account for a large part of outflows in low and lower-middle income countries, there is also a large amount of misinvoicing in manufactured goods.

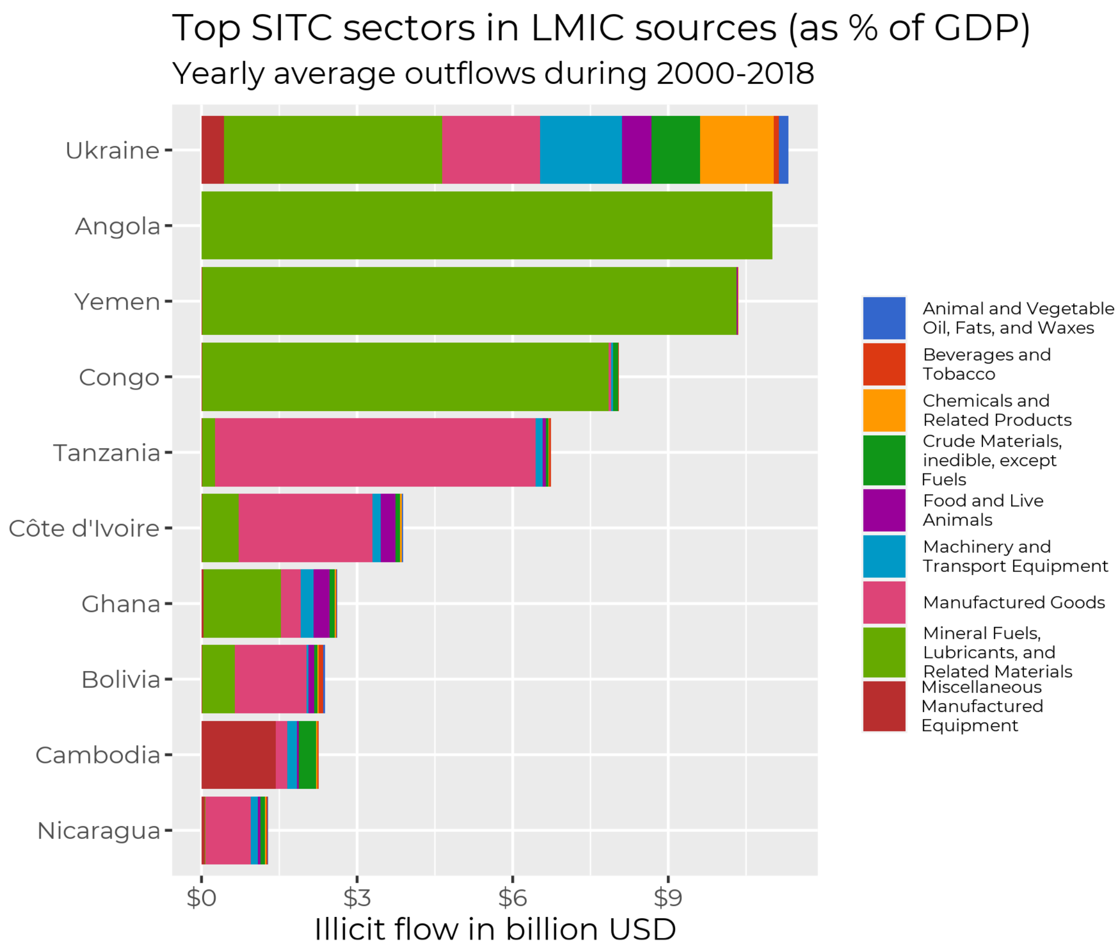


Figure 3.15: Sectoral breakdown of outflows in top 10 countries (as % of GDP).

The distributional implications of trade misinvoicing are also important to consider. According to Figure 3.16, most of the outflows from low and lower-middle income countries accrue to rich countries that have a GNI per capita greater than $30,000. Furthermore, within the lower tranche of the LMIC classification (below $2,000), outflows tend to go to comparatively poorer countries than outflows from the higher tranche of LMIC countries (above $2,000).

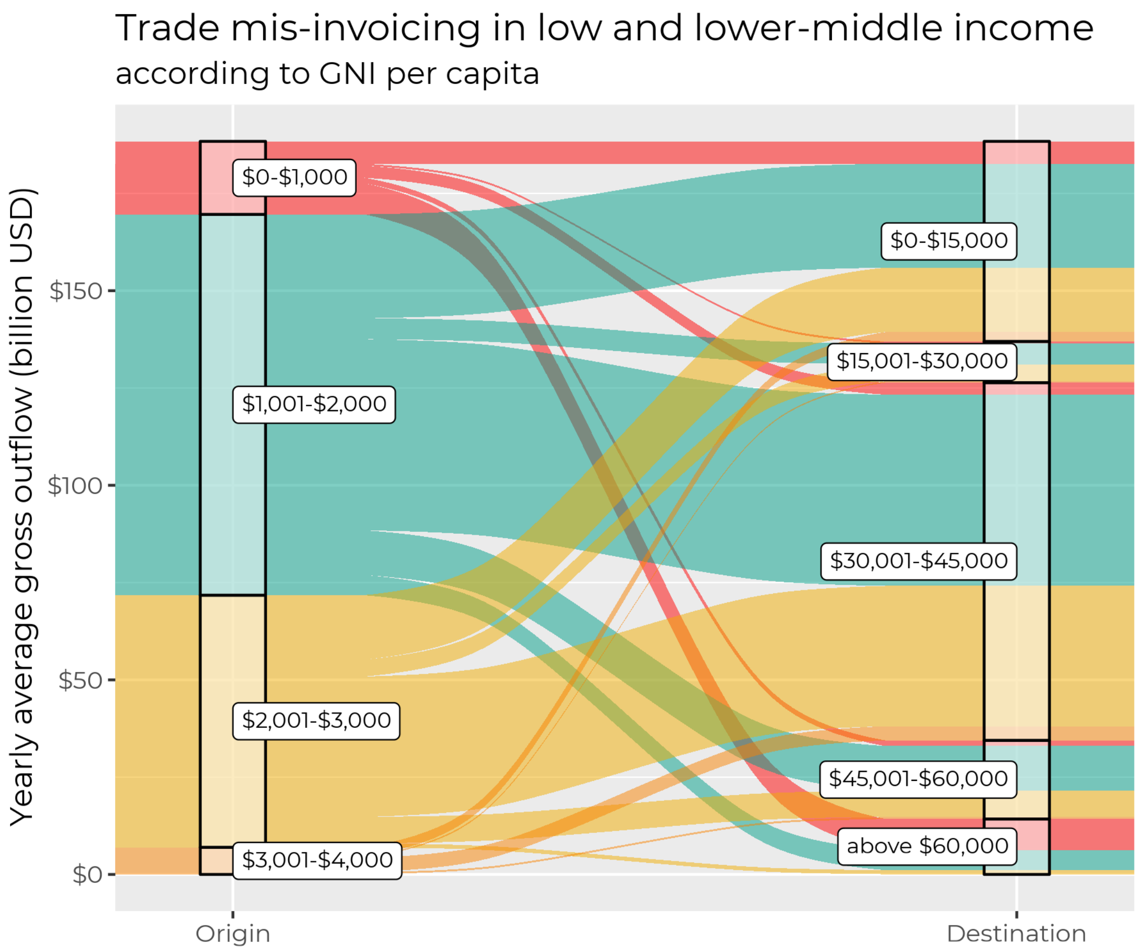


Figure 3.16: Breakdown of outflows from low and lower-middle income countries by GNI per capita.

Finally, Figure 3.17 displays the top destinations of outflows from low and lower-middle income countries. This is a mixed group which includes countries that are trading hubs, emerging economies, those that have a high degree of financial secrecy, and those that have a high presence of multinational companies. Countries that have many multinational corporations may be a significant destination for illicit financial flows that represent repatriated profits. As noted earlier, multinational corporations frequently use trade misinvoicing to transfer finance between parts of their multinational group located in different countries in order to evade fiscal and regulatory constraints.

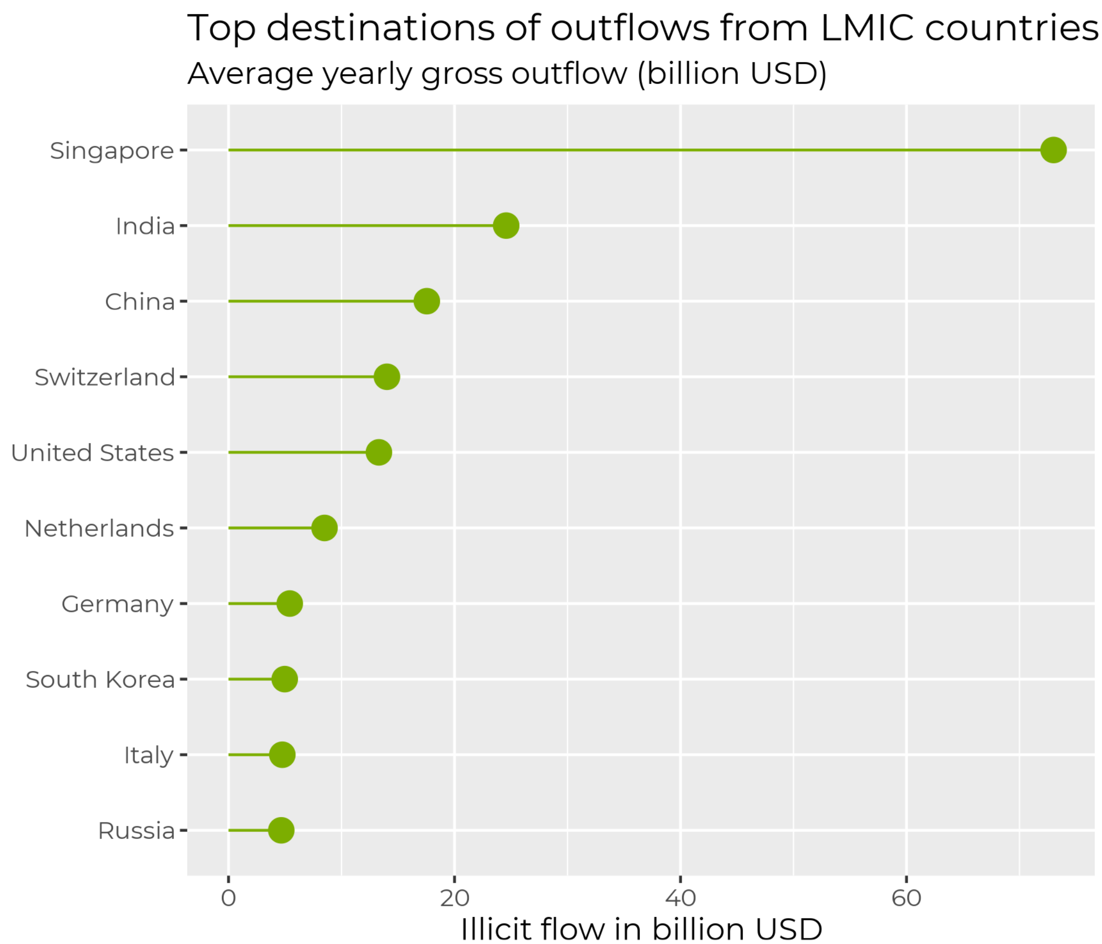


Figure 3.17: Top destinations of outflows from low and lower-middle income countries.

3.6 Discussion

The methodology of the "atlas of misinvoicing" does, of course, carry limitations. The estimates presented here are likely to be an underestimate of the true extent of the phenomenon of trade misinvoicing. First, they do not cover misinvoicing of the trade in services. Second, the method will not pick up misinvoicing where the distortion is repeated consistently at export and import (so-called "same-invoice faking" – see Kar (2010)). In particular, the mirror trade gaps approach will not capture when the importer and exporter collude at both ends of the transaction to submit over-valued invoices, and so the resulting mirror declarations match (World Customs Organization, 2018). For example, an importer can create a secret fund abroad to evade taxes or domestic financial controls by creating a subsidiary shell company in a foreign country. The importer can then remit an over-valued payment to the exporter by depositing the funds into the bank accounts of the (foreign) exporting company's shareholders. As a shareholder of the shell company exporter, the importer can then withdraw the illicit proceeds in small amounts at a time in ATMs in their country (World Customs Organization, 2018), a process known as "smurfing".[35] Third, using a higher level of commodity aggregation will likely result in "within-sector" netting which would underestimate the extent of misinvoicing. On the one hand, the methodological choice of using a higher level of commodity aggregation (at the 2-digit HS code) rather than more disaggregated commodity data such as the 4- or 6-digit codes is justified to avoid any false positive identification of misinvoicing due to genuine mistakes on how to classify a certain good when many similar options exist. However, the trade-off is that the higher level of aggregation will cancel out some over-invoicing of products with the under-invoicing of other products if they both fall under the same HS chapter (Kravchenko, 2018), and thus the method will miss genuine cases of misinvoicing. As a result, the estimates presented here are conservative and should be interpreted as a lower bound of the true extent of illicit financial flows from trade misinvoicing.

Moreover, the controls for delayed shipment arrivals and asymmetric reporting of re-exports might not be completely adequate, since the method assumes a linear relationship, and these phenomena are likely to have different effects across countries. However,

[35] This type of manipulation would only be detectable by exchanging information on the ultimate beneficial ownership of the traders, a proposed policy initiative in the fight against IFFs which is the subject of ongoing political negotiation.

the approach of econometrically controlling for these effects has the advantage of not requiring data on the quantities or weights being shipped (where the data coverage in Comtrade is much more scarce than for prices). Errors in trade data that are not corrected by the adjustments may also negatively affect the quality of the estimates, as might price fluctuations during shipment (Forstater, 2016) – though, errors should have only a minimal effect on the net estimates, since there would be errors both in estimating inflows and outflows which would tend to cancel each other out in the aggregate.

The paper offers several innovations by relaxing many of the existing (and sometimes implicit) assumptions in the literature. First, transportation and freight costs are no longer assumed to be a constant value; they are estimated econometrically, in a way that controls for trade misinvoicing that might have been missed by previous estimates of transport costs. Second, instead of assuming that declarations from developed economies are more accurate than declarations from poor countries, the relative trustworthiness of country declarations is empirically determined through the harmonization procedure. Importantly, the paper does not directly equate observed trade irregularities with trade misinvoicing. Nor does the paper assume that only the portion of trade discrepancies that *are* explained by predictors of illicitness are related to trade misinvoicing. Rather, the "residual" approach of the paper makes the assumption that trade gaps that cannot be explained by non-illicit or benign reasons are the result of either deliberate misinvoicing or statistical noise.

The coefficient estimates for the estimated licit and illicit margins should be interpreted cautiously and are likely to be correlational, and not causal. By not elucidating the causal mechanisms of observed trade gaps, this complicates the partition of the trade gap into its respective licit and illicit components. Moreover, the act of partitioning assumes that predictors of discrepancies can be attributed to either legitimate or illegitimate reasons, but not both at the same time. However, the estimand of interest here is not the causal effect of those predictors on the trade discrepancies; rather, it is the population quantity of the amount of trade misinvoicing. In that sense, the respective *groups* of coefficient estimates are of interest because they provide marginal effects that hold constant the other type of predictors. But individual coefficient estimates are not directly used to ascribe illicit intent to an observed trade gap. Different specifications of the gravity models that take advantage of the panel structure of the data could also be explored to increase the plausibility of causal identification of the coefficients, but there are some difficulties. Including country-time fixed effects as is sometimes recommended in the gravity literature (e.g., Yotov et al. (2016)) would absorb the variation in other predictors of interest to illicit flows that are country-specific and vary across time, such as the governance variables or many additional potential variables relating to country institutions and national policies. Future work should be directed towards conducting further sensitivity checks about the robustness of the results to the changing of assumptions and predictors used in the methodology, including the treatment of outliers, and the inclusion of additional predictors of illicit determinants.

There are two broad interpretations of the concept of "illicit financial flows" in the literature: a narrow, legalistic one where IFFs are defined as international transfers of funds that were or are illegally obtained, transferred or used; and a broad definition, which understands such flows to be any international transfers of wealth that are harmful to development (UNECA, 2018a; Blankenburg and Khan, 2012; Cobham and Janský, 2020). Of course, this begs the bigger question of what is desirable, and the decision as to what counts as an illicit financial flow becomes political, linked to what form of development one considers to be positive; and it therefore cannot be easily answered by technocrats. But the question of what illicit financial flows are, and to what extent we should tackle them, was political to begin with anyway – it must be, since, as this paper has shown, it has profound consequences for the international distribution of wealth, creating winners and losers, and it will therefore sharply divide opinion along political lines. Nevertheless, this paper hopes to provide a significant contribution to the discussion on trade misinvoicing and its likely extent. Moreover, the estimates presented here still find that the magnitude of IFFs through trade misinvoicing is substantive, broadly in line with the findings of other estimates. This suggests that existing estimates of trade misinvoicing are not, as some authors have suggested, an artefact of these statistical phenomena. Instead, the results support the argument that trade misinvoicing is real, substantial, and the conduit for hundreds of billions of dollars of illicit financial flows every year, suggesting that combating illicit financial flows should be an urgent priority for policy-makers.

3.7 Conclusion

This paper has presented the "atlas of trade misinvoicing", an original dataset of estimates for 167 countries during 2000-2018 that provides both broad country coverage and disaggregated estimates by year and by sector. Academics might find the dataset useful as a new dependent variable or might wish to use estimates of illicit flows as an additional control variable in econometric work looking at globalization, investment, and development.

Moreover, the paper offers a new methodology that seeks to mitigate some of the main concerns of the literature on trade misinvoicing estimates. In particular, the method adopts both a "residual" and a "harmonization" approach that adjusts for sources of illicit and non-illicit discrepancies in trade data and for the quality of a country's declaration in order to provide a more accurate estimate.

This paper demonstrates how the "atlas" can be used in further analysis by identifying leading sources, destinations, and commodities involved in trade misinvoicing. Natural resources lead the commodities affected by trade misinvoicing in developing countries, while the main destinations appear to be either countries with a high level of financial secrecy or countries in which many multinational corporations are based. Illicit financial flows deplete government revenues, weaken governance, and erode state institutions. Inflows are also detrimental to development since they are untaxed and invisible to governments. The estimates presented here are conservative and should be interpreted as a lower-bound of possible misinvoicing. This paper provides empirical confirmation that illicit outflows and inflows are pervasive across developing countries. In order to meet the challenge of the 2030 Agenda for Sustainable Development and to realize the SDGs, reducing illicit financial flows will be crucial for domestic resource mobilization.

Bibliography

ACTSA (2019). The money drain: how trade misinvoicing and unjust debt undermine economic and social rights in Southern Africa. *ACTSA Briefing Paper*, pages 1–16.

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

Anderson, J. E. and Van Wincoop, E. (2003). Gravity with gravitas: A solution to the border puzzle. *American Economic Review*, 93(1):170–192.

Andreas, P. (2015). International Politics and the Illicit Global Economy. *Perspectives on Politics*, 13(3):782–788.

Baker, R., Clough, C., Kar, D., LeBlanc, B., and Simmons, J. (2014). *Hiding in Plain Sight: Trade Misinvoicing and the Impact of Revenue Loss in Ghana, Kenya, Mozambique, Tanzania, and Uganda: 2002-2011*. Global Financial Integrity, Washington, DC.

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

Bhagwati, J. (1964). On the Underinvoicing of Imports. *Bulletin of the Oxford University Institute of Economics & Statistics*, 27(4):389–397.

Blankenburg, S. and Khan, M. (2012). Governance and Illicit Flows. In Reuter, P., editor, *Draining Development? Controlling Flows of Illicit Funds from Developing Countries*, chapter 2, pages 21–68. World Bank Group, Washington, DC.

Chang, H.-J. (2005). *Why Developing Countries Need Tariffs? How WTO NAMA Negotiations Could Deny Developing Countries' Right to a Future*. South Centre, Geneva, Switzerland.

Cobham, A. and Janský, P. (2020). *Estimating Illicit Financial Flows - A Critical Guide to the Data, Methodologies, and Findings*. Oxford University Press, Oxford, UK.

Cobham, A., Janský, P., and 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. *World Bank Research Observer*, 35(1):44–86.

Conte, M., Cotterlaz, P., and Mayer, T. (2021). *The CEPII Gravity Database*. Centre d'Études Prospectives et d'Informations Internationales (CEPII), Paris, France.

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

De Boyrie, M. E., Pak, S. J., and Zdanowicz, J. S. (2005). Estimating the magnitude of capital flight due to abnormal pricing in international trade: The Russia-USA case. *Accounting Forum*, 29(3):249–270.

Devarajan, S. (2013). Africa's statistical tragedy. *Review of Income and Wealth*, 59:9–16.

Disdier, A. C. and Head, K. (2008). The puzzling persistence of the distance effect on bilateral trade. *The Review of Economics and Statistics*, 90(1):37–48.

Doumbia, D. and Lauridsen, M. L. (2019). Closing the SDG Financing Gap — Trends and Data. *International Finance Corporation Note*, 73:1–8.

Dunning, T. (2008). *Crude Democracy: Natural Resource Wealth and Political Regimes.* Cambridge University Press, Cambridge, UK.

ECLAC (2016). *Economic Survey of Latin America and the Caribbean 2016: The 2030 Agenda for Sustainable Development and the challenges of financing for development.* United Nations Economic Commission for Latin America and the Caribbean, Santiago, Chile.

ESCWA (2018). *Illicit Financial Flows in the Arab Region.* United Nations Economic and Social Commission for Western Asia, Beirut, Lebanon.

Eunomix Research (2017). *A review of the UNCTAD report on trade misinvoicing, with a full counterfactual on South African exports.* Eunomix, Johannesburg, South Africa.

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

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

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

Fisman, R. and 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.

Forstater, M. (2016). Illicit Flows and Trade Misinvoicing: Are we looking under the wrong lamppost? *CMI Insights*, 5:1–8.

Forstater, M. (2018). Illicit financial flows, trade misinvoicing, and multinational tax avoidance: the same or different? *Center for Global Development Policy Paper*, 123:1–33.

Gara, M., Giammatteo, M., and 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.

Gaulier, G. and Zignago, S. (2010). *BACI: International Trade Database at the Product-Level (the 1994-2007 Version)*. Centre d'Études Prospectives et d'Informations Internationales (CEPII), Paris, France.

Griffiths, P. (2003). *The Economist's Tale: A Consultant Encounters Hunger and the World Bank*. Zed Books, London, UK.

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, Addis Ababa, Ethiopia.

Hong, K., H Pak, C., and J Pak, S. (2014). Measuring abnormal pricing – an alternative approach. *Journal of Money Laundering Control*, 17(2):203–218.

Hong, K. P. and Pak, S. J. (2017). Estimating Trade Misinvoicing from Bilateral Trade Statistics: The Devil is in the Details. *International Trade Journal*, 31(1):3–28.

Ibis Ghana and Africa Centre for Energy Policy (2015). *Illicit Financial Flows and the Extractive Industry in Ghana*. Africa Centre for Energy Policy, Accra, Ghana.

Iimi, A. (2007). Escaping from the Resource Curse: Evidence from Botswana and the Rest of the World. *IMF Staff Papers*, 54(4):663–699.

Jean, S., Mitaritonna, C., and Vatan, A. (2018). Institutions and Customs Duty Evasion. *CEPII Working paper*, 24:1–33.

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

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

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

Jha, R. and Truong, D. N. (2015). Estimates of Trade Mis-invoicing and Their Macroeconomic Outcomes for the Indian Economy. *Review of Economics & Finance*, 5(2):19–34.

Kar, D. (2010). *The Drivers and Dynamics of Illicit Financial Flows from India: 1948-2008*. Global Financial Integrity, Washington, DC.

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

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

Kellenberg, D. and Levinson, A. (2019). Misreporting trade: Tariff evasion, corruption, and auditing standards. *Review of International Economics*, 27(1):106–129.

Kravchenko, A. (2018). *Where and how to dodge taxes and shift money abroad using trade misinvoicing A beginner's guide*. United Nations Economic and Social Commission for Asia and the Pacific, Bangkok, Thailand.

Leblanc, B. (2014, May 9). Trade Misinvoicing, or How to Steal from Africa. *Think Africa Press*.

Lewis, D., McNeill, R., and Shabalala, Z. (2019, April 24). Exclusive: Gold worth billions smuggled out of Africa. *Reuters*.

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

Mevel, S., 'Ofa, S. V., and Karingi, S. (2013). Quantifying illicit financial flows from Africa through trade mis-pricing and assessing their incidence on African economies. In *16th GTAP Conference*, pages 1–33, Shanghai, China.

Miao, G. and Fortanier, F. (2017). Estimating Transport and Insurance Costs of International Trade. *OECD Statistics Working Papers*, 80:3–34.

Mikuriya, K. (2018). Capital flight in trade payment. *The Global Governance Project*.

Moore, M. (2007). How Does Taxation Affect the Quality of Governance? *Institute of Development Studies (IDS) Working Paper*, 280:1–37.

Morgenstern, O. (1950). *On the Accuracy of Economic Observations*. Princeton University Press, Princeton, NJ.

Morgenstern, O. (1974). On the accuracy of economic observations: foreign trade statistics. In Bhagwati, J., editor, *Illegal Transactions in International Trade*, chapter 7, pages 87–122. North-Holland, Amsterdam.

Ndikumana, L. and Boyce, J. K. (2010). Measurement of Capital Flight: Methodology and Results for Sub-Saharan African Countries. *African Development Review*, 22(4):471–481.

Ndikumana, L. and Boyce, J. K. (2018). Capital Flight from Africa - Updated Methodology and New Estimates. *PERI Research Report*, pages 1–12.

Nitsch, V. (2012). Trade Mispricing and Illicit Flows. In Reuter, P., editor, *Draining Development? Controlling Flows of Illicit Funds from Developing Countries*, chapter 10, pages 309–334. World Bank Group, Washington, DC.

Nitsch, V. (2016). Trillion dollar estimate: Illicit financial flows from developing countries. *Darmstadt Discussion Papers in Economics*, 227:1–5.

Oliver, S., Jablonski, R., and Hastings, J. V. (2017). The tortuga disease: The perverse effects of illicit foreign capital. *International Studies Quarterly*, 61(2):312–327.

Patnaik, I., Sen Gupta, A., and Shah, A. (2008). *Trade misinvoicing: A channel for de facto capital account openness*. National Institute for Public Finance and Policy, New Delhi, India.

Picard, J. (2003). Can We Estimate the Global Scale and Impact of Illicit Trade? In Miklaucic, M. and Brewer, J., editors, *Convergence - Illicit Networks and National Security in the Age of Globalization*, chapter 3, pages 37–60. National Defense University Press, Washington, DC.

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

Rodrik, D. (2018). What do trade agreements really do? *Journal of Economic Perspectives*, 32(2):73–90.

Ross, M. L. (1999). The Political Economy of the Resource Curse. *World Politics*, 51(2):297–322.

Ross, M. L. (2015). What Have We Learned about the Resource Curse? *Annual Review of Political Science*, 18(1):239–259.

Sachs, J. D. and Warner, A. (1995). Economic Reform and the Process of Global Integration. *Brookings Papers on Economic Activity*, 1995(1):1–118.

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

Sarraf, M. and Jiwanji, M. (2001). Beating the Resource Curse: The Case of Botswana. *World Bank Environment Department Working Paper*, 83:1–22.

Schuster, C. and 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, Geneva, Switzerland.

South African Revenue Service (2016). SARS Media Statement on UNCTAD Report.

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

Tauberer, J. (2014). *Open Government Data: The Book*. Second edition.

Tax Justice Network (2018). Financial Secrecy Index.

Tørsløv, T. R., Wier, L. S., and Zucman, G. (2018). The Missing Profits of Nations. Working paper, National Bureau of Economic Research.

UN General Assembly (2015). Transforming our world: the 2030 Agenda for Sustainable Development.

UNCTAD (2016). *Trade Misinvoicing in Primary Commodities in Developing Countries: The cases of Chile, Côte d'Ivoire, Nigeria, South Africa and Zambia*. United Nations Conference on Trade and Development, Geneva, Switzerland.

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, Geneva, Switzerland.

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, Addis Ababa, Ethiopia.

UNECA (2018a). *A study on the global governance architecture for combating illicit financial flows*. United Nations Economic Commission for Africa, Addis Ababa, Ethiopia.

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

UNECA (2019). Multinational corporations, tax avoidance and evasion and natural resources management. In *Economic Report on Africa 2019*, chapter 6, pages 118–137. United Nations Economic Commission for Africa, Addis Ababa, Ethiopia.

United Nations Inter-agency Task Force on Financing for Development (2019). *Report of the Inter-agency Task Force on Financing for Development: Financing for Sustainable Development Report 2019*. United Nations, New York, NY.

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

UNODC and 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, Vienna, Austria and Geneva, Switzerland.

Van Rensburg, D. (2016, August 1). How wrong the UN was on SA gold ”smuggling”. *City Press*.

Vézina, P. L. (2015). Illegal trade in natural resources: Evidence from missing exports. *International Economics*, 142:152–160.

Vicard, V. (2015). Profit shifting through transfer pricing: evidence from French firm level trade data. *Document de Travail, Banque de France*, 555:5–27.

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

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

Yotov, Y. V. (2012). A simple solution to the distance puzzle in international trade. *Economics Letters*, 117(3):794–798.

Yotov, Y. V., Piermartini, R., Monteiro, J.-A., and 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, Geneva, Switzerland.

Zdanowicz, J. S. (2004). US trade with the world and Al Qaeda watch list countries - 2001 - an estimate of money moved out of and into the US due to suspicious pricing in international trade.

Zdanowicz, J. S. (2009). Trade Based Money Laundering and Terrorist Trade Based Money Laundering and Terrorist. *Review of Law and Economics*, 5(2):855–878.

Chapter 4

Machine learning for missing data on illicit trade

4.1 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 (Jerven, 2009; Devarajan, 2013; Sandefur and 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 (Paige et al., 2020; Jerven, 2013; Devarajan, 2013; Beegle et al., 2016), and who are particularly afflicted by the harmful consequences of illicit financial flows (Reuter, 2012; UNECA, 2017; UNCTAD, 2020). 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 4.2 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 4.2.1 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 4.2.2 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 4.2.3 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 4.3 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 4.3.1), studies of trade-based money laundering (section 4.3.2), and analyses of the determinants of trade misinvoicing (section 4.3.3). The rest of the paper presents the data, methodology, and findings. Section 4.4 describes the outcome and predictor variables in more detail. Section 4.5 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 4.6. Section 4.7 discusses potential applications of the method and its limitations, and section 4.8 concludes.

4.2 The application of machine learning methods to illicit finance

4.2.1 Predicting the "atlas" measure of trade misinvoicing

Defining 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; Reuter, 2012; High Level Panel on Illicit Financial Flows from Africa, 2015; Cobham and Janský, 2020). 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 (Rose, 2014; Nelleman and INTERPOL Environmental Crime Programme, 2012), and grand corruption (van der Does de Willebois et al., 2011; Findley et al., 2020). 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 (Morse, 2019; FATF, 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,

[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 and Janský, 2020; Tørsløv et al., 2018; UNECA, 2018a).

e.g., Christian Aid (2009); Cobham et al. (2020); 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 and Christensen, 2013; Shaxson, 2011). Andersen et al. (2017) found that exogenous increases in petroleum rents were associated with an increase in hidden wealth in 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); 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 (O'Hare et al., 2014; High Level Panel on Illicit Financial Flows from Africa, 2015; UNECA, 2017; UNCTAD, 2020).

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 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

[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)).

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 (Jerven, 2013; Devarajan, 2013; Jerven and Johnston, 2015; Sandefur and 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 esti-

[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.

mate 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 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.

4.2.2 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

[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

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 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); Kleinberg et al. (2017); Ge et al. (2020); 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

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 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 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 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 policymakers, 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 counterterrorism 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

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.

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 and 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 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.

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

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

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. 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,

[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.

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, 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.

4.2.3 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 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

[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.

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 (FATF, 2021; Brainard, 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; Tiwari et al., 2020; Labib 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 and 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 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

[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.

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 and Balsa, 2015). 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 (Raza and Haider, 2011; Badal-Valero et al., 2018). 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. Wohl and Kennedy (2018); Quimba and Barral (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 and 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 and Spiess, 2017; Storm et al., 2020). Section 4.3 below critically reviews the literature in order to identify a set of variables that are likely to realize high predictive returns for misinvoiced trade.

4.3 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 4.3.1 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 4.3.2 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 4.3.3 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.

4.3.1 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 and Van 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 and Van Wincoop, 2003; Tinbergen, 1962; Ferwerda et al., 2013; Disdier and Head, 2008).

The model is often presented in its logarithmic form to linearize the parameters in order to allow for estimation with linear regression models: $\ln V_{ij}^X = \ln \beta_0 + \beta_1 \ln Y_i + \beta_2 \ln N_i + \beta_3 \ln Y_j + \beta_4 \ln N_j + \beta_5 \ln D_{ij} + \beta_6 P_{ij}$ (Tinbergen, 1962; McCallum, 1995). Variations include assuming that $\beta_2 = \beta_4 = 0$ to remove population size from the equation (Tinbergen, 1962; Ferwerda et al., 2013; Bergstrand, 1985), 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 (Ferwerda et al., 2013; Disdier and Head, 2008).[13]

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 (Bergstrand, 1985; Anderson and Van Wincoop, 2003).

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

[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 them, and G is the gravitational constant (Disdier and 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).

(Disdier and 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 Van 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 and Van Wincoop, 2003; Feenstra, 2004; Piermartini and Yotov, 2016; Yotov et al., 2016; Santos Silva and Tenreyro, 2006) 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 (Yotov, 2012; Piermartini and Yotov, 2016; 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 and 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 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 and 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 and Yotov, 2016). Others suggest that the Gamma PML estimator is preferable under certain conditions given by the data (Head and Mayer, 2014). Clearly, developing theory-consistent estimation methods remains an open problem in the economics of international trade (Head and 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 and Irwin, 1995) or estimating multiplicative models with PPML and GPML (e.g., Disdier and Head (2008); 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.

4.3.2 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). 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

[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.

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 and Unger, 2009). Likewise, corruption is expected to have ambiguous effects on whether a host country 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

[16] And perhaps more so, to exceptional societal challenges such as the deepening and perpetuation of global inequalities (Piketty and 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.

and 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 and 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) (Ferwerda et al., 2013, 2020; Collin, 2019). 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 and 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.

4.3.3 The market and regulatory abuse dimension

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 (2014); Cobham and Janský (2020) 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 4.1 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 and Davis, 2020). When the value of imports is overstated, excess funds leave the country disguised as a form of trade payment (Schuster and 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 and 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 and 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 and 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 and Khan, 2012; Cobham, 2014; Cobham and Janský, 2020; Kar and 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 and Khan, 2012; Spanjers and Salomon, 2017; Salomon, 2019). 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 and 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 4.1: Typology of trade misinvoicing manipulations by trade flow and direction.

Determinants of trade misinvoicing

As illustrated by Table 4.1, 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 and 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 and 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 (Worku et al., 2016; Jean and Mitaritonna, 2010), 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 (Worku et al., 2016; Rijkers et al., 2017; Carrère and Grigoriou, 2015).

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, 2001b). 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 4.3.1); the "illicit premium" dimension introduced by the Walker model (see section 4.3.2) 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 4.3.3). The data used to represent these variables is presented in the next section.

4.4 Data

4.4.1 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.

4.4.2 Predictor variables

Section 4.3 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 C.1 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.

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

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)

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`

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`)

4.5 Methods

4.5.1 Pre-processing the data

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, 2018b). 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, 2018a).

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 (Sandefur and Glassman, 2015; Devarajan, 2013; Jerven, 2009; Jerven and Johnston, 2015; Jerven, 2016). 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 4.1 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 in Africa (Andersen et al., 2017). If the gross outflows are netted of the inflows, the findings from the "atlas"

[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).

database indicate that Africa experienced mostly net outflows from 2009 onwards.

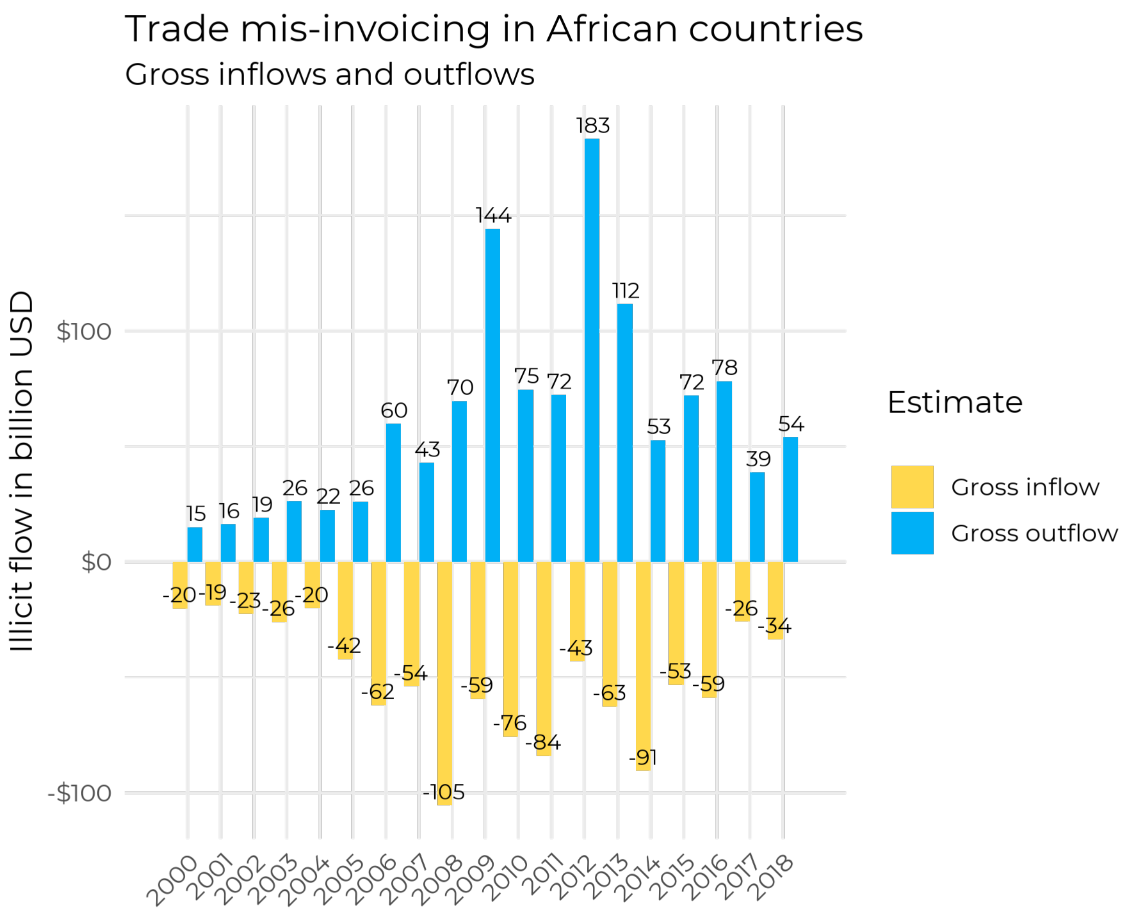


Figure 4.1: Yearly gross inflows (negative value) and gross outflows in Africa.

The incidence of outflows and inflows varies substantially across countries in the continent, as shown in figures 4.2 and 4.3. 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.

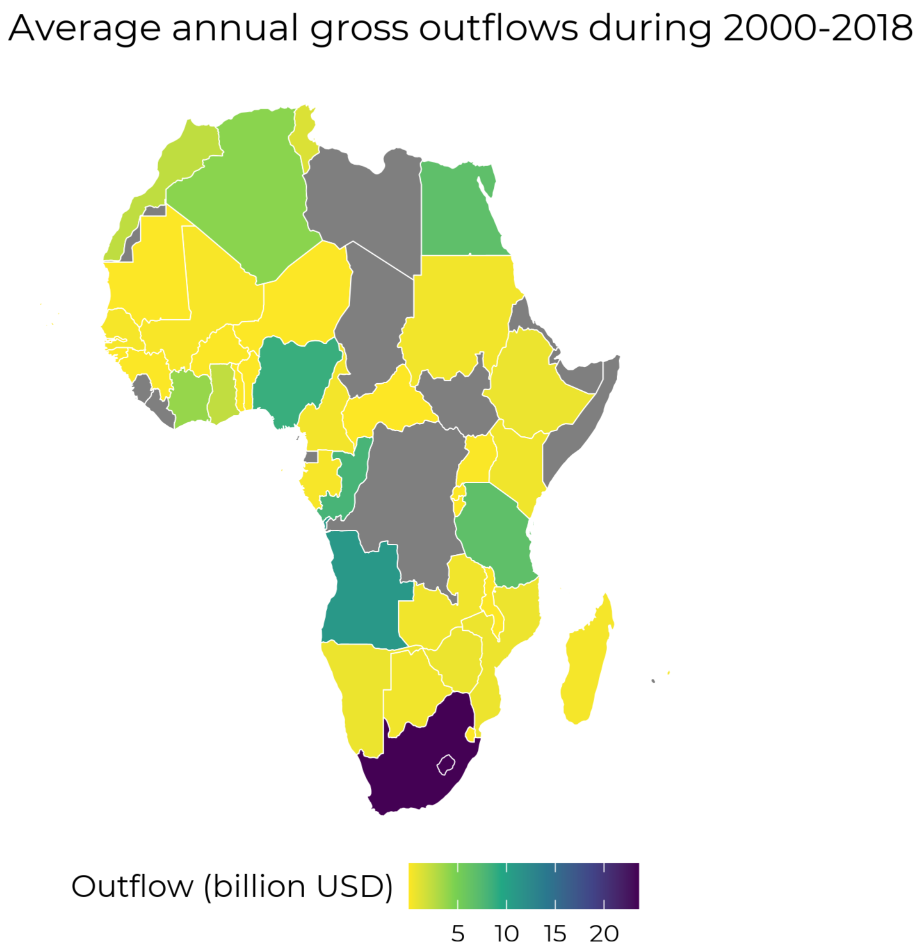


Figure 4.2: Average annual gross (without reversals) outflows during 2000-2018. Countries with missing data are in grey.

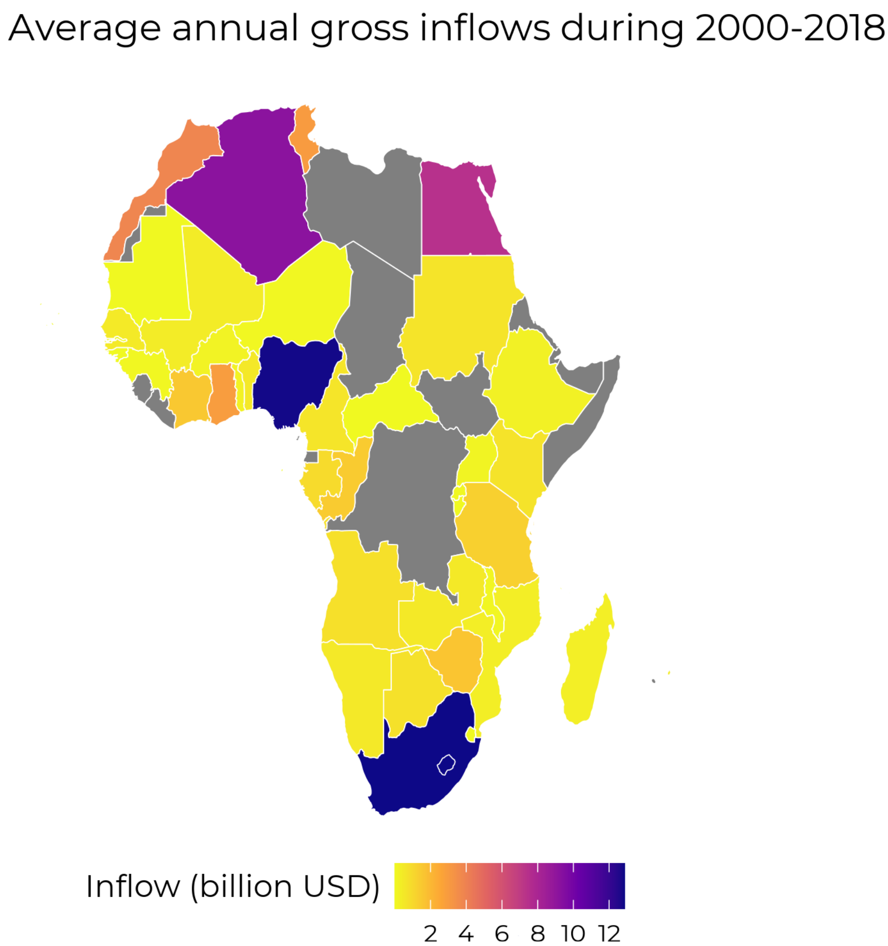


Figure 4.3: Average annual gross (without reversals) inflows during 2000-2018. The legend refers to the absolute value of inflows. Countries with missing data are in grey.

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 4.2.

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

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 4.4.

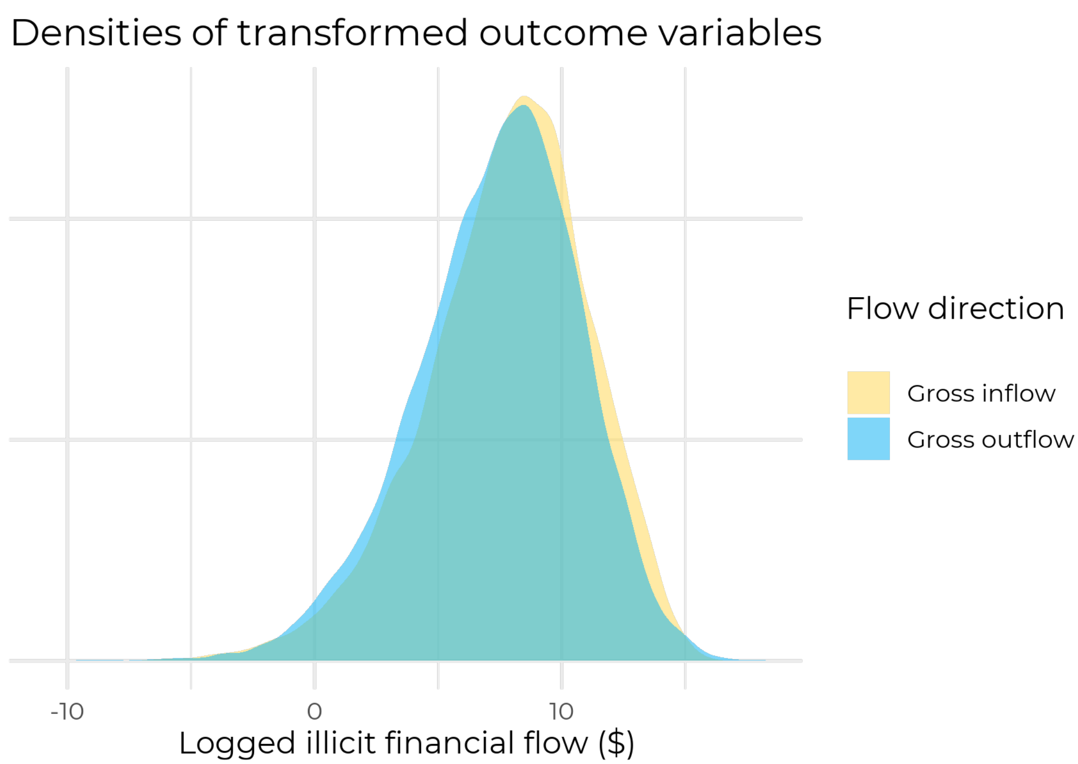


Figure 4.4: Distribution of logged outcome variable for gross illicit outflows and gross illicit inflows. Data are pooled for all African countries and years 2000-2018.

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

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

Random Forest models.

The dichotomous variables (dummy indicators taking a value of either 0 or 1) in the data are: `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 4.5 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.

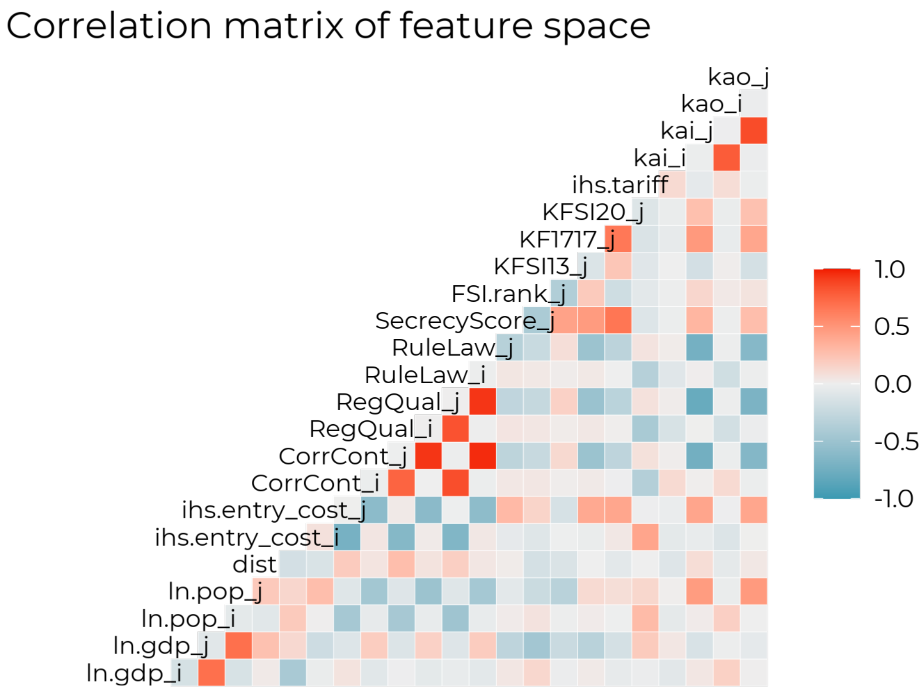


Figure 4.5: Correlation matrix of the feature space for continuous predictors, after transformations. Unilateral variables relating to reporters and partners have the suffixes _i and _j respectively. Variables that do not have a suffix are bilateral features.

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 4.3.

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

4.5.2 Tuning and training the machine learning models

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 (2001a)). 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 (James et al., 2013; Hastie et al., 2017). 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 (Hastie et al., 2017; Breiman, 2001a). 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.

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

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

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.

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

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

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 C.2 of the appendix. The Random Forest model was fit using the *scikit-learn* library.[22]

4.6 Results

4.6.1 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 4.4 reports two types of performance metrics for both the error of the model (using the Mean Square Error) 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

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

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 and 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 4.4: 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 4.6 and 4.7 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.3 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.

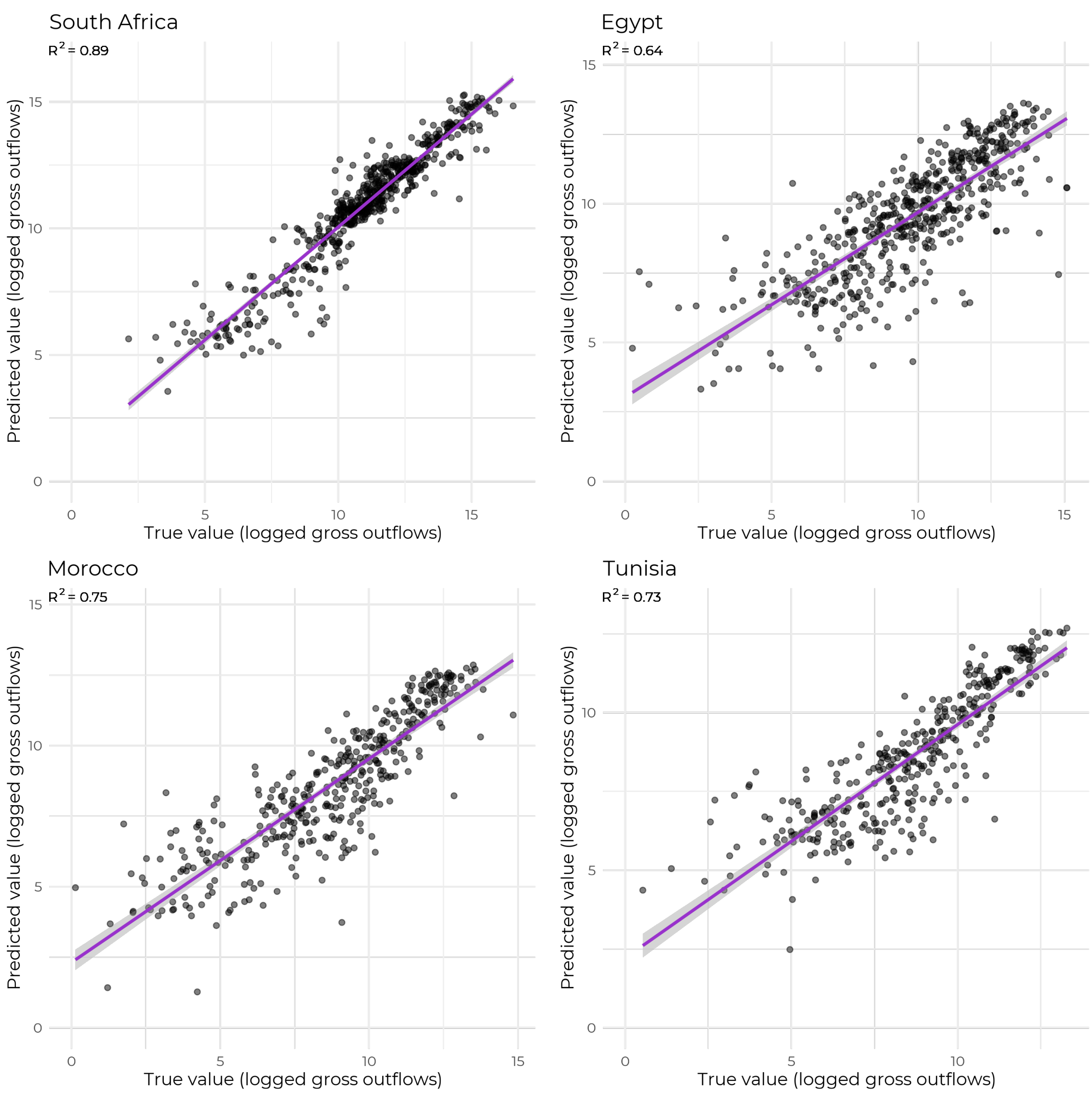


Figure 4.6: Cross-validated out-of-fold predictions of logged outflows for countries by Random Forest model trained on pooled model of African countries over 2000-2018.

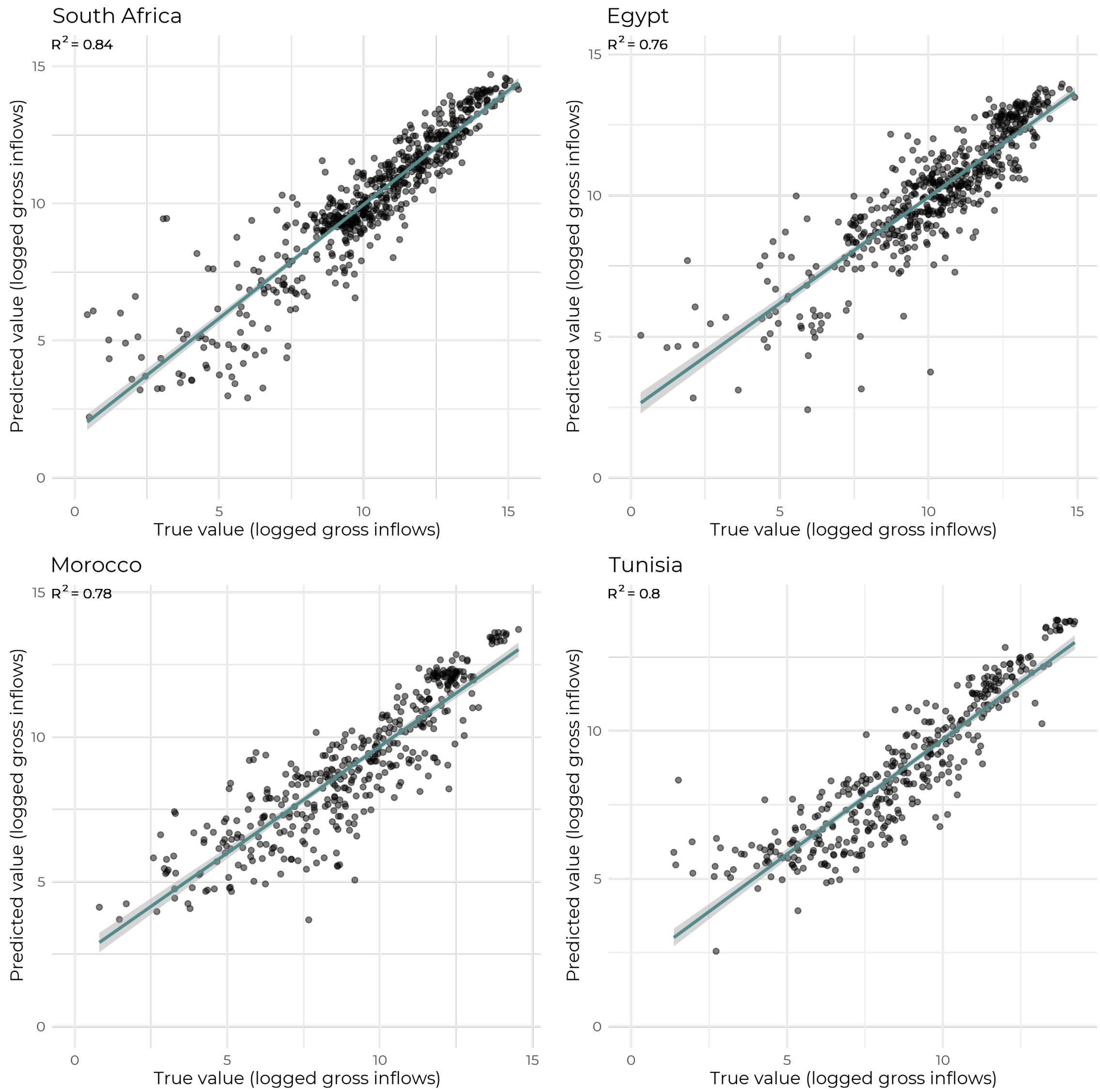


Figure 4.7: Cross-validated out-of-fold predictions of logged inflows for countries by Random Forest model trained on pooled model of African countries over 2000-2018.

4.6.2 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 4.8.

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.

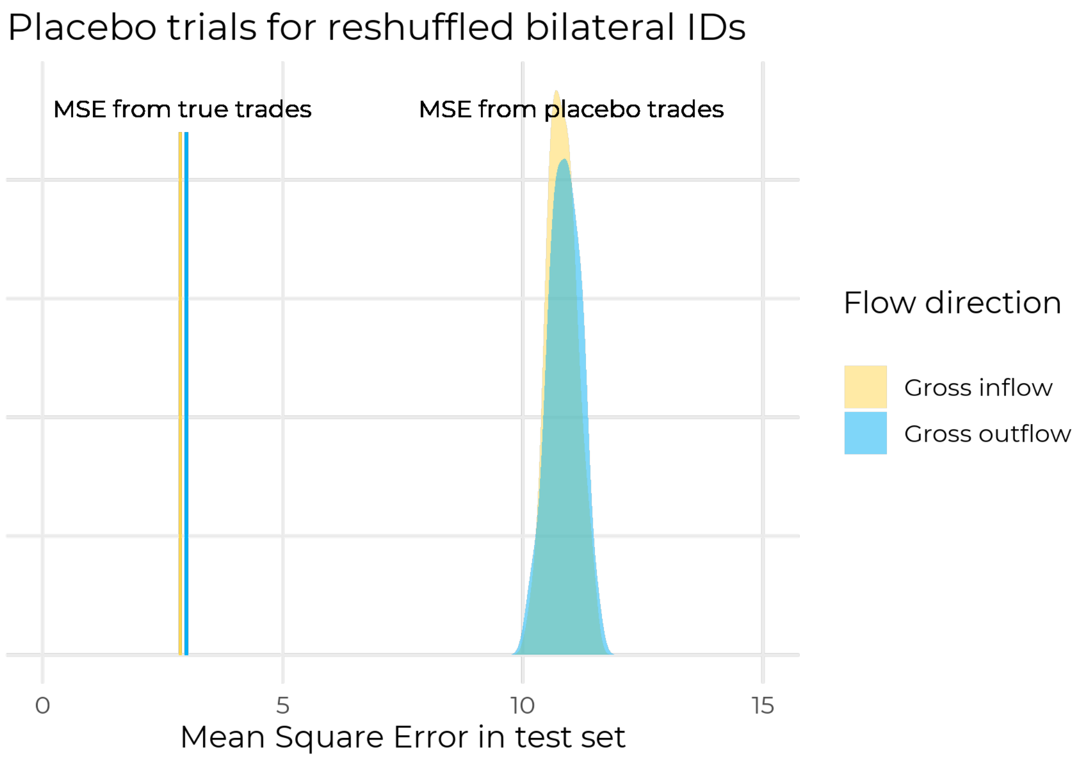


Figure 4.8: Mean Square Error on the independent test set of model trained on true bilateral transactions, and of 100 placebo models trained on randomly reshuffled bilateral trades.

4.6.3 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

[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.

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 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 4.5. 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.

[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.

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

Table 4.5: Estimates of the models' ability to generalize across borders.

4.6.4 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:

$$\begin{aligned}\log Y_{ijt} = \alpha + \boldsymbol{Gravity\beta} + \boldsymbol{Governance\gamma} \\ + \boldsymbol{RegulEnvironment\lambda} + \boldsymbol{FinancialIntegrity\pi} + \epsilon_{ijt}\end{aligned} \tag{4.1}$$

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 4.3; 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:

$$
\begin{aligned}
\log \texttt{GER\_Tot\_IFF}_{ijt} = \alpha &+ \beta_1 \log \texttt{GDP}_{it} + \beta_2 \log \texttt{GDP}_{jt} + \\
&+ \beta_3 \texttt{comlang}_{ij} + \beta_4 \texttt{col45}_{ij} + \beta_5 \texttt{RTA}_{ij} \\
&+ \gamma_1 \texttt{CorrCont}_{it} + \gamma_2 \texttt{CorrCont}_{jt} + \gamma_3 \texttt{RegQual}_{jt} \\
&+ \lambda_1 \texttt{FATF}_{it} + \lambda_2 \texttt{FATF}_{jt} \\
&+ \pi_1 \texttt{tariff}_{ijt} + \pi_2 \texttt{kao}_{it} + \pi_3 \texttt{kai}_{jt} + \epsilon_{ijt}
\end{aligned}
\tag{4.2}
$$

$$\begin{aligned}\log \texttt{In\_GER\_Tot\_IFF}_{ijt} = \alpha &+ \beta_1 \log \texttt{GDP}_{it} + \beta_2 \log \texttt{GDP}_{jt} + \\ &+ \beta_3 \texttt{comlang}_{ij} + \beta_4 \texttt{col45}_{ij} + \beta_5 \texttt{RTA}_{ij} \\ &+ \gamma_1 \texttt{CorrCont}_{it} + \gamma_2 \texttt{CorrCont}_{jt} + \gamma_3 \texttt{RegQual}_{it} \\ &+ \lambda_1 \texttt{FATF}_{it} + \lambda_2 \texttt{FATF}_{jt} \\ &+ \pi_1 \texttt{tariff}_{ijt} + \pi_2 \texttt{kai}_{it} + \pi_3 \texttt{kao}_{jt} + \epsilon_{ijt}\end{aligned} \tag{4.3}$$

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 C.2 of the appendix. Performance metrics are reported in table 4.6 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**	
	Outflows	Inflows	Outflows	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 4.6: Predictive performance of the linear regression models on illicit trade outcomes. The full model is rank deficient.

4.7 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., Jean et al. (2016); Henderson et al. (2012); Chen and Nordhaus (2011); 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.

4.8 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.

Bibliography

Abugre, C., Cobham, A., Lépissier, A., Etter-Phoya, R., Meinzer, M., Monkam, N., and Mosioma, A. (2020). *Vulnerability and Exposure to Illicit Financial Flows risk in*

Africa. 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), Addis Ababa, Ethiopia.

Alexandre, C. and Balsa, J. (2015). Client Profiling for an Anti-Money Laundering System. *arXiv*, 1510.00878(cs.LG).

Andersen, J. J., Johannesen, N., Lassen, D. D., and 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.

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

Anderson, J. E. and Van Wincoop, E. (2003). Gravity with gravitas: A solution to the border puzzle. *American Economic Review*, 93(1):170–192.

Athey, S. (2017). Beyond prediction: Using big data for policy problems. *Science*, 355:483–485.

Badal-Valero, E., Alvarez-Jareño, J. A., and 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.

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

Batarseh, F. A., Gopinath, M., and Monken, A. (2020). Artificial Intelligence Methods for Evaluating Global Trade Flows. *International Finance Discussion Papers, Board of Governors of the Federal Reserve System*, 1296:1–31.

Batarseh, F. A., Gopinath, M., Nalluru, G., and Beckman, J. (2019). Application of Machine Learning in Forecasting International Trade Trends. In *AAAI FSS-19: Artificial Intelligence in Government and Public Sector*, pages 1–4, Arlington, VA.

Beegle, K., Christiaensen, L., Dabalen, A., and Gaddis, I. (2016). The State of Data for Measuring Poverty. In *Poverty in a Rising Africa*, pages 25–56. World Bank Group, Washington, DC.

Bergstrand, J. H. (1985). The Gravity Equation in International Trade: Some Microeconomic Foundations and Empirical Evidence. *The Review of Economics and Statistics*, 67(3):474–481.

Bhagwati, J. (1964). On the Underinvoicing of Imports. *Bulletin of the Oxford University Institute of Economics & Statistics*, 27(4):389–397.

Blankenburg, S. and Khan, M. (2012). Governance and Illicit Flows. In Reuter, P., editor, *Draining Development? Controlling Flows of Illicit Funds from Developing Countries*, chapter 2, pages 21–68. World Bank Group, Washington, DC.

Brainard, L. (2021). Supporting Responsible Use of AI and Equitable Outcomes in Financial Services. In *AI Academic Symposium, hosted by the Board of Governors of the Federal Reserve System*, Washington, DC.

Breiman, L. (2001a). Random Forests. *Machine Learning*, 45:5–32.

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

Buehn, A. and Eichler, S. (2011). Trade misinvoicing: The dark side of world trade. *World Economy*, 34(8):1263–1287.

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.

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

Chandler, D., Levitt, S. D., and List, J. A. (2011). Predicting and preventing shootings among at-risk youth. *American Economic Review*, 101(3):288–292.

Chen, X. and Nordhaus, W. D. (2011). Using luminosity data as a proxy for economic statistics. *Proceedings of the National Academy of Sciences*, 108(21):8589–8594.

Christensen, J. (2012). The hidden trillions: Secrecy, corruption, and the offshore interface. *Crime Law and Social Change*, 57(3):325–343.

Christian Aid (2009). *False profits: robbing the poor to keep the rich tax-free.* London, UK.

Clausing, K. A. (2003). Tax-motivated transfer pricing and US intrafirm trade prices. *Journal of Public Economics*, 87(9-10):2207–2223.

Cobham, A. (2014). Benefits and Costs of the IFF Targets for the Post-2015 Development Agenda. *Copenhagen Consensus Center Assessment Paper*, pages 1–31.

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

Cobham, A. and Janský, P. (2020). *Estimating Illicit Financial Flows - A Critical Guide to the Data, Methodologies, and Findings.* Oxford University Press, Oxford, UK.

Collin, M. (2019). Illicit Financial Flows: Concepts, Measurement, and Evidence. *World Bank Research Observer*, 35(1):44–86.

Conte, M., Cotterlaz, P., and Mayer, T. (2021). *The CEPII Gravity Database*. Centre d'Études Prospectives et d'Informations Internationales (CEPII), Paris, France.

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

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

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*, 59:9–16.

Disdier, A. C. and Head, K. (2008). The puzzling persistence of the distance effect on bilateral trade. *The Review of Economics and Statistics*, 90(1):37–48.

Efimov, D., Xu, D., Kong, L., Nefedov, A., and Anandakrishnan, A. (2020). Using generative adversarial networks to synthesize artificial financial datasets. In *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, pages 1–8, Vancouver, Canada.

Eichengreen, B. and 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.

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, New York, NY.

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

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

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

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

Fernández, A., Klein, M. W., Rebucci, A., and 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., and Bikker, J. A. (2013). Gravity models of trade-based money laundering. *Applied Economics*, 45(22):3170–3182.

Ferwerda, J., van Saase, A., Unger, B., and Getzner, M. (2020). Estimating money laundering flows with a gravity model-based simulation. *Scientific Reports*, 10(1):1–11.

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

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

Fisman, R. and 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.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3-5):75–174.

Gara, M., Giammatteo, M., and 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.

Ge, Q., Huang, X., Fang, S., Guo, S., Liu, Y., Lin, W., and Xiong, M. (2020). Conditional Generative Adversarial Networks for Individualized Treatment Effect Estimation and Treatment Selection. *Frontiers in Genetics*, 11(585804):1–17.

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

Gonzalez-Lira, A. and 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., and Bengio, Y. (2014). Generative Adversarial Nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Gopinath, M., Batarseh, F. A., and 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., and Han, J. (2014). Outlier Detection for Temporal

Data: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, 26(9):2250–2267.

Haberly, D. and Wójcik, D. (2014). Regional blocks and imperial legacies: Mapping the global offshore FDI network. *Economic Geography*, 91(3):251–280.

Haberly, D. and Wójcik, D. (2015). Tax havens and the production of offshore FDI: An empirical analysis. *Journal of Economic Geography*, 15(1):75–101.

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

Head, K. and Mayer, T. (2014). Gravity Equations: Workhorse, Toolkit, and Cookbook. In Gopinath, G., Helpman, E., and Rogoff, K., editors, *Handbook of International Economics*, volume 4, chapter 3, pages 131–195. North Holland.

Henderson, J. V., Storeygard, A., and Weil, D. N. (2012). Measuring Economic Growth from Outer Space. *American Economic Review*, 102(2):994–1028.

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, Addis Ababa, Ethiopia.

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

Javorcik, B. S. and Narciso, G. (2008). Differentiated products and evasion of import tariffs. *Journal of International Economics*, 76(2):208–222.

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

Jean, S. and 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.

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

Jerven, M. (2016). Trapped between tragedies and miracles: Misunderstanding African economic growth. *Development Policy Review*, 34(6):911–915.

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

Joaristi, M., Serra, E., and Spezzano, F. (2019). Detecting suspicious entities in Offshore Leaks networks. *Social Network Analysis and Mining*, 9(1):1–15.

Jullum, M., Løland, A., Huseby, R. B., Ånonsen, G., and Lorentzen, J. (2020). Detecting money laundering transactions with machine learning. *Journal of Money Laundering Control*, 23(1):173–186.

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

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

Kellenberg, D. and Levinson, A. (2019). Misreporting trade: Tariff evasion, corruption, and auditing standards. *Review of International Economics*, 27(1):106–129.

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

Kleinberg, J., Ludwig, J., Mullainathan, S., and Obermeyer, Z. (2015). Prediction policy problems. *American Economic Review*, 105(5):491–495.

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

Lépissier, A. and Cobham, A. (2019). *Risk Measures for Illicit Financial Flows Dataset*. DOI:10.5281/zenodo.3371739.

Li, Y., Duan, D., Hu, G., and 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.

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., and Tsiatis, A. (2014). *Handbook of Missing Data Methodology*. Chapman and Hall/CRC, New York, NY.

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

Mullainathan, S. and Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106.

Nelleman, C. and 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, Norway.

OECD (2017). *OECD Transfer Pricing Guidelines for Multinational Enterprises and Tax Administrations*. OECD Publishing, Paris, France.

O'Hare, B., Makuta, I., Bar-Zeev, N., Chiwaula, L., and 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.

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

Paige, J., Fuglstad, G.-A., Riebler, A., and 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*, pages 1–31.

Patnaik, I., Sen Gupta, A., and Shah, A. (2012). Determinants of Trade Misinvoicing. *Open Economies Review*, 23(5):891–910.

Paula, E. L., Ladeira, M., Carvalho, R. N., and 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*, pages 954–960.

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

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

Pinkovskiy, M. and 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.

Quimba, F. M. A. and 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. and Haider, S. (2011). Suspicious activity reporting using Dynamic Bayesian Networks. *Procedia Computer Science*, 3:987–991.

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

Rijkers, B., Baghdadi, L., and Raballand, G. (2017). Political Connections and Tariff Evasion Evidence from Tunisia. *World Bank Economic Review*, 31(2):459–482.

Rose, A. K. and Spiegel, M. M. (2007). Offshore Financial Centres: Parasites or Symbionts? *The Economic Journal*, 117(523):1310–1335.

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

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

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

Santos Silva, J. and 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., and Yu, X. (2016). Detection of money laundering groups using supervised learning in networks. *arXiv*, 1608.00708(cs.SI):1–11.

Schuster, C. and 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, Geneva, Switzerland.

Senator, T. E., Goldberg, H. G., Wooton, J., Cottini, M. A., Khan, A. F. U., Klinger, C. D., Llamas, W. M., Marrone, M. P., and 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, New York, NY.

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

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

Storm, H., Baylis, K., and Heckelei, T. (2020). Machine learning in agricultural and applied economics. *European Review of Agricultural Economics*, 47(3):849–892.

Sudjianto, A., Nair, S., Yuan, M., Zhang, A., Kern, D., and Cela-Díaz, F. (2010). Statistical methods for fighting financial crimes. *Technometrics*, 52(1):5–19.

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, New York, NY.

Tiwari, M., Gepp, A., and Kumar, K. (2020). A review of money laundering literature: the state of research in key areas. *Pacific Accounting Review*, 32(2):271–303.

Tørsløv, T. R., Wier, L. S., and Zucman, G. (2018). The Missing Profits of Nations. *National Bureau of Economic Research Working Paper Series*, 24701:1–29.

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, Geneva, Switzerland.

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, Addis Ababa, Ethiopia.

UNECA (2018a). *A study on the global governance architecture for combating illicit financial flows*. United Nations Economic Commission for Africa, Addis Ababa, Ethiopia.

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

UNECA (2019). Multinational corporations, tax avoidance and evasion and natural resources management. In *Economic Report on Africa 2019*, chapter 6, pages 118–137. United Nations Economic Commission for Africa, Addis Ababa, Ethiopia.

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, Vienna, Austria.

UNODC and 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, Vienna, Austria and Geneva, Switzerland.

van der Does de Willebois, E., Halter, E. M., Harrison, R. A., Park, J. W., and 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, Washington, DC.

Varma, S. and Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC bioinformatics*, 7:91.

Vézina, P. L. (2015). Illegal trade in natural resources: Evidence from missing exports. *International Economics*, 142:152–160.

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

Walker, J. and Unger, B. (2009). Measuring global money laundering: "The Walker Gravity Model". *Review of Law and Economics*, 5(2):821–853.

West, J. and Bhattacharya, M. (2016). Intelligent financial fraud detection: A comprehensive review. *Computers and Security*, 57:47–66.

Wohl, I. and 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., and 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.

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

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

Yotov, Y. V. (2012). A simple solution to the distance puzzle in international trade. *Economics Letters*, 117(3):794–798.

Yotov, Y. V., Piermartini, R., Monteiro, J.-A., and 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, Geneva, Switzerland.

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.

Chapter 5

Conclusion

This dissertation is located in the intellectual tradition of applied social science research in observational settings and substantively advances scholarship on the “wicked” problems of climate change and illicit finance. By breaking off facets of the wicked problem into discrete “tame” problems, I identify three types of questions in the study of climate change and illicit finance that can be solved with discrete inferential tasks: a causal evaluation problem, a measurement problem, and a missing data problem.

These problems have distinct estimands that require different classes of estimators and modes of statistical inference. The dissertation operates in turn within the scientific frameworks of causal, descriptive, and predictive inference. The first chapter evaluates the causal effect of a climate mitigation policy in the UK on carbon emissions in a quasi-experimental setting by exploiting the fact that the UK was the first among its peers to adopt the climate policy treatment. Methodologically, the chapter shows how to conduct climate impact evaluations without relying on unrealistic Business-As-Usual scenarios for the counterfactual. The second chapter originates the “atlas” measure of illicit trade that can be used by academics and practitioners in future work. A measure to provide credible descriptive inferences of illicit trade is a fundamental first step in advancing the incipient academic field of illicit finance. Methodologically, the chapter proposes innovations to remedy long-standing criticisms that existing estimates of illicit trade are illusions created by artefacts of the recording procedures of international trade. The third chapter uses machine learning to address missing data gaps in the “atlas” database by demonstrating that illicit trade outcomes can reliably be recovered even if data on the observed trade flow is not available for training. Methodologically, the chapter demonstrates that models trained on readily available data on country-level features are able to meaningfully discern structure in the dyadic nature of countries’ relationships that are predictive of illicit trade.

The substantive contributions of this dissertation further our understanding of climate policy instruments and of the incidence of trade misinvoicing, particularly in developing countries. Insights from the first chapter suggest that voluntary climate reforms that make concessions to domestic producers, as a result of a negotiated bargaining process, are still able to meaningfully reduce emissions in the absence of a binding global climate agreement and despite departures from first-best economic theory. Findings from the second chapter provide empirical confirmation that developing countries are particularly afflicted by trade misinvoicing, and suggest that combating illicit financial flows will be an integral part of domestic resource mobilization in poor countries, if they are to reduce their dependence on foreign assistance and obtain new sources of financing for development. Results from the third chapter indicate that off-the-shelf machine learning algorithms can reliably be trained using researcher-compiled aggregate economic data to address gaps in official government statistics in countries that have weak institutions for statitistical data collection, and 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.

In the prologue, I identified climate change and illicit finance as "wicked" problems to reveal the limitations of epistemologies that assume that there is a single generative process in nature for these phenomenona that can be known to be true. Yet, researchers wishing to generate policy-relevant insights that are obtained using rigorous empirical research can parse the wicked problems by explicitly seeking to ameliorate a specific dimension of the problem instead. By cultivating a polyglot technical repertoire and abstaining from the fetishization of one inferential framework over another, this dissertation has shown how to deploy innovative quantitative techniques that are appropriately directed to the inferential target.

Appendix A

Appendix for Chapter 2

A.1 Imputation procedure for missing data

Data on Germany's emissions per capita are missing in the World Bank's World Development Indicators (WDI) database prior to 1991. The underlying source of these data is the Carbon Dioxide Information Analysis Center (CDIAC) at Oak Ridge, which provides the most widely used inventory of national CO_2 emissions. We reconstruct Germany's emissions for the missing years by sourcing emissions from the Federal Republic of Germany and the German Democratic Republic directly from the CDIAC database. Emissions per capita are derived using the population indicator "SP.POP.TOTL" from the WDI database.

Data on Kuwait's emissions per capita are missing in the WDI database for the years 1992-1994, yet the WDI does have data on emissions in kilotons for those years. The WDI is also missing data on Kuwait's population for those years, so we turn to the underlying source of the population data in the WDI, which is the United Nations World Population Prospects (WPP) database. We use the 2019 WPP database to obtain data on Kuwait's total population for 1992-1994. We compute emissions per capita for Kuwait for 1992-1994 by dividing emissions from the indicator "EN.ATM.CO2E.KT" (multiplied by 1,000) by the population data from WPP.

The WDI and CDIAC do not have data on Liechtenstein's CO_2 emissions. Therefore, we obtain data from Liechtenstein's National Inventory Report in 2017 to the UNFCCC (Principality of Liechtenstein, 2017) on CO_2 emissions (excluding emissions from Land Use and Land Use Change in order to be comparable with CDIAC/WDI data) for 1990-2000. Following the procedure for Germany and Kuwait, we compute CO_2 emissions in metric tons per capita by dividing emissions (appropriately converted to metric tons) by total population.

A.2 Make-up of the synthetic UK

In order to construct our main donor pool, we proceed as follows. There are 85 countries that were either OECD members or classified as high or upper middle income countries in 2001 by the World Bank. We exclude the 8 countries that were treated in 2001 (which includes the UK). We exclude the 10 countries that had missing data on CO_2 emissions between 1990 and 2000. We also exclude countries that had a population smaller than 250,000 in 2001. The 51 countries that remain form our donor pool.

As a robustness check, we also restrict the donor pool to OECD or high income countries in 2001 ($n = 32$), and then to OECD members in 2001 ($n = 22$). The full list of countries can be found in table A.3.

Figure A.1 displays the weights applied to each country in the donor pool.

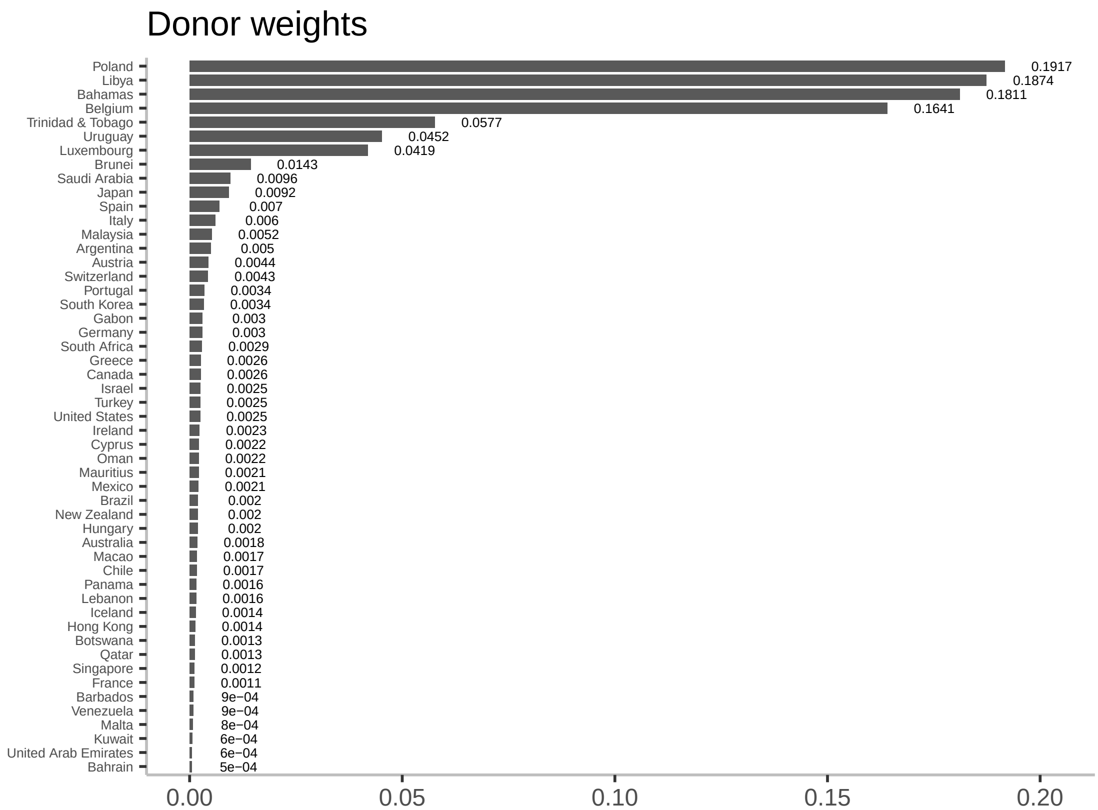


Figure A.1: Weights applied to donor countries.

Figure A.2 displays the per capita emissions trajectories in the 8 countries that make up 88% of the weights used to construct the synthetic UK.

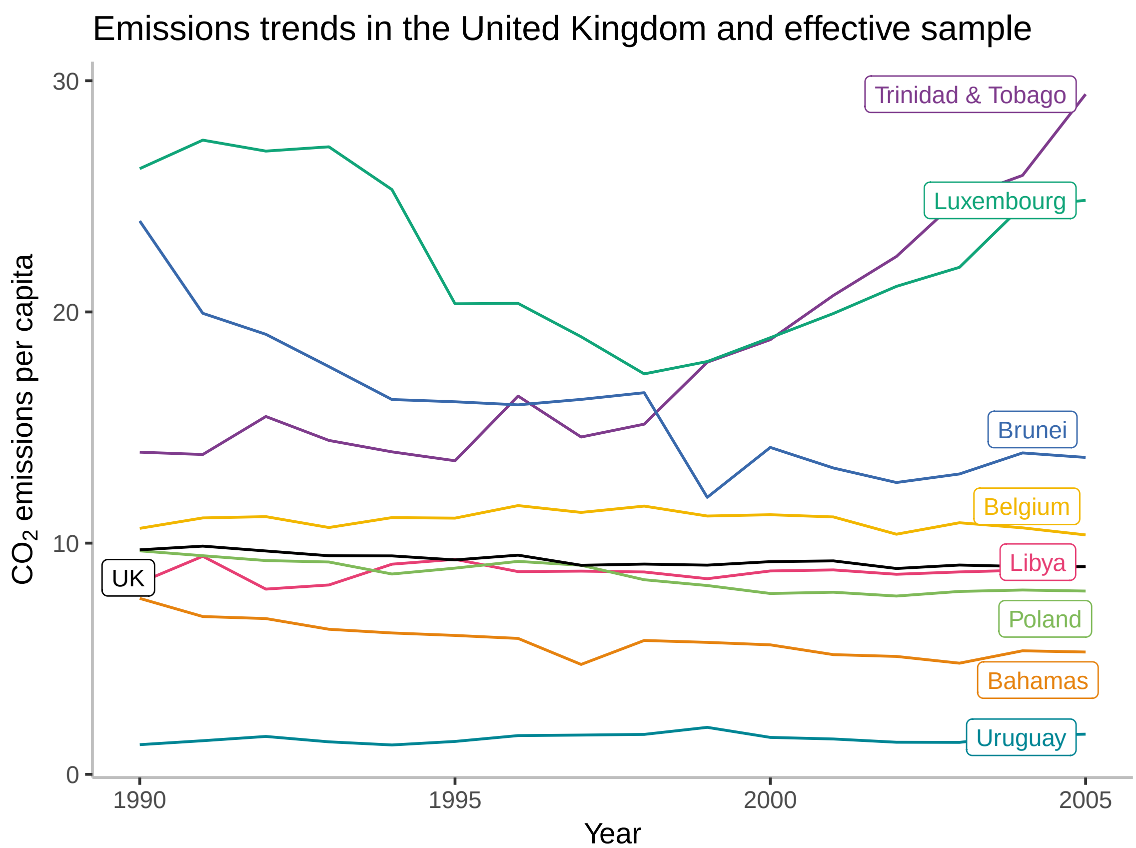


Figure A.2: CO_2 per capita emissions in the UK and in the effective sample of countries used to construct the synthetic UK. The donor pool comprises countries that were OECD, high or upper middle income countries in 2001, excluding countries with population less than 250,000 in 2001.

While both Luxembourg and Trinidad and Tobago had increasing emissions trends, our results are not dependent on having them in the donor pool, and the treatment effects remain comparable. If we were to drop Luxembourg from the donor pool, we estimate a treatment effect of -8.5% ($p = 0.02$), and if we were to drop Trinidad and Tobago from the donor pool, we estimate a treatment effect of -6.3% ($p = 0.059$). To recall, our main treatment effect is -9.8% ($p = 0.02$).

A.3 Balance

The synthetic UK achieves much better balance on the predictor variables than an unweighted sample of OECD, high and upper middle income countries. Table A.1 below shows that the pre-treatment values of the outcome variable as simulated by the synthetic UK are very similar to those actually observed in the UK in that time. By contrast, the pre-treatment values in the whole sample, when taking an unweighted mean, differ markedly from those observed in the UK.

Predictor	Treated UK	Synthetic UK	Sample Mean	Weight
1990 emissions per capita	9.711	9.714	9.101	0.061
1991 emissions per capita	9.871	9.861	8.94	0.022
1992 emissions per capita	9.661	9.654	9.353	0.153
1993 emissions per capita	9.455	9.469	9.905	0.101
1994 emissions per capita	9.448	9.451	9.949	0.095
1995 emissions per capita	9.275	9.268	9.897	0.099
1996 emissions per capita	9.480	9.477	9.818	0.079
1997 emissions per capita	9.043	9.042	10.041	0.082
1998 emissions per capita	9.094	9.101	10.022	0.096
1999 emissions per capita	9.048	9.049	9.91	0.092
2000 emissions per capita	9.200	9.199	10.297	0.122

Table A.1: Pre-treatment values of the predictor variables in the UK (column 2), its synthetic control (column 3), and in the unweighted sample (column 4); and weights applied to those predictor variables (column 5).

Figure 2 in the main text displays the difference in means in pre-treatment values of the dependent variable between the UK and the weighted synthetic counterfactual, and between the UK and the unweighted mean sample of OECD, high and upper middle income countries in 2001. Table A.2 reports summary balance statistics. The p-value for the two sample t-test indicates that we fail to reject the null hypothesis that the difference in means between the pre-treatment emissions in the UK and in the synthetic UK is 0; and the p-value for the Kolmogorov-Smirnov test suggests that we fail to reject the null that the pre-treatment values of the UK and its synthetic control come from the same distribution.

Balance statistic	
p-value two sample t-test	0.9813836
p-value Kolmogorov Smirnov test	1
Mean difference in QQ plots	0.0416667
Median difference in QQ plots	0.0416667
Maximum difference in QQ plots	0.0833333

Table A.2: Balance statistics between pre-treatment values of the dependent variable in the UK and its synthetic counterpart. p-values are reported for a two sample t-test for a difference in means and for a Kolmogorov-Smirnov test for whether the samples come from different distributions. QQ statistics are reported for the empirical CDF of both samples.

A.4 Falsification tests

In the main text we present results of a placebo test where we re-assign treatment to countries which we know to be unaffected by treatment and where we should expect to see null results. To increase our confidence that our results are not the product of chance, we should like to see the UK's treatment effect lie on the outer edges of that null distribution.

Since a large MSPE indicates a poor fit between the placebo unit and its synthetic counterpart, we cannot use these placebos are meaningful comparisons. In the main text, we present the result of the placebo test where we discard placebos with pre-treatment MSPE larger than 30. Figures A.3 and A.4 below display the gaps in the UK and in placebos where the pre-treatment MSPE cut-off is 50 and 100, respectively.

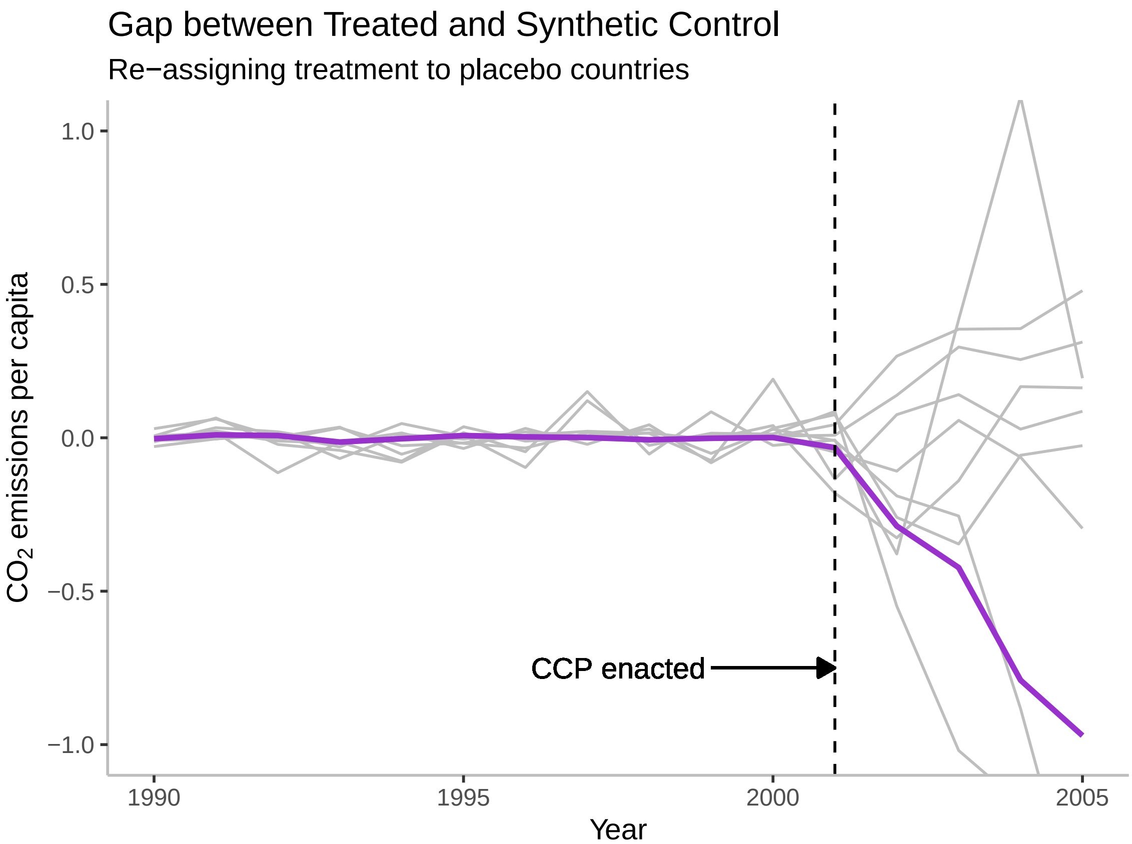


Figure A.3: Gaps in emissions per capita between the treated unit and its synthetic counterpart. The thick purple line represents the gaps for the UK. The grey lines represent the distribution of placebo treatment effects. Countries with a pre-treatment MSPE greater than 50 times that of the UK have been excluded (see Methods for details).

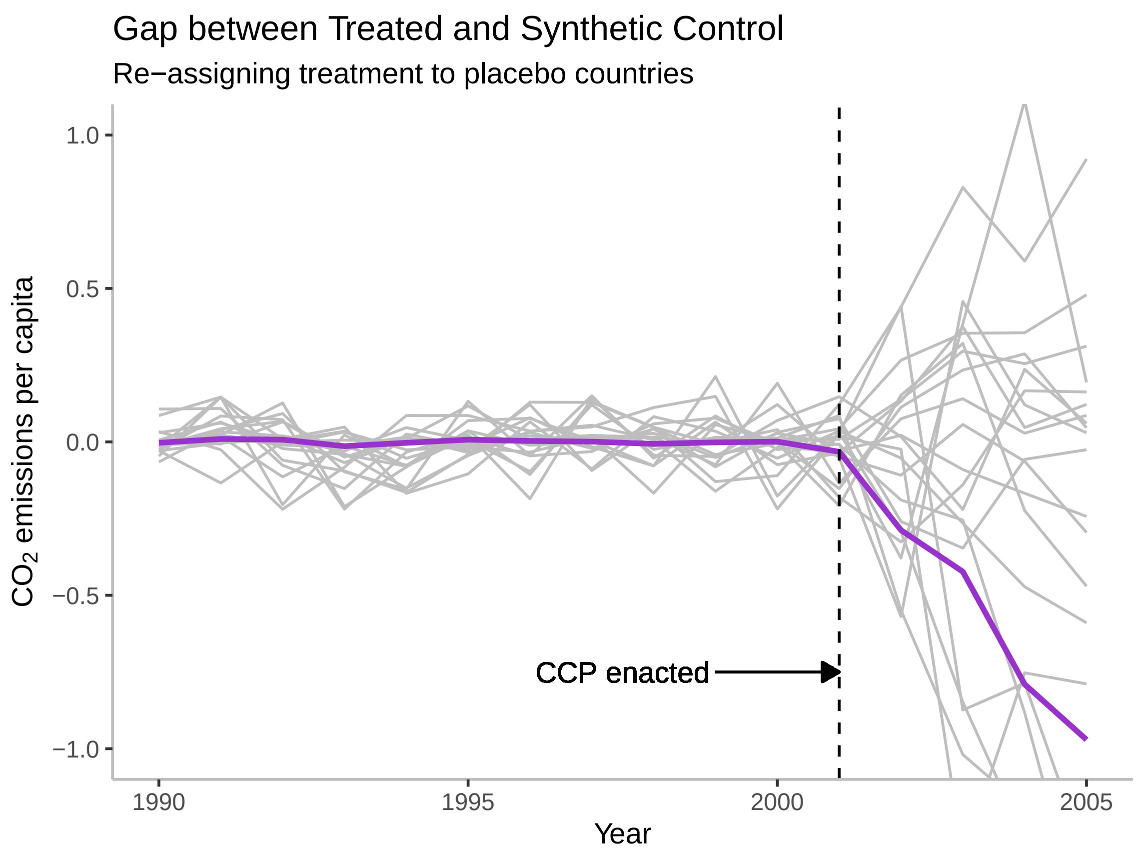


Figure A.4: Gaps in emissions per capita between the treated unit and its synthetic counterpart. The thick purple line represents the gaps for the UK. The grey lines represent the distribution of placebo treatment effects. Countries with a pre-treatment MSPE greater than 100 times that of the UK have been excluded (see Methods for details).

We then use a test statistic that obviates the need to decide on a cut-off point: the ratio of post-treatment to pre-treatment MSPE for each country. Figure 4 in the main text shows that the ratio in the UK lies at the end of the right tail of that distribution, which indicates that the effect is likely not the result of chance.

Finally, figure A.5 displays the ratio for each country in the sample. The UK has the largest ratio statistic out of all countries in the sample. If we were to pick a country at random under uniform sampling from the entire sample, the probability of obtaining a ratio statistic as large as the UK's is $1/51 \approx 0.02$.

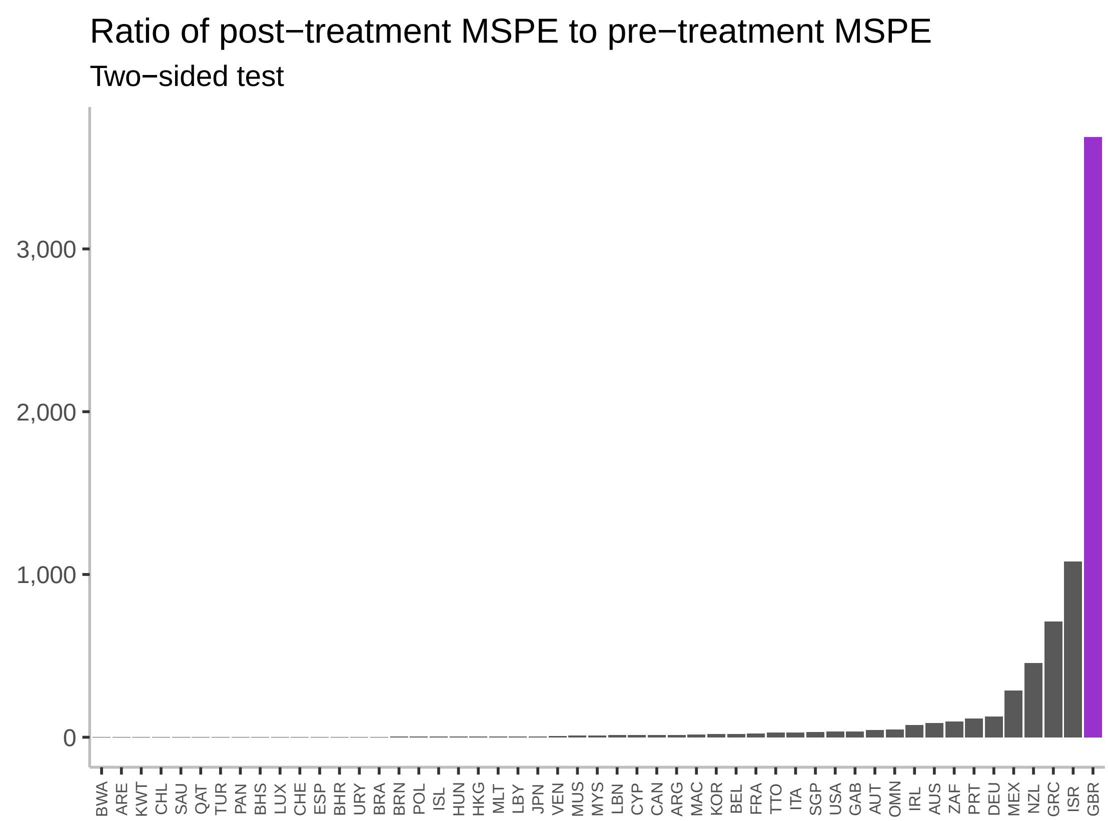


Figure A.5: Ratio of post-treatment Mean Square Prediction Error (MSPE) to pre-treatment MSPE for all countries in the sample.

A.5 Robustness check: placebo treatment year

We conduct placebo "in time" checks, where we assign treatment to the years prior to the passage of the CCP, in which we expect to see no treatment effect. We report the results of those "in time" placebo tests in figure 6 in the main text, and in figures A.6, A.7, A.8, A.9, and A.10 below. We use the same donor pool as reported in the main text ($n = 51$), but only include the pre-(placebo)-treatment years as predictor variables. A large positive placebo effect would weaken our confidence in our results. As shown in the pages below, our analysis passes the "in time" placebo test, except for the year 1998 where there is a positive placebo effect (figure A.7). The placebo test for 1997 looks like it is positive (figure A.8), though this placebo run demonstrates a poor fit between the UK and its placebo synthetic control and therefore should be discarded as uninformative.

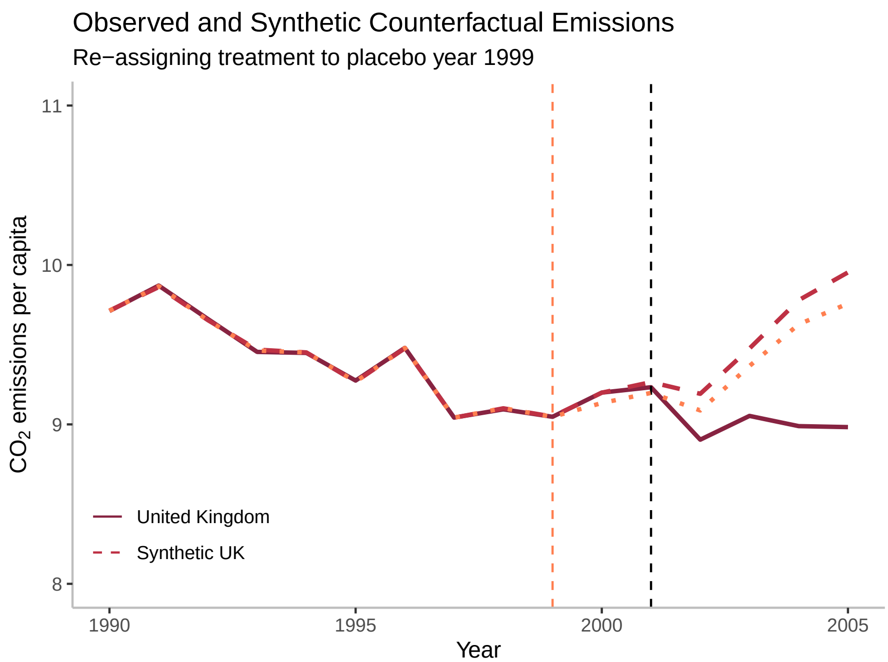


Figure A.6: Observed and synthetic counterfactual emissions for a placebo run where treatment occurs in 1999.

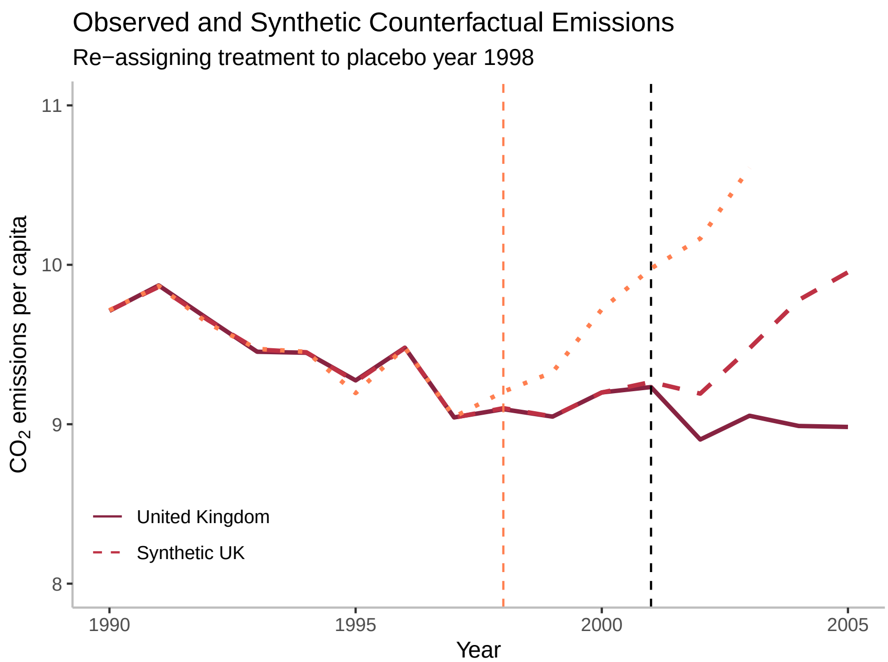


Figure A.7: Observed and synthetic counterfactual emissions for a placebo run where treatment occurs in 1998.

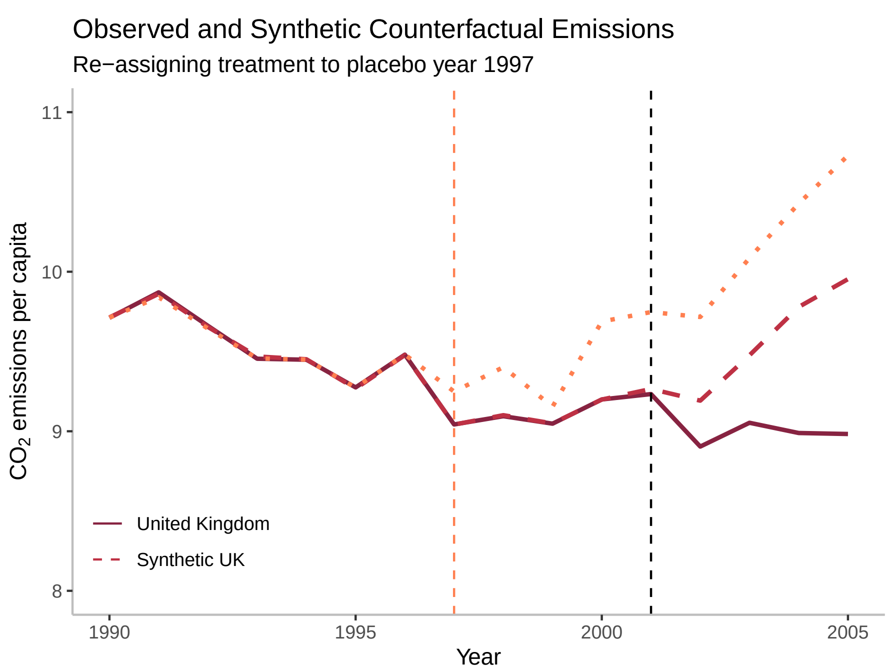


Figure A.8: Observed and synthetic counterfactual emissions for a placebo run where treatment occurs in 1997.

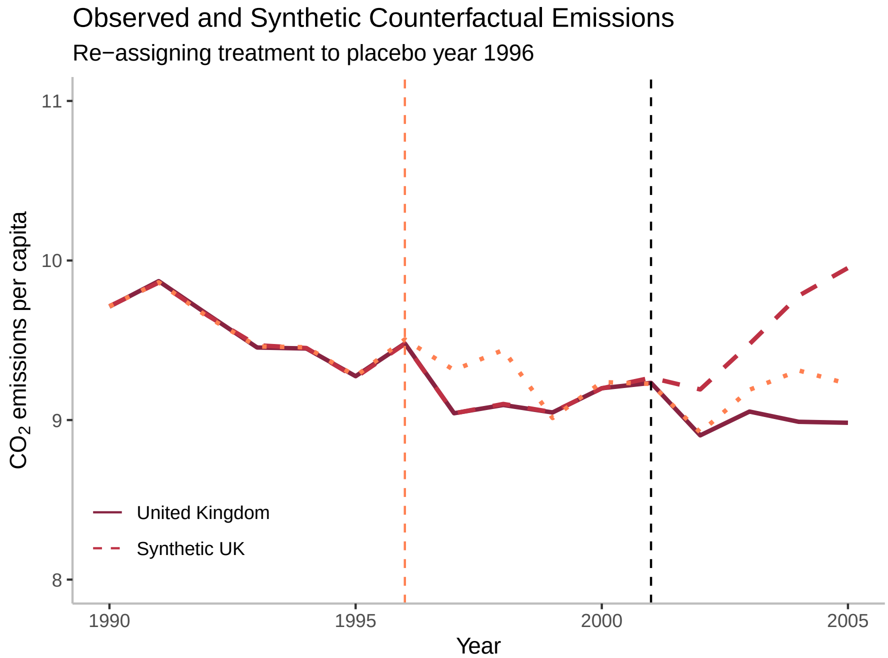


Figure A.9: Observed and synthetic counterfactual emissions for a placebo run where treatment occurs in 1996.

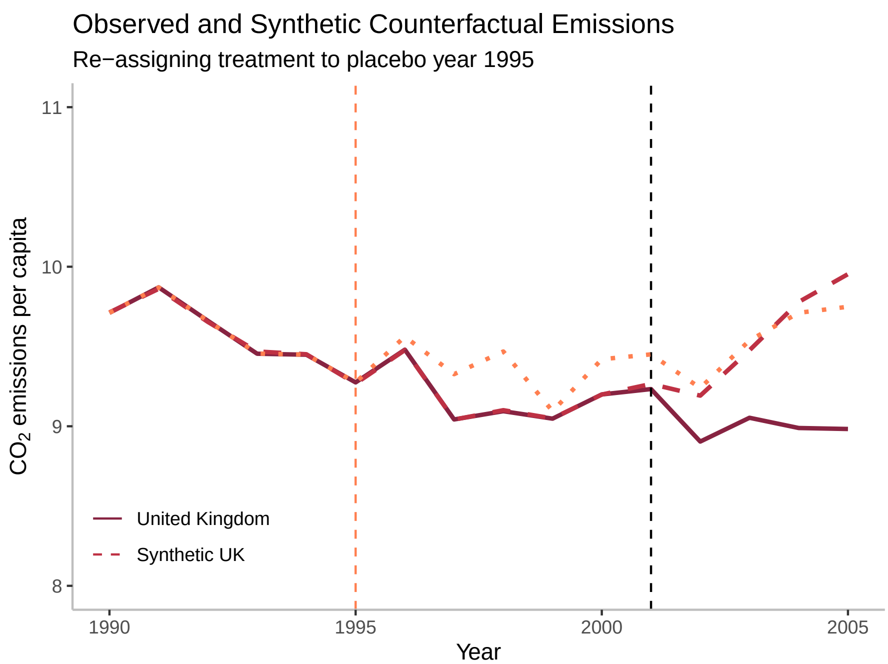


Figure A.10: Observed and synthetic counterfactual emissions for a placebo run where treatment occurs in 1995.

A.6 Robustness check: placebo countries

Single-country placebos are provided on the following pages in figures A.11-A.60. The SCM algorithm failed when running a placebo test on Barbados, so the falsification test is out of 51 rather than out of 52. This happens in cases where the system is singular and the algorithm fails to converge.

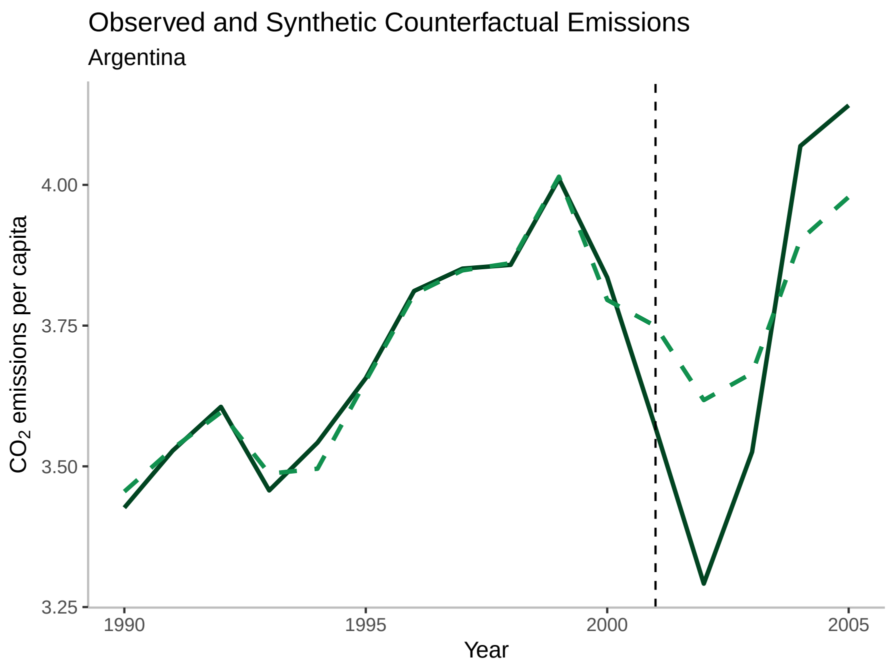


Figure A.11: Observed and synthetic counterfactual emissions for placebo country Argentina.

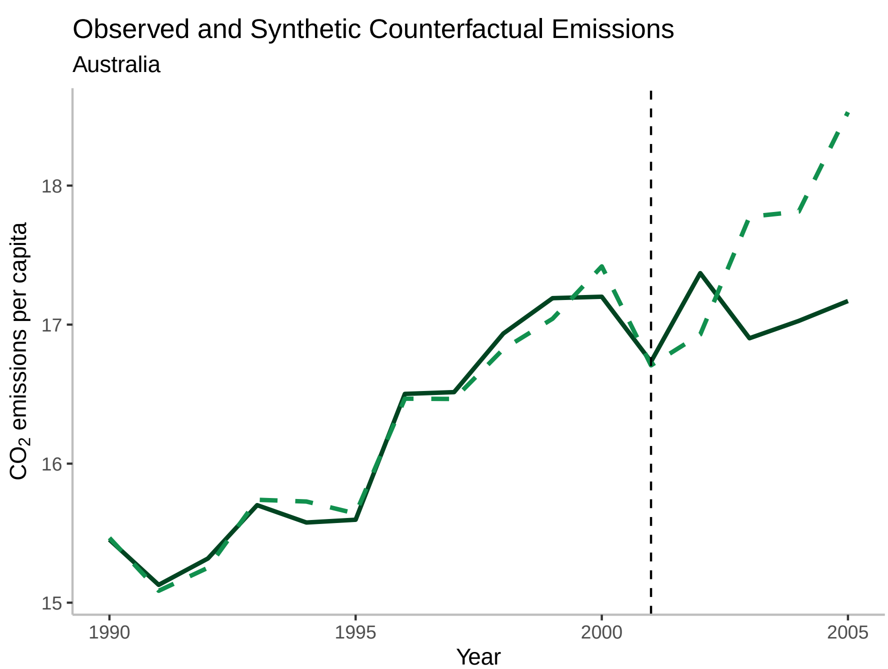


Figure A.12: Observed and synthetic counterfactual emissions for placebo country Australia.

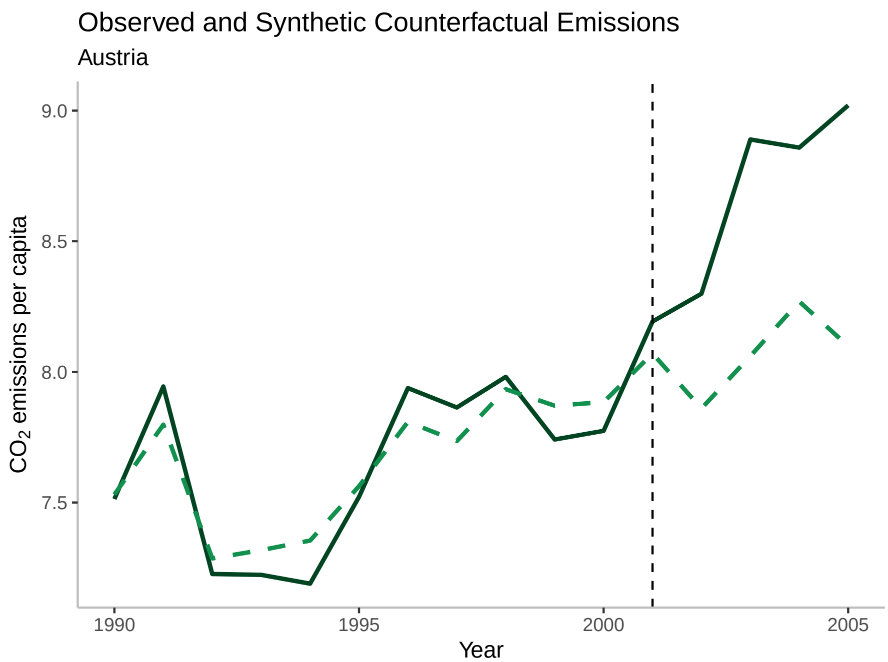


Figure A.13: Observed and synthetic counterfactual emissions for placebo country Austria.

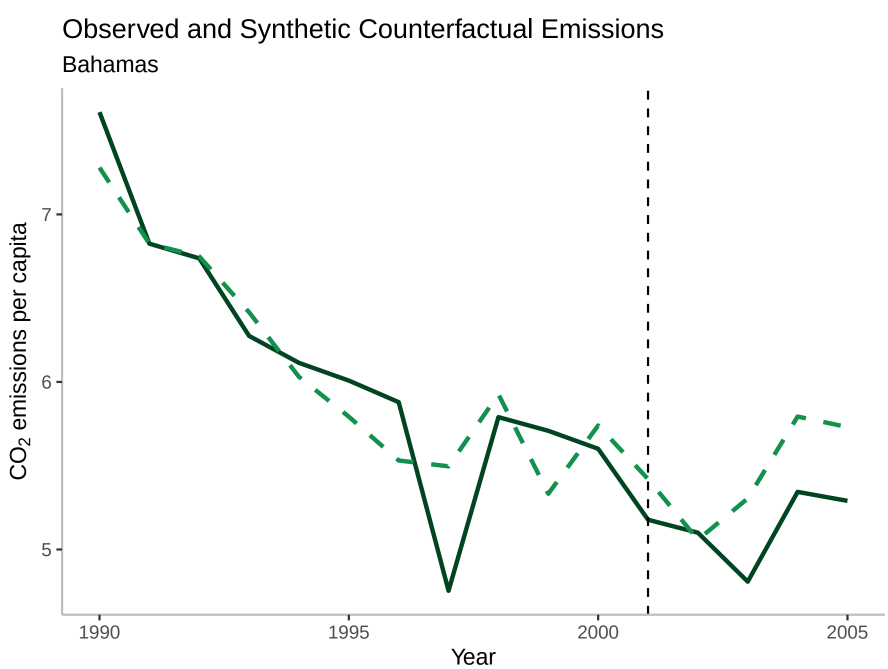


Figure A.14: Observed and synthetic counterfactual emissions for placebo country Bahamas.

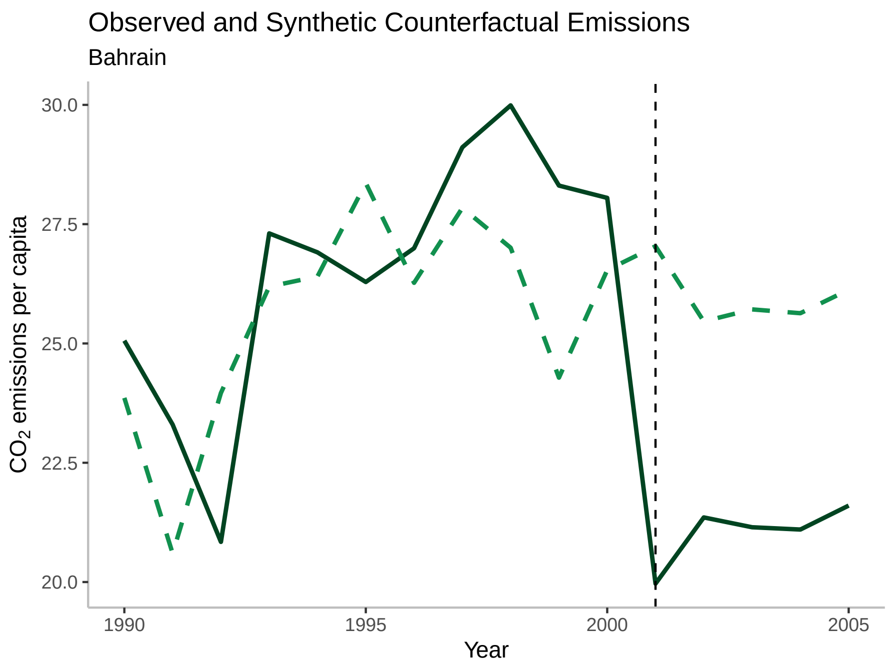


Figure A.15: Observed and synthetic counterfactual emissions for placebo country Bahrain.

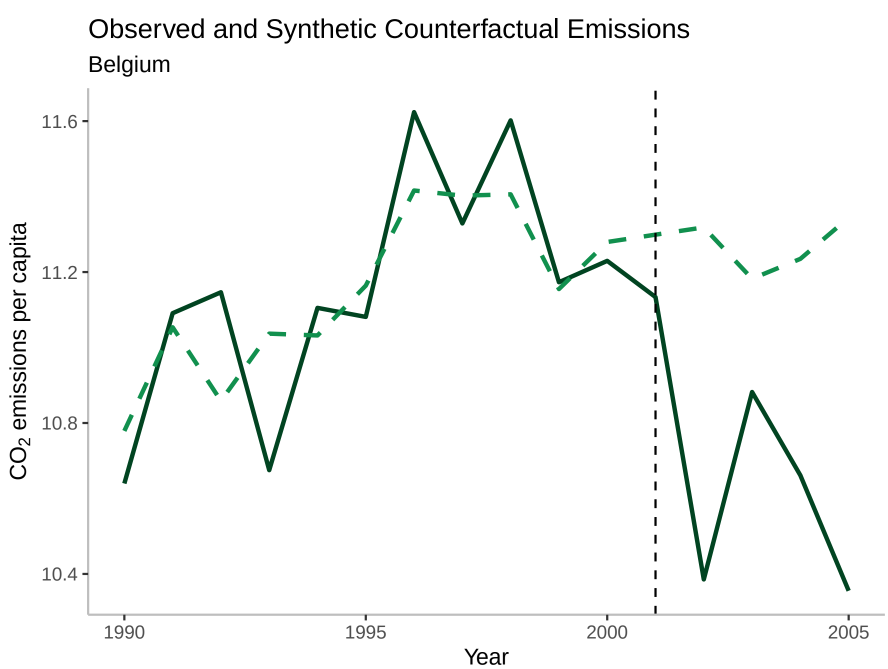


Figure A.16: Observed and synthetic counterfactual emissions for placebo country Belgium.

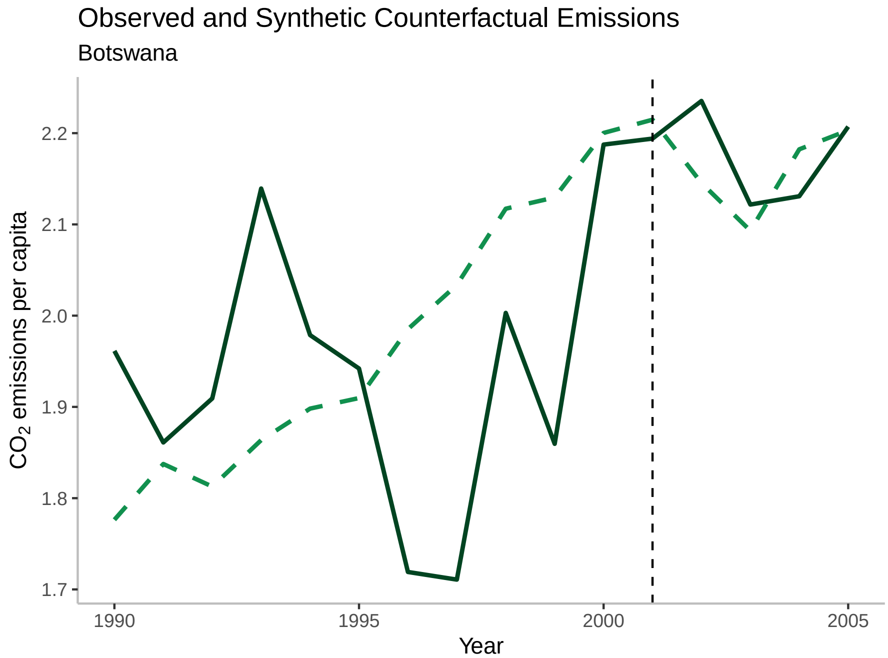


Figure A.17: Observed and synthetic counterfactual emissions for placebo country Botswana.

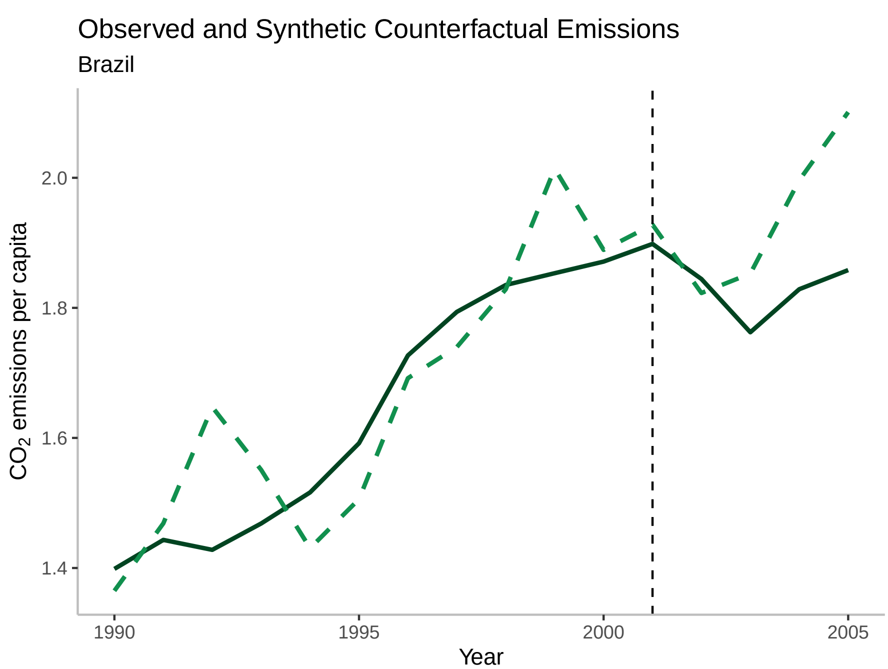


Figure A.18: Observed and synthetic counterfactual emissions for placebo country Brazil.

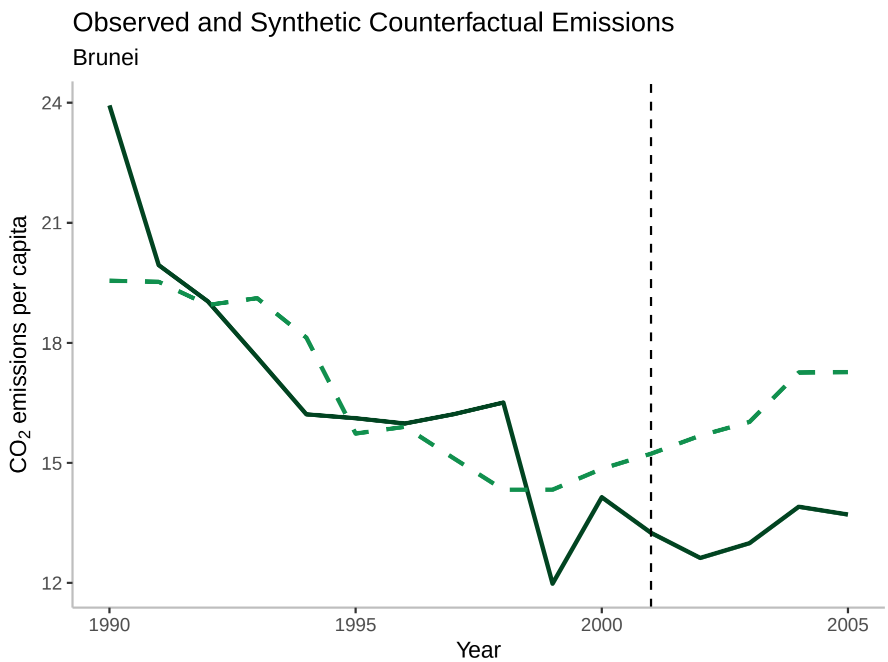


Figure A.19: Observed and synthetic counterfactual emissions for placebo country Brunei.

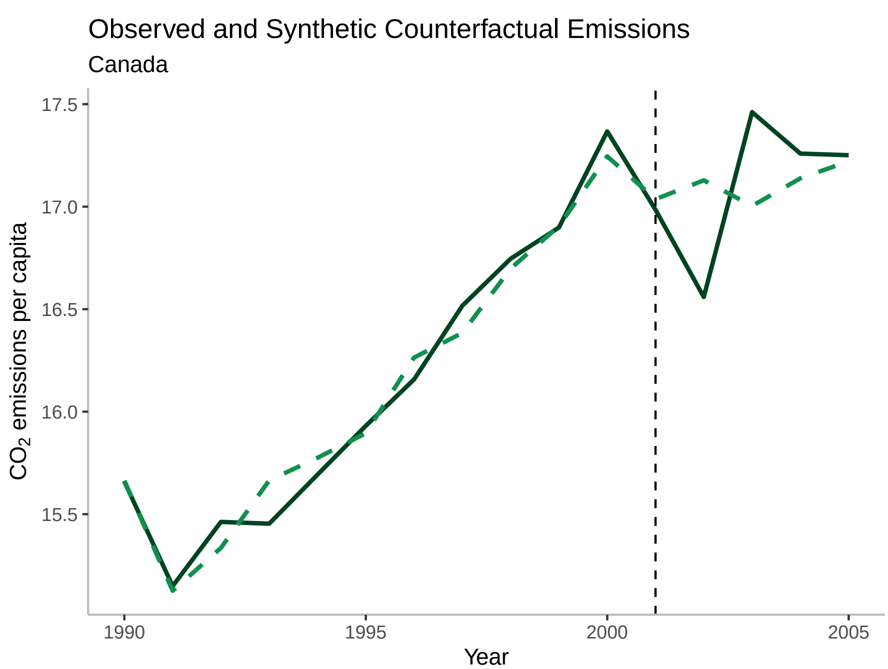


Figure A.20: Observed and synthetic counterfactual emissions for placebo country Canada.

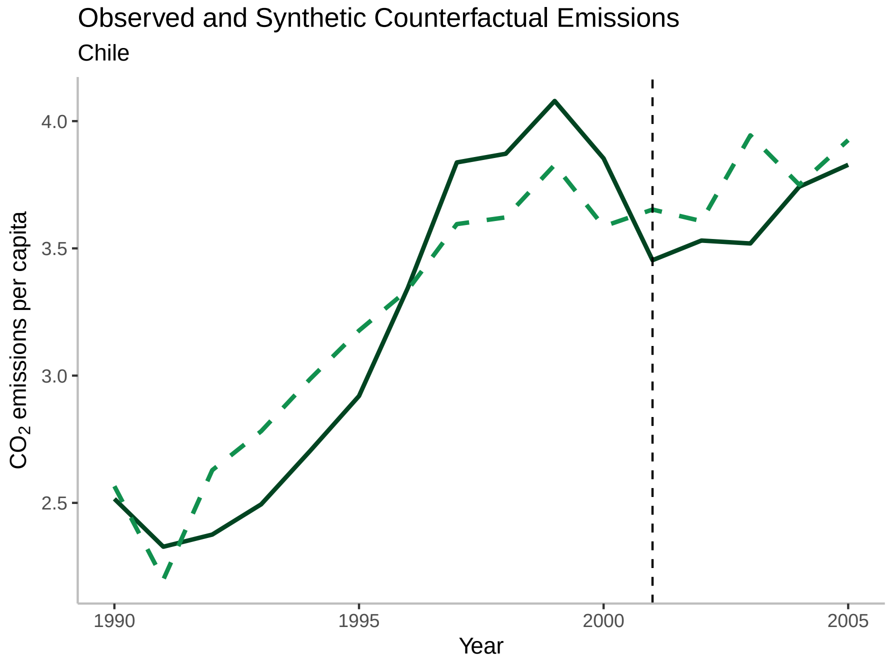


Figure A.21: Observed and synthetic counterfactual emissions for placebo country Chile.

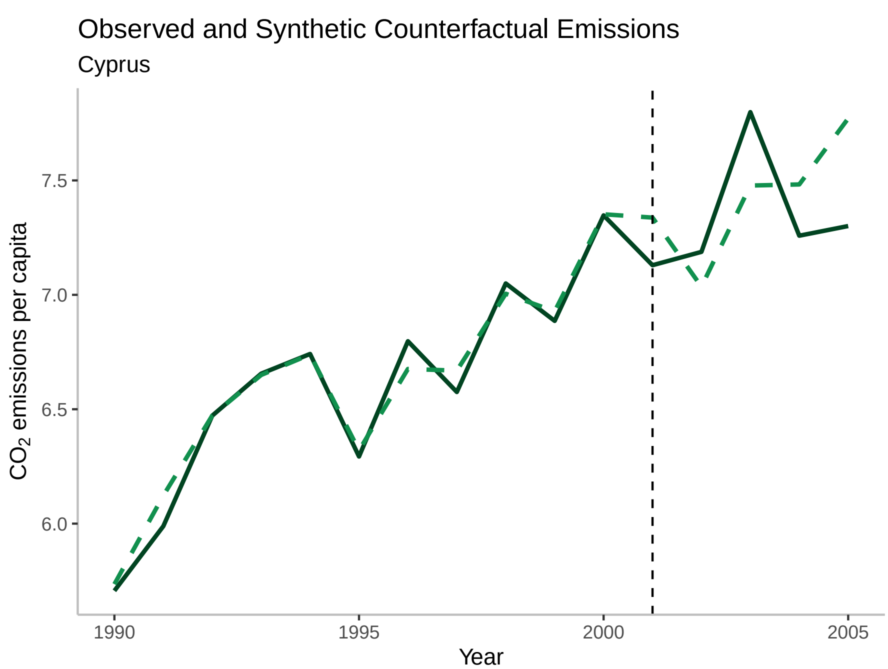


Figure A.22: Observed and synthetic counterfactual emissions for placebo country Cyprus.

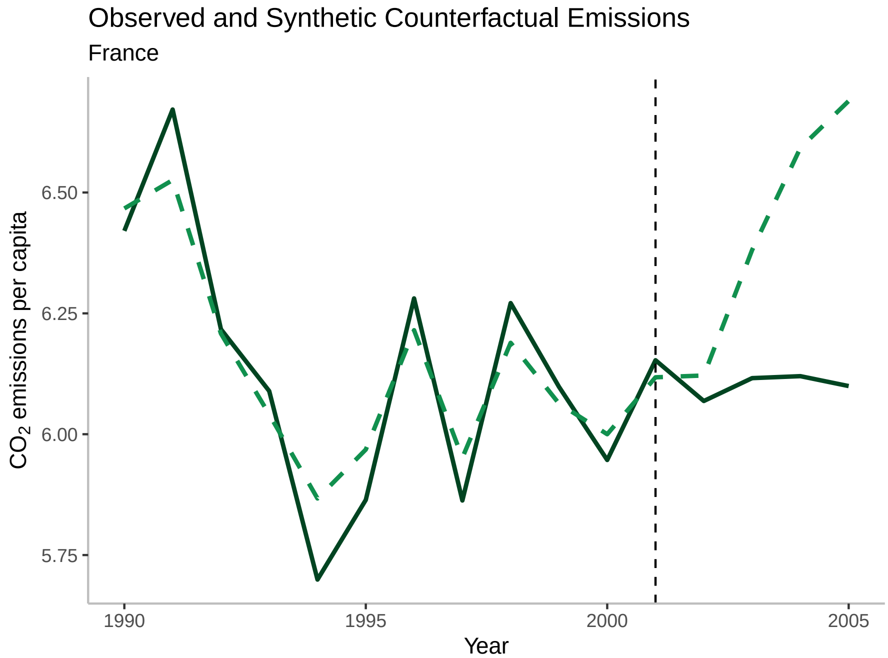


Figure A.23: Observed and synthetic counterfactual emissions for placebo country France.

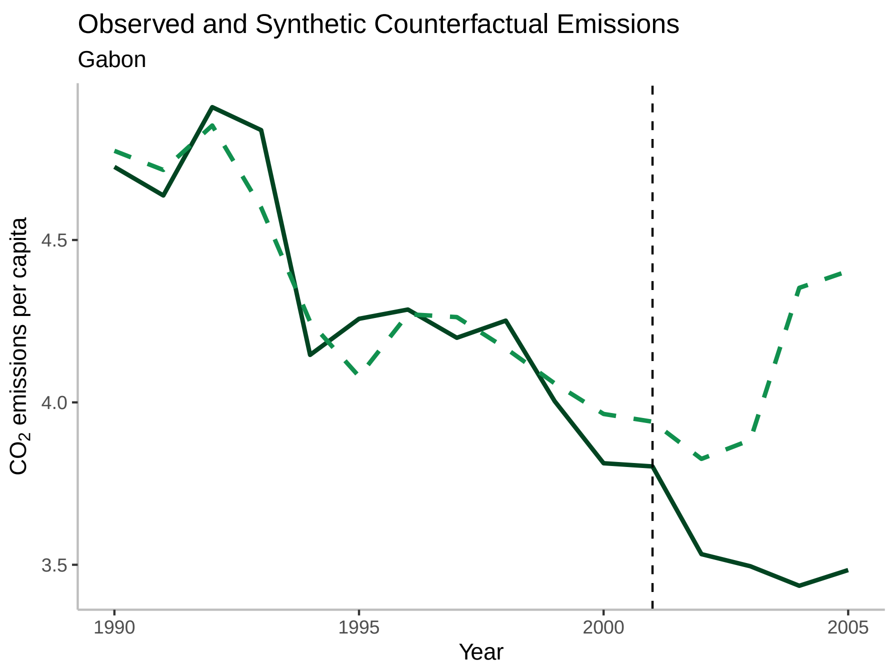


Figure A.24: Observed and synthetic counterfactual emissions for placebo country Gabon.

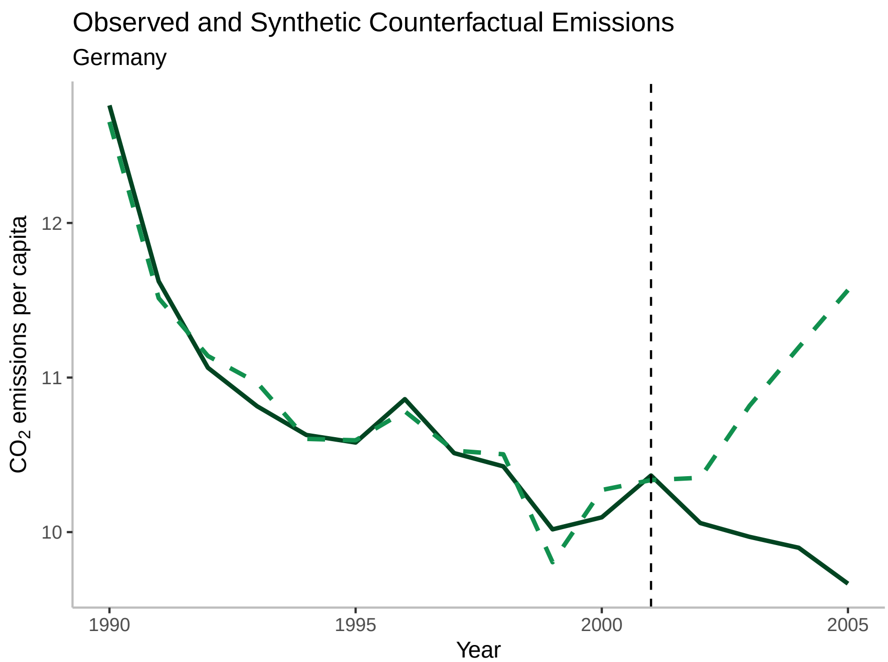


Figure A.25: Observed and synthetic counterfactual emissions for placebo country Germany.

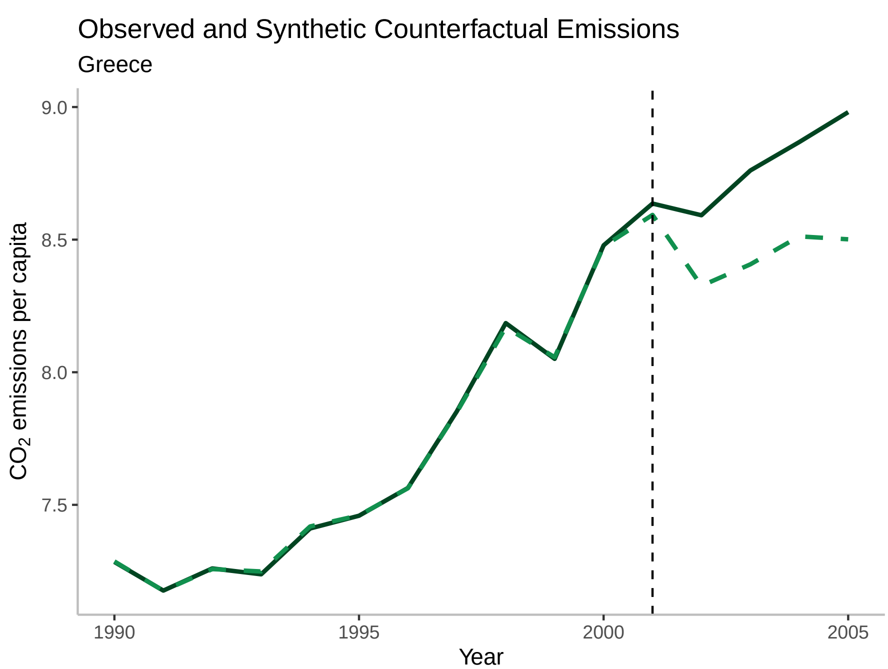


Figure A.26: Observed and synthetic counterfactual emissions for placebo country Greece.

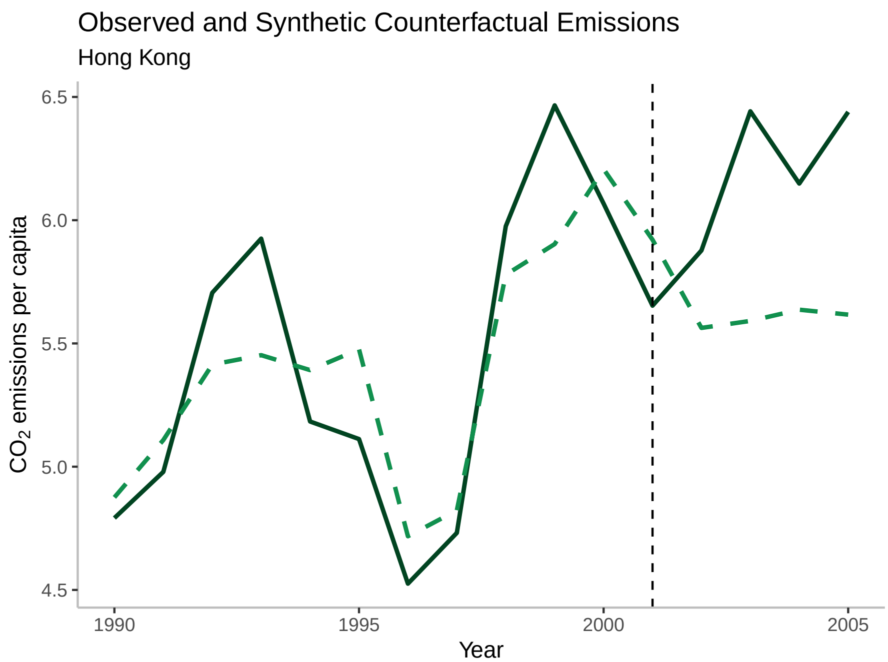


Figure A.27: Observed and synthetic counterfactual emissions for placebo country Hong Kong.

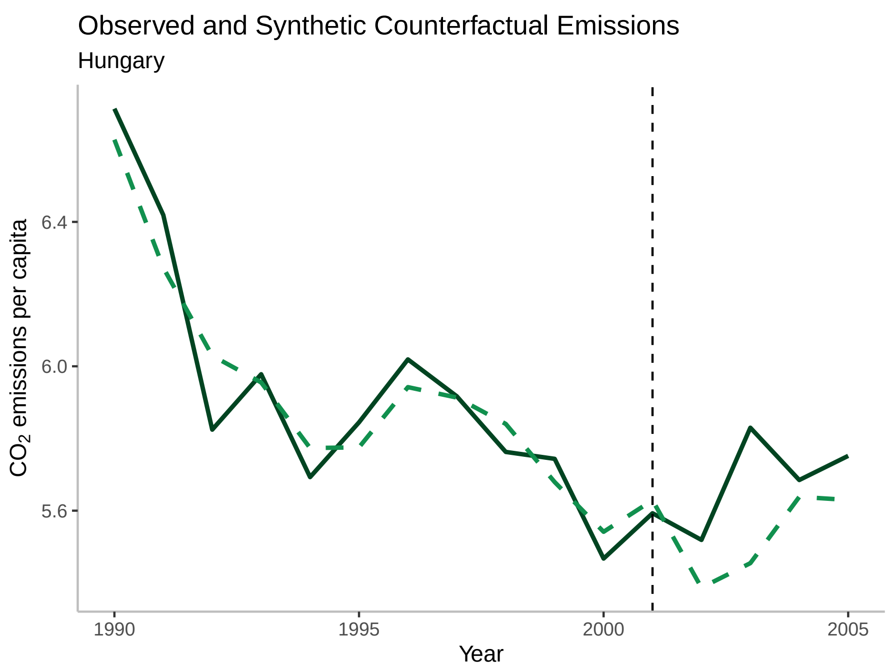


Figure A.28: Observed and synthetic counterfactual emissions for placebo country Hungary.

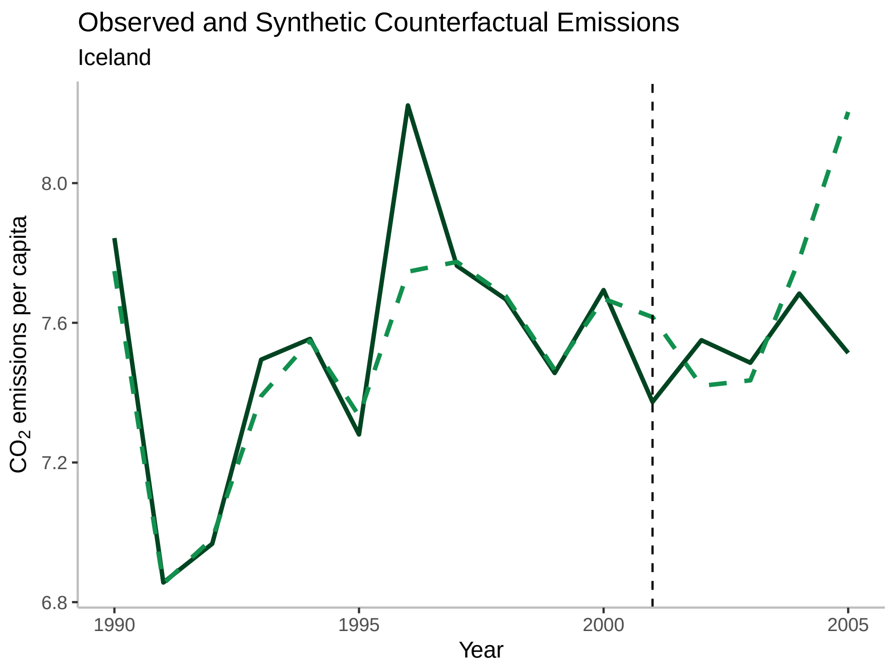


Figure A.29: Observed and synthetic counterfactual emissions for placebo country Iceland.

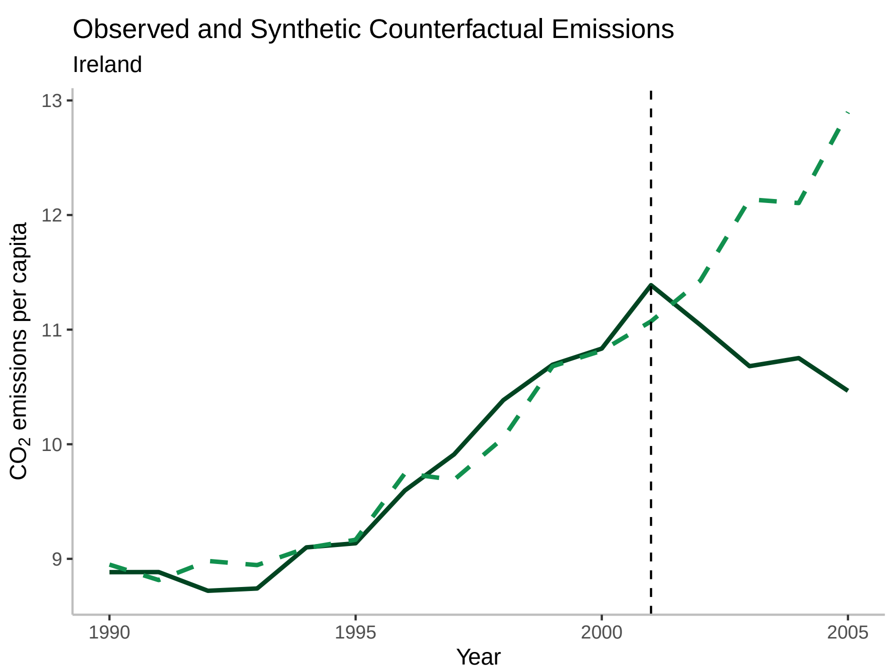


Figure A.30: Observed and synthetic counterfactual emissions for placebo country Ireland.

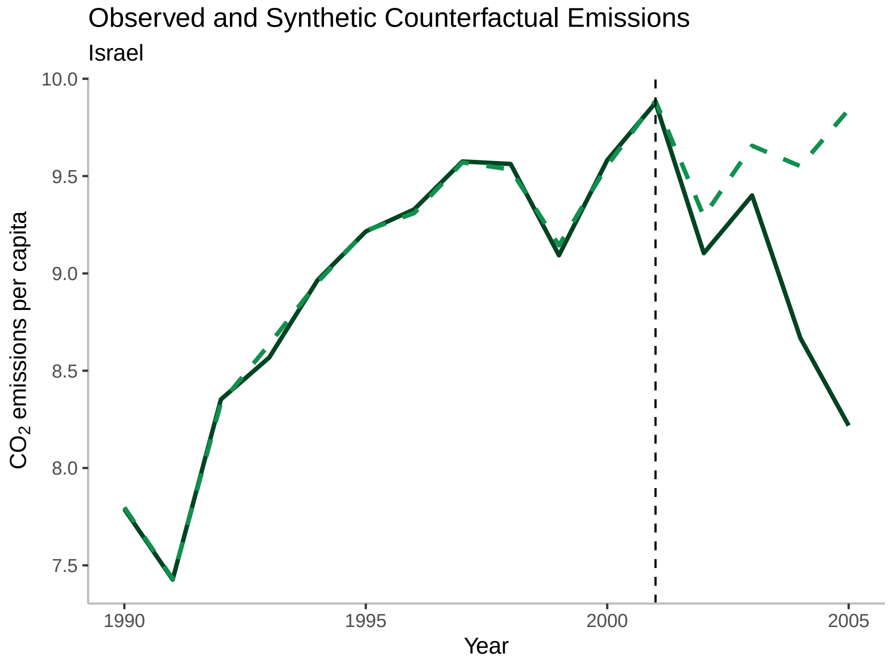


Figure A.31: Observed and synthetic counterfactual emissions for placebo country Israel.

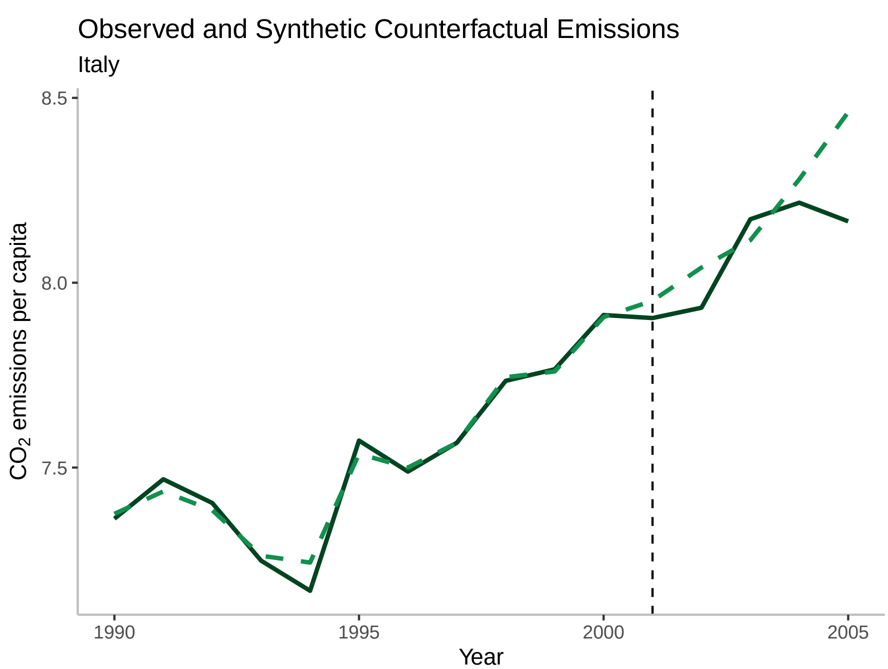


Figure A.32: Observed and synthetic counterfactual emissions for placebo country Italy.

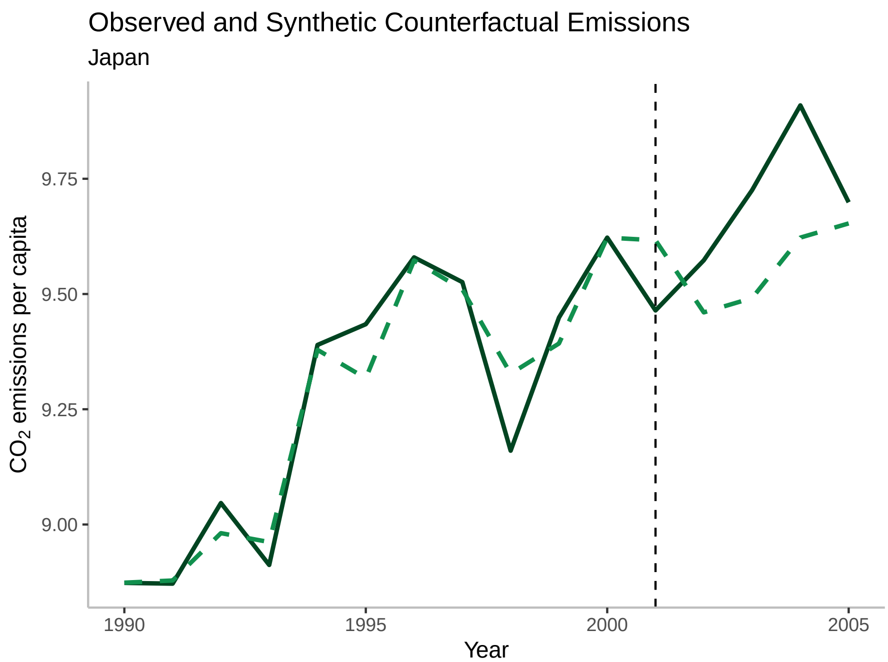


Figure A.33: Observed and synthetic counterfactual emissions for placebo country Japan.

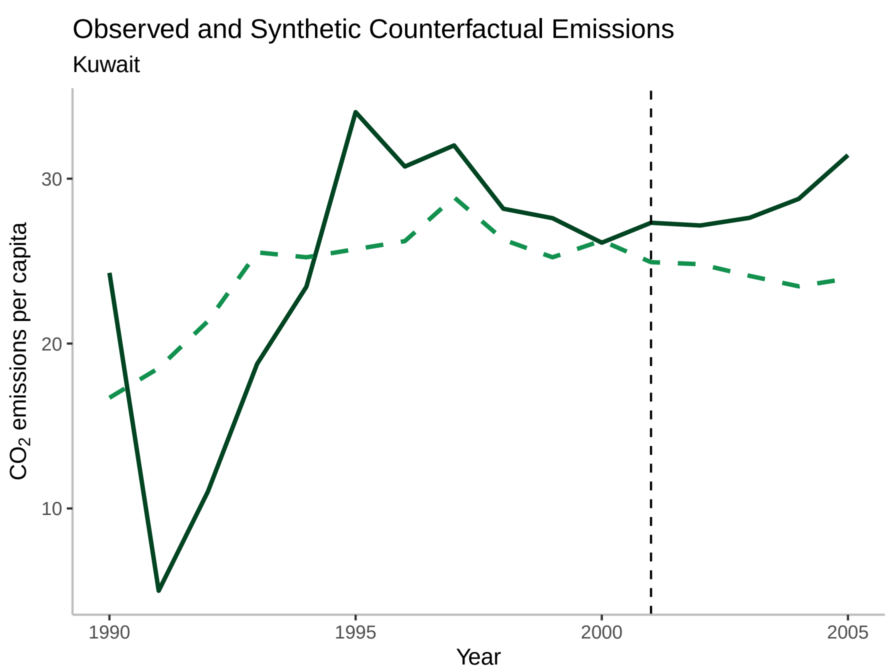


Figure A.34: Observed and synthetic counterfactual emissions for placebo country Kuwait.

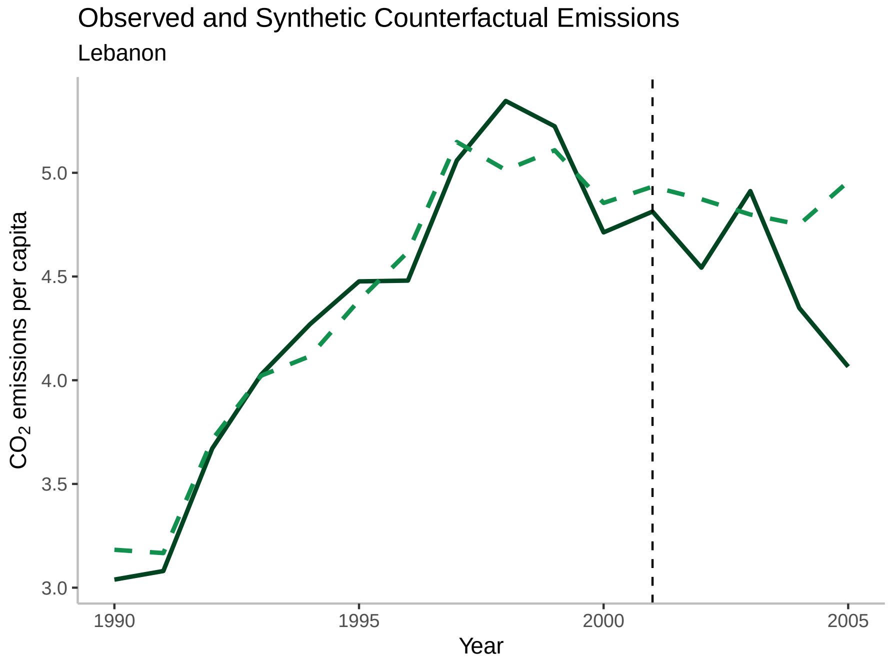


Figure A.35: Observed and synthetic counterfactual emissions for placebo country Lebanon.

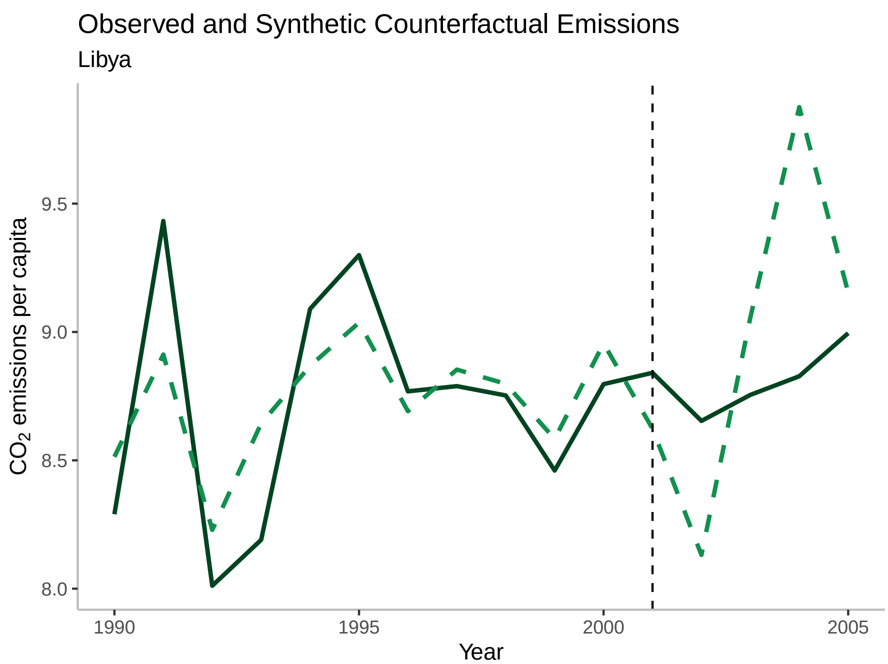


Figure A.36: Observed and synthetic counterfactual emissions for placebo country Libya.

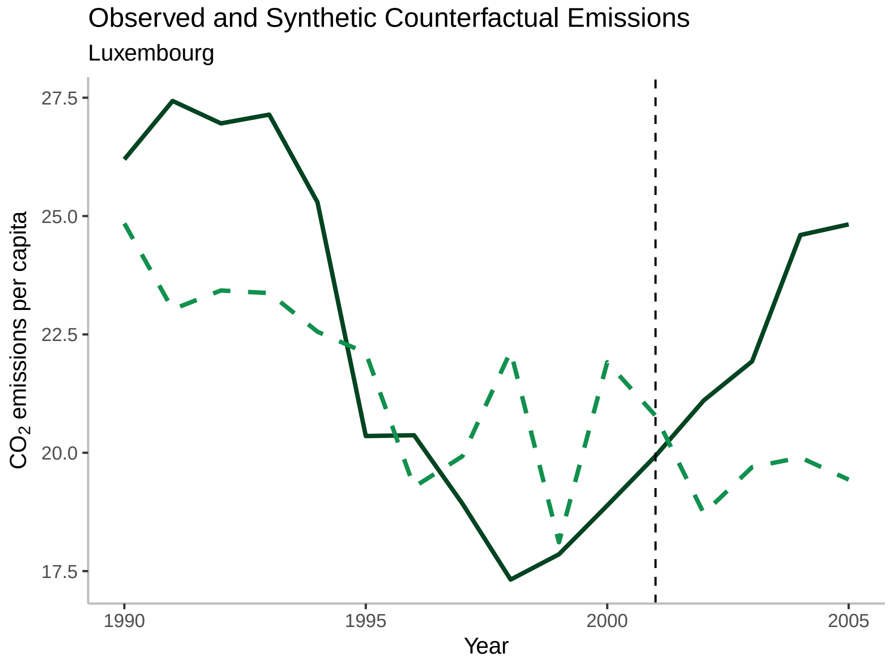


Figure A.37: Observed and synthetic counterfactual emissions for placebo country Luxembourg.

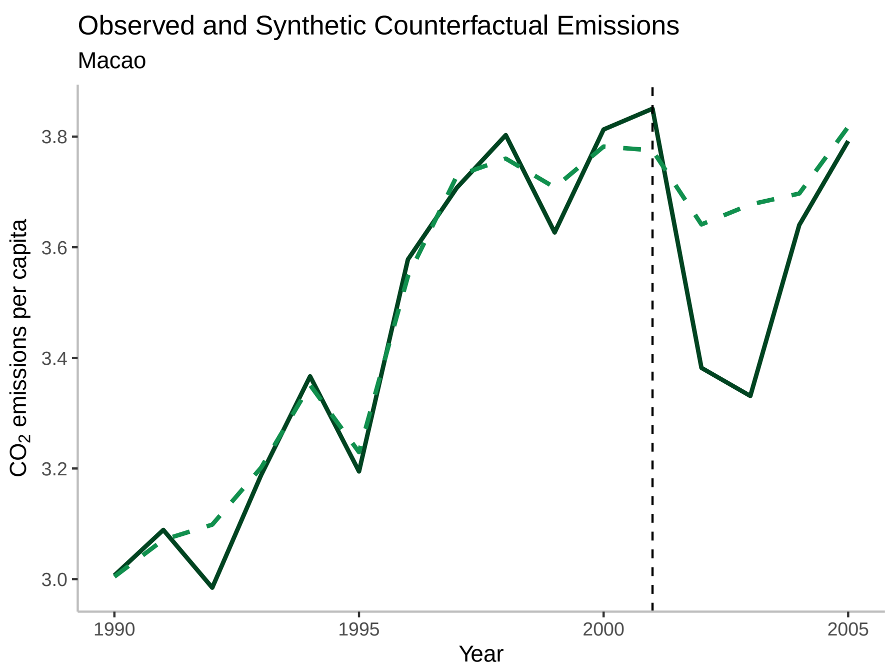


Figure A.38: Observed and synthetic counterfactual emissions for placebo country Macao.

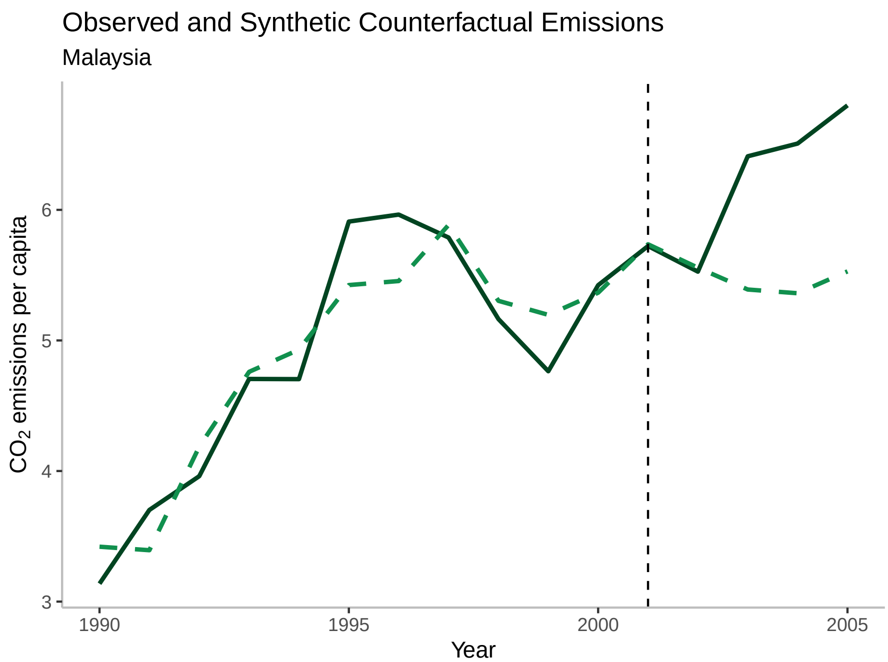


Figure A.39: Observed and synthetic counterfactual emissions for placebo country Malaysia.

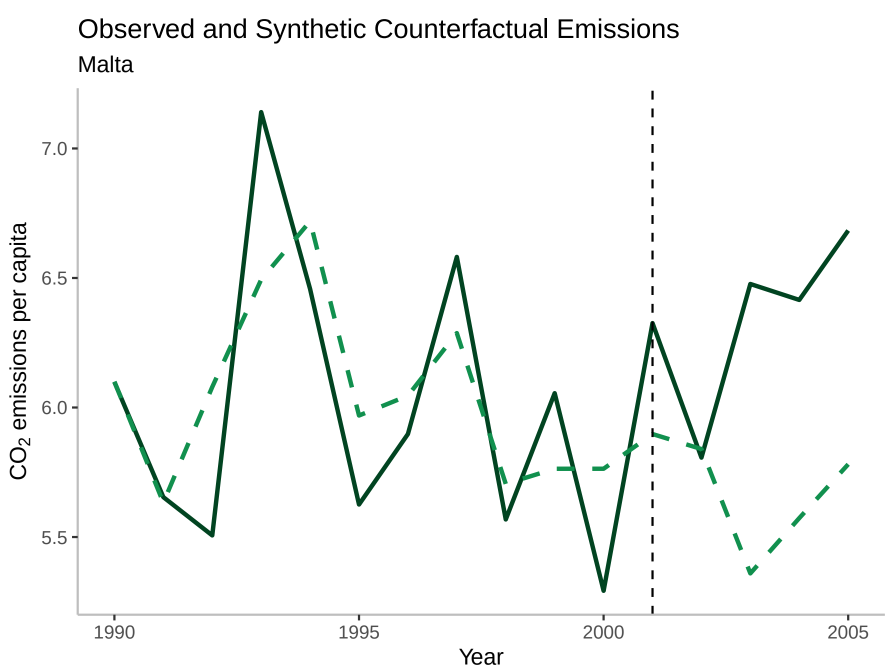


Figure A.40: Observed and synthetic counterfactual emissions for placebo country Malta.

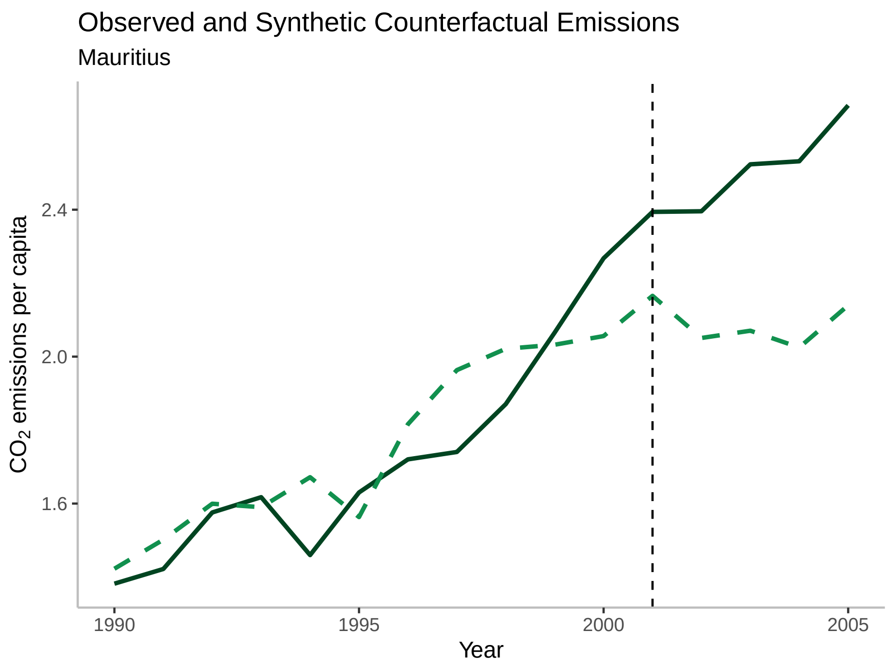


Figure A.41: Observed and synthetic counterfactual emissions for placebo country Mauritius.

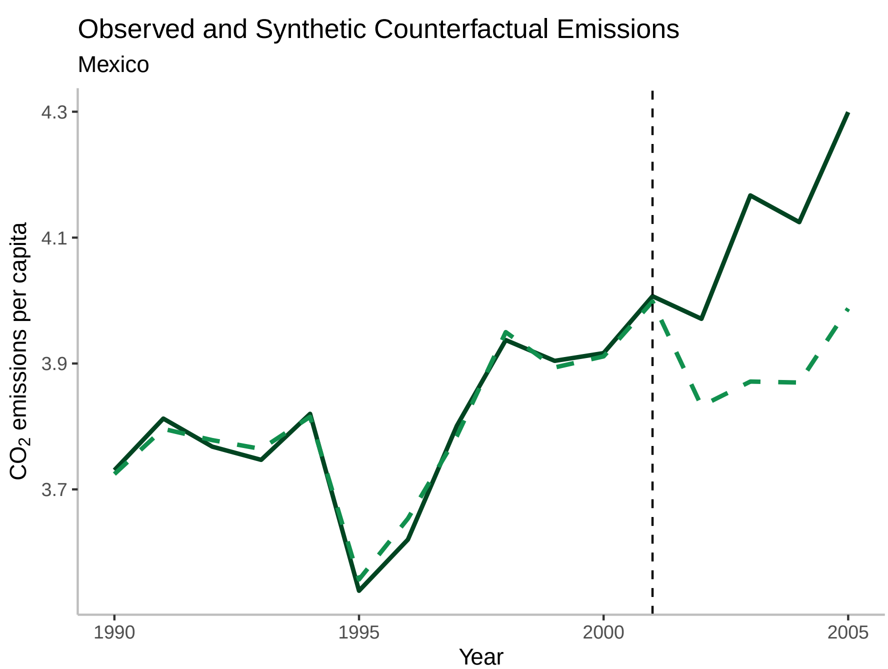


Figure A.42: Observed and synthetic counterfactual emissions for placebo country Mexico.

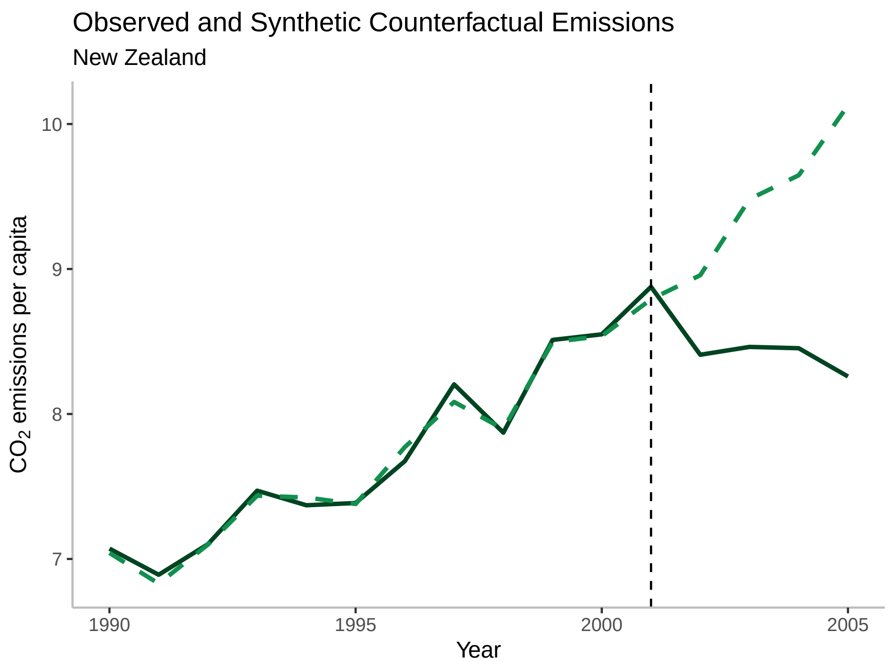


Figure A.43: Observed and synthetic counterfactual emissions for placebo country New Zealand.

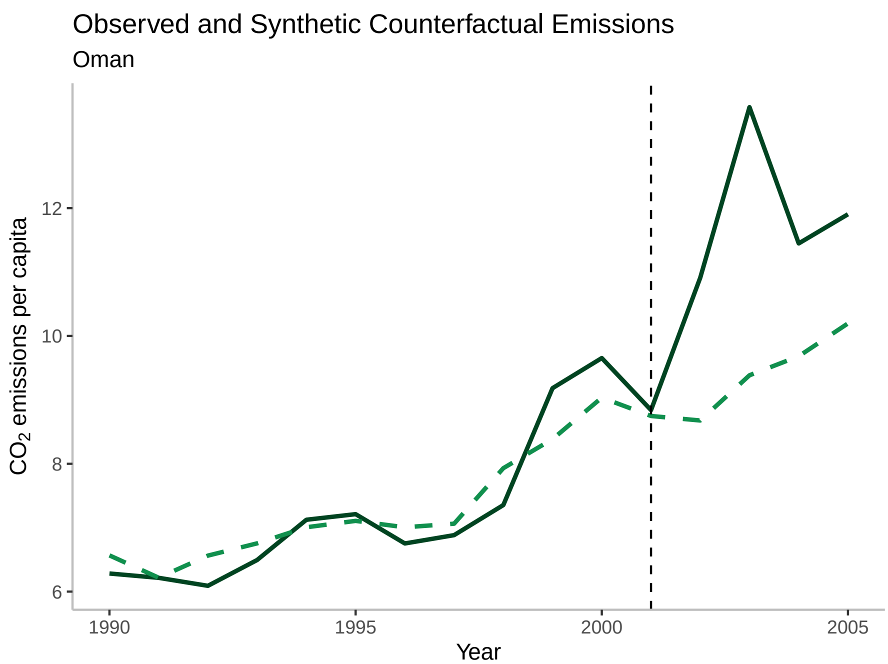


Figure A.44: Observed and synthetic counterfactual emissions for placebo country Oman.

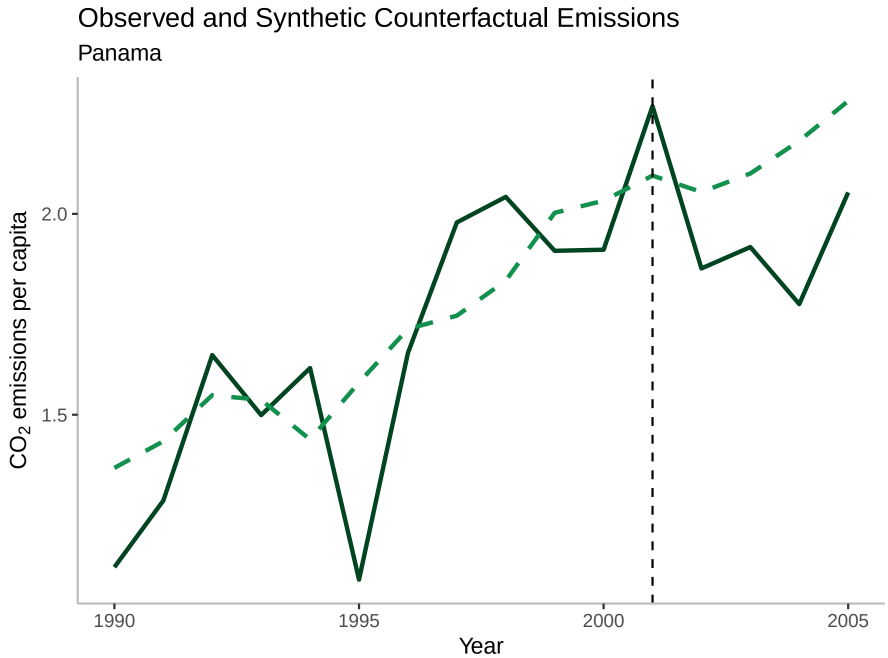


Figure A.45: Observed and synthetic counterfactual emissions for placebo country Panama.

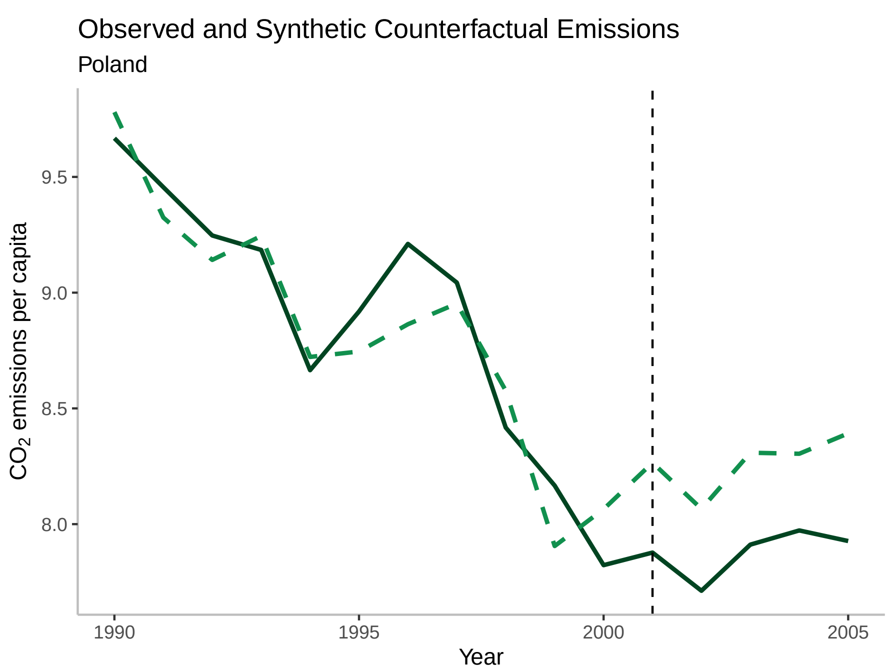


Figure A.46: Observed and synthetic counterfactual emissions for placebo country Poland.

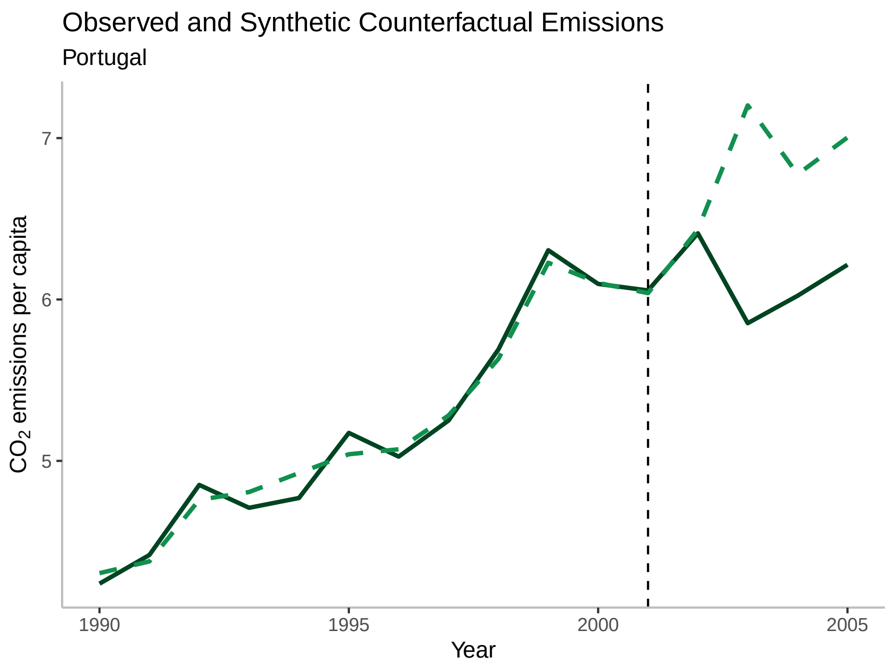


Figure A.47: Observed and synthetic counterfactual emissions for placebo country Portugal.

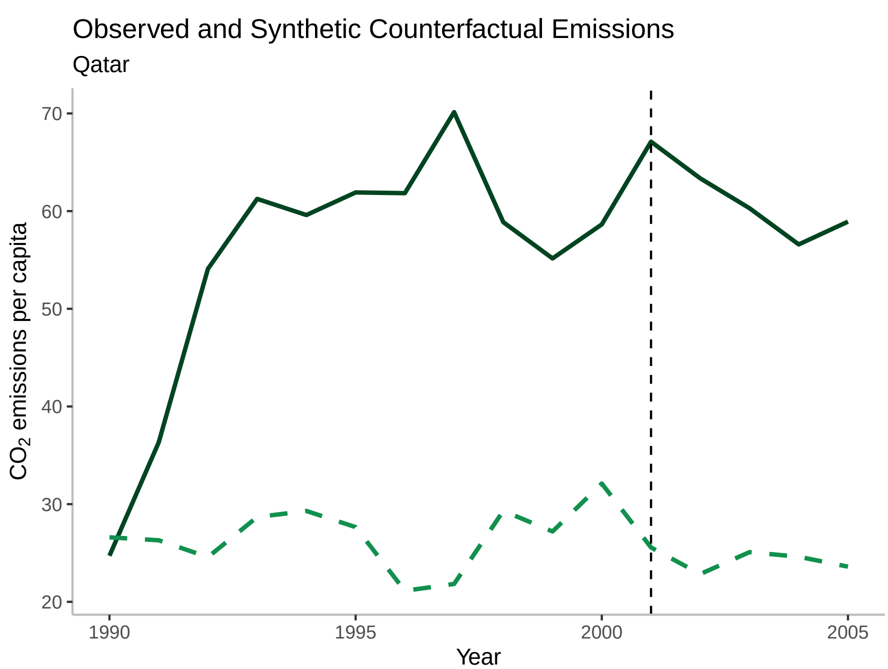


Figure A.48: Observed and synthetic counterfactual emissions for placebo country Qatar.

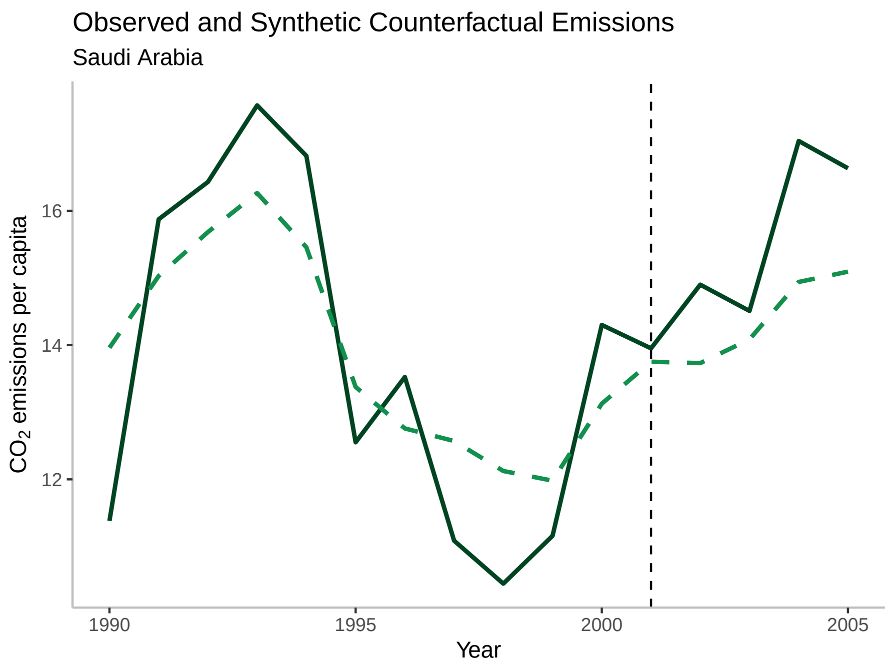


Figure A.49: Observed and synthetic counterfactual emissions for placebo country Saudi Arabia.

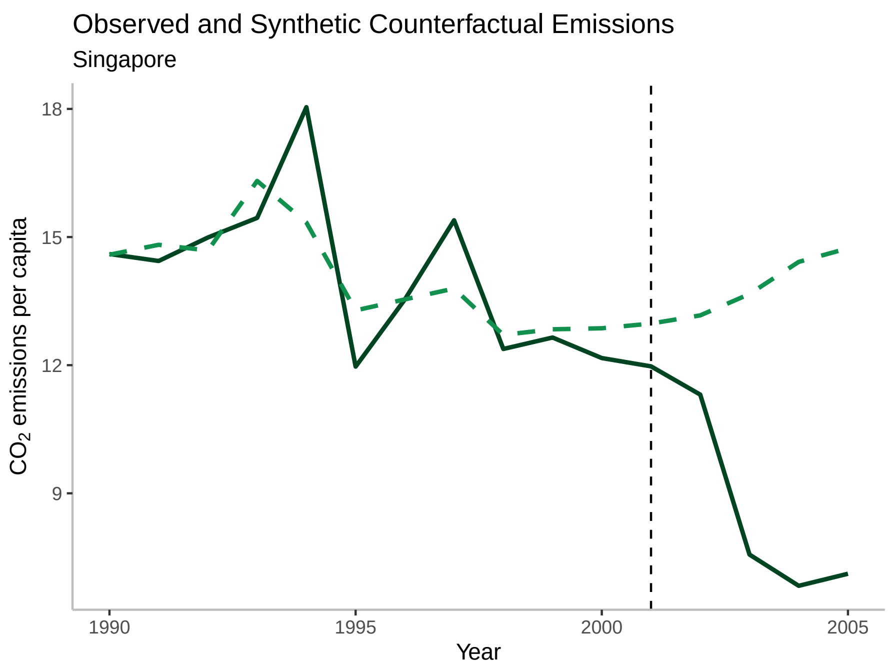


Figure A.50: Observed and synthetic counterfactual emissions for placebo country Singapore.

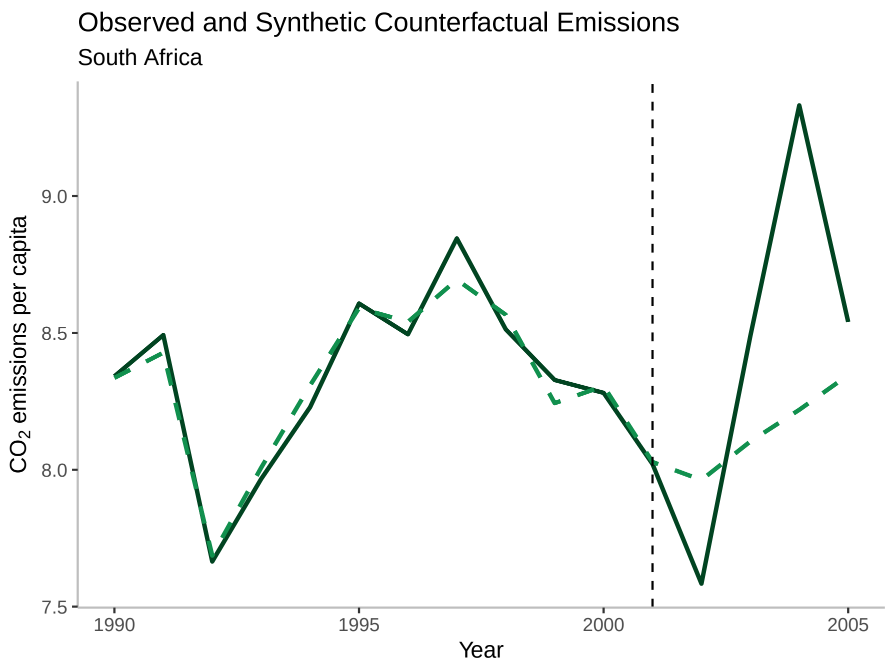


Figure A.51: Observed and synthetic counterfactual emissions for placebo country South Africa.

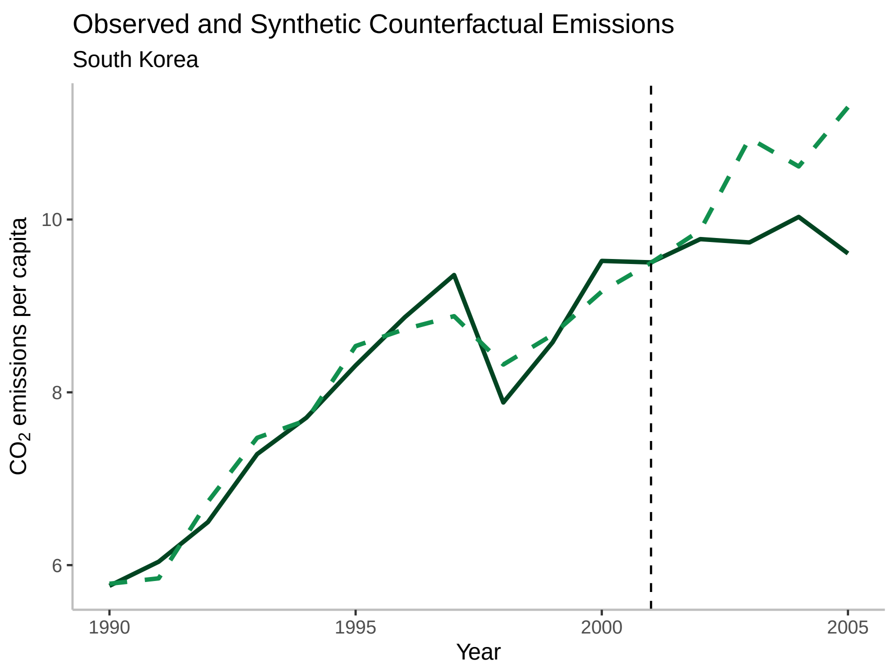


Figure A.52: Observed and synthetic counterfactual emissions for placebo country South Korea.

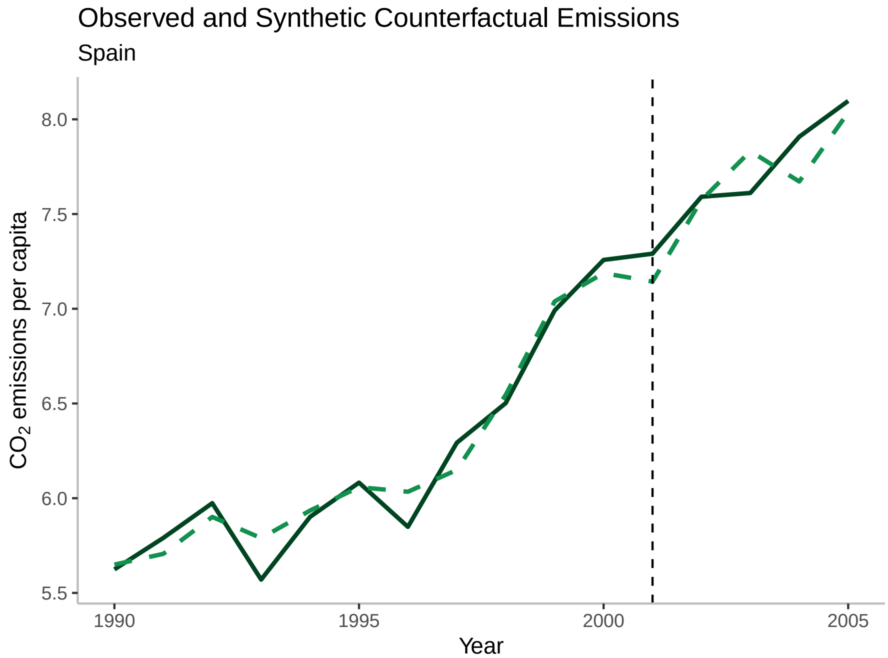


Figure A.53: Observed and synthetic counterfactual emissions for placebo country Spain.

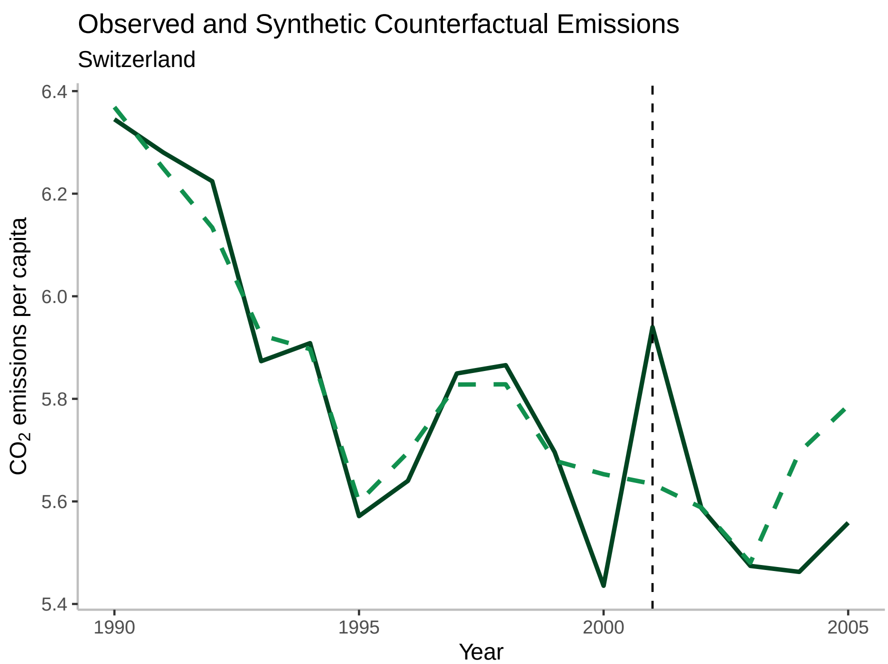


Figure A.54: Observed and synthetic counterfactual emissions for placebo country Switzerland.

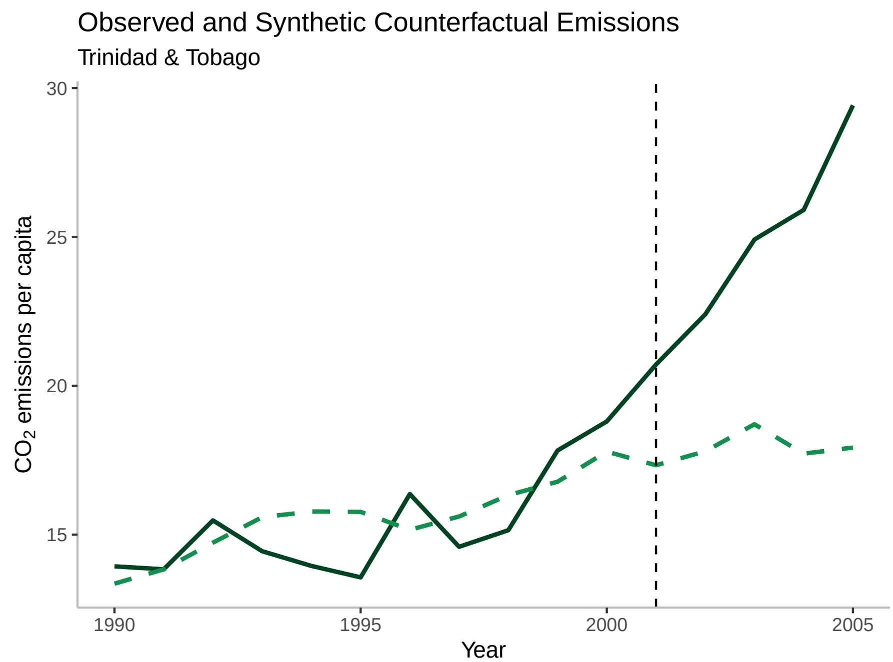


Figure A.55: Observed and synthetic counterfactual emissions for placebo country Trinidad and Tobago.

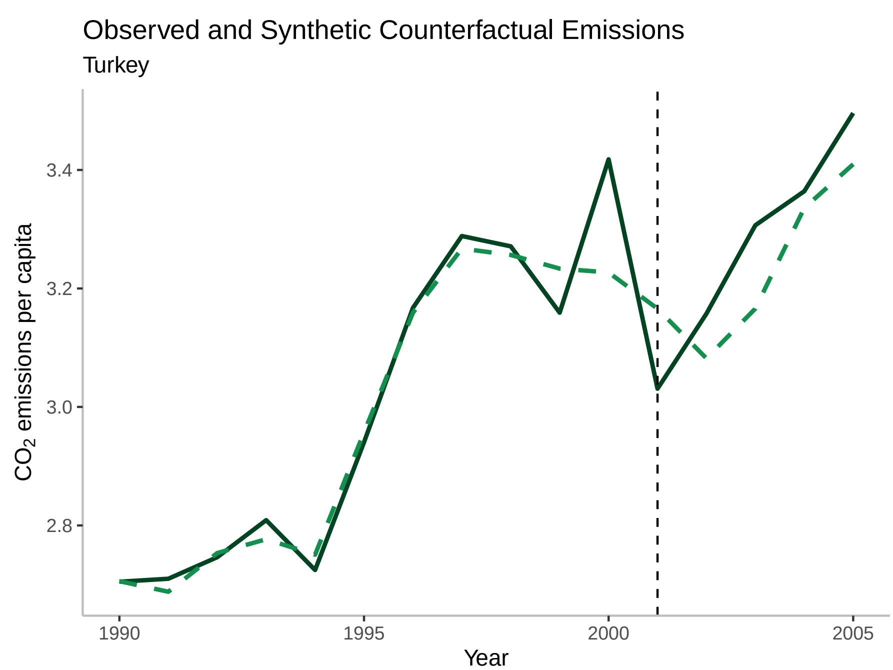


Figure A.56: Observed and synthetic counterfactual emissions for placebo country Turkey.

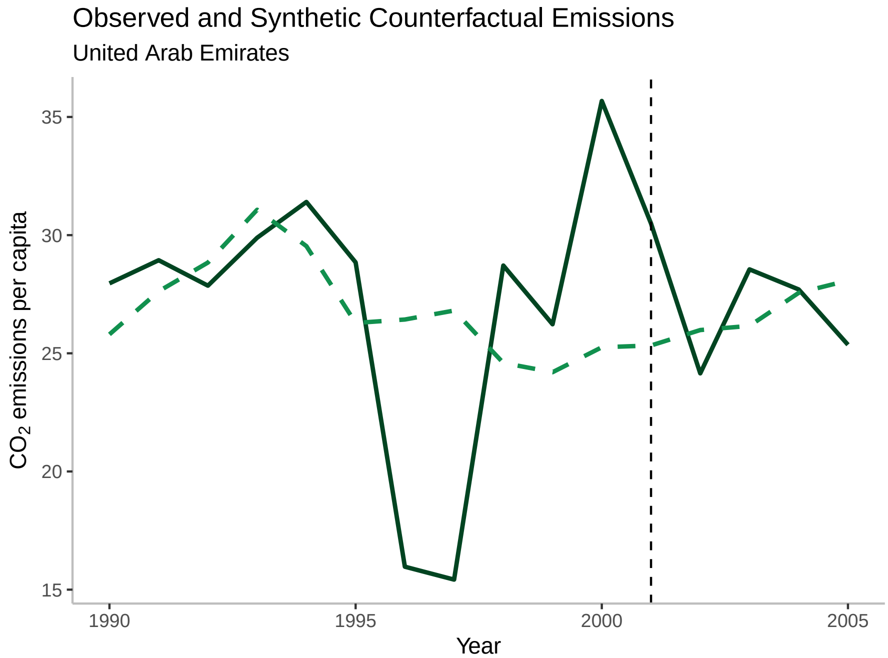


Figure A.57: Observed and synthetic counterfactual emissions for placebo country United Arab Emirates.

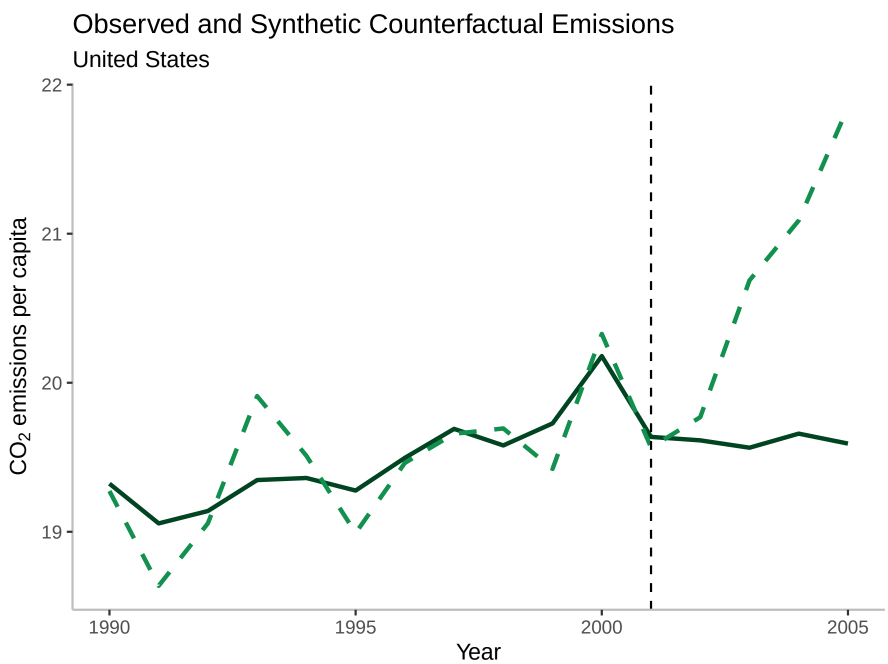


Figure A.58: Observed and synthetic counterfactual emissions for placebo country United States.

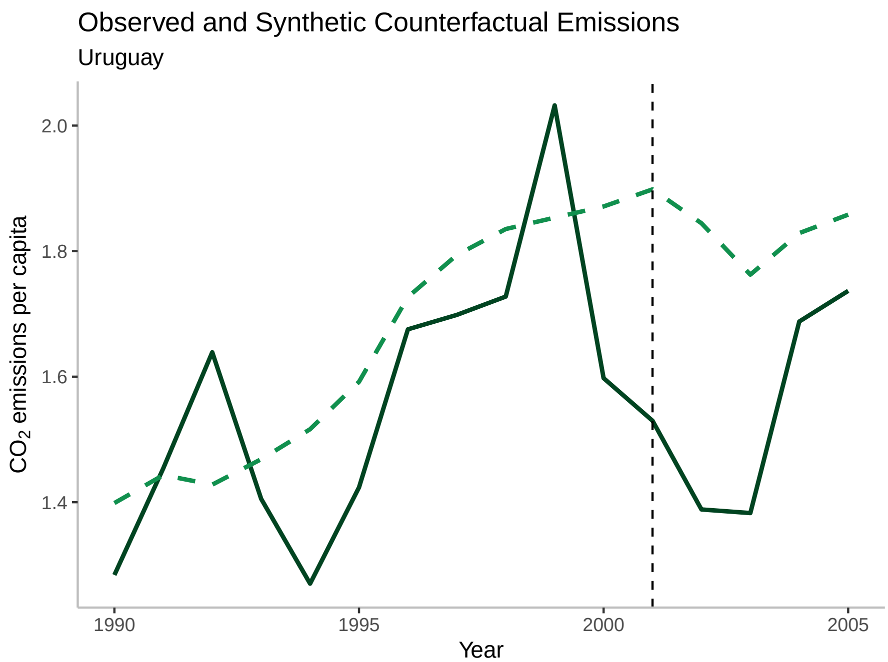


Figure A.59: Observed and synthetic counterfactual emissions for placebo country Uruguay.

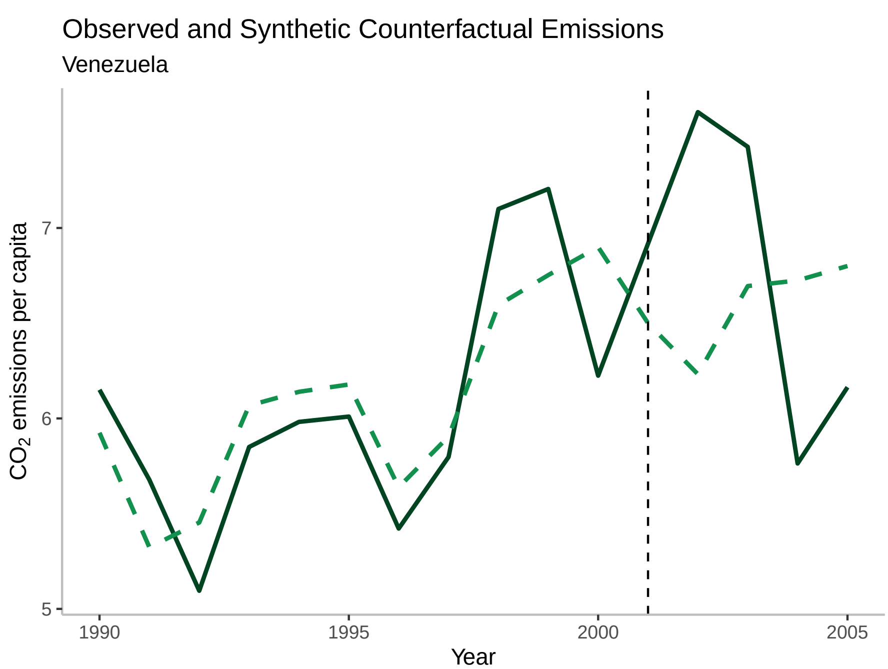


Figure A.60: Observed and synthetic counterfactual emissions for placebo country Venezuela.

A.7 Alternative specifications

A.7.1 Summary of all specifications

Table A.3 displays the weights applied to each country in the donor pool for the alternate specifications. A "-" denotes that the particular country is not included in the donor pool because it does not meet the criteria for inclusion (e.g. being a high income country in 2001), while a "NA" denotes that data is missing for a particular country. The latter case applies only to specification 2 which uses other covariates.

Full results for the alternate specifications 2-11 are provided in sections G.2-G.11. For these alternate specifications, we report: the donor weights, estimated treatment effects, statistical inference, and robustness checks.

Figures A.63, A.69, A.75, A.81, A.87, A.93, A.99, A.105, A.111, and A.117 display the results of the placebo test where treatment is iteratively re-assigned to each country in the donor pool. The thick purple line represents the gaps in emissions between the UK and its synthetic control (as estimated by the alternate specification). The thin grey lines represent the gaps in emissions between placebo countries and their corresponding synthetic counterpart. Only the countries where the pre-treatment MSPE for each placebo is less than 5 times greater than the pre-treatment MSPE of the UK are displayed to avoid graphing unnecessary noise. In each case, the line representing the gaps for the UK is unusually large relative to the gaps of the placebo countries.

Figures A.64, A.70, A.76, A.82, A.88, A.94, A.100, A.106, A.112, and A.118 display the results of the "leave-one-out" robustness check for each alternate specification 2-11. When iteratively dropping one country from the donor pool used to estimate each alternate specification, we see that the magnitude and sign of the gap in emissions between the

UK and its counterpart remains large and negative.

As we did for the main specification, we also look at the empirical distribution of the ratio of post- to pre-treatment MSPE in the UK and in placebo countries in order to test the statistical significance of our findings. We conduct both a two-sided and a one-sided test, where the alternative hypotheses are that the CCP had a non-zero effect on emissions per capita, and that the CCP led to a decrease in emissions per capita, respectively. Given that the stated intention of a climate policy is to reduce emissions, a directional hypothesis is sometimes appropriate, and we thus also report a one-sided test. When we randomly re-assign treatment to all countries in the sample, and restrict our attention to those countries that have a negative treatment effect (i.e., where treatment resulted in a decrease in emissions), we find that the UK has the largest ratio statistic in every alternate specification from 2 to 11. Figures A.65, A.71, A.77, A.83, A.89, A.95, A.101, A.107, A.113, and A.119 display the empirical distribution of this ratio statistic.

Figures A.66, A.72, A.78, A.84, A.90, A.96, A.102, A.108, A.114, and A.120 display the ratio of the post- to pre-treatment MSPE for the UK and all placebo countries for alternate specifications 2-11.

We also compute pseudo p-values by dividing the number of countries for which we observe a ratio statistic at least as large as the UK's by the total number of countries in the sample. We computed a ratio statistic for the UK, and also for each country in the donor pool where a placebo test was run. We should not expect this ratio to be large in the placebo countries (other than by chance alone). Therefore, these non-parametric p-values represent the probability of observing an effect as large as the UK's if the null distribution were true. This probability is less than 0.05 in alternate specifications 4-11, and is barely larger in specifications 2 ($p = 0.053$) and 3 ($p = 0.058$).

Tables A.4, A.6, A.8, A.10, A.12, A.14, A.16, A.18, A.20, and A.22 display summary statistics and the weights applied to the pre-treatment covariates included in alternate specifications 2-11. The covariates included as predictors of the synthetic UK vary with each specification in order to test the robustness of our results. In those tables, column 2 ("Treated UK") displays the observed values of the predictors in the UK prior to treatment in 2001. These pre-treatment means in the UK are significantly different than those of the unweighted sample of donor countries considered for each specification, shown in column 4 ("Sample Mean"). Thus, using an unweighted sample of donor countries as a counterfactual for the pre-CCP emissions of the UK would not identify the causal effect of the treatment. Therefore, we apply the weights given in column 5 ("Weight") in order to yield a trajectory of carbon emissions for the synthetic control. The pre-treatment means of the covariates for the synthetic control are given in column 3 ("Synthetic UK"). Examining these tables makes it clear that the pre-treatment means of the covariates are remarkably similar between those observed in the UK and those simulated by the synthetic UK.

We test this impression formally by conducting a set of statistical balance tests, reported in tables A.5, A.7, A.9, A.11, A.13, A.15, A.17, A.19, A.21, and A.23, for each alternate specification 2-11. Those tables report the balance statistics between the pre-treatment values of the dependent variable (which varies according to specification, i.e., CO_2 emissions per capita, CO_2 emissions rescaled to a 1990, or to a 2000 baseline, in order to assess the robustness of our findings). The first two rows report the p-values for a two-sample t-test of equality of means, and for a Kolmogorov-Smirnov (KS) test for equality of probability distributions, respectively. In both cases, the way to interpret these statistics is to look for a statistically *insignificant* result. That is, we want to fail to reject the null hypothesis that the means of the outcome variable in the UK and the synthetic UK are equal (t-test) or the null hypothesis that the two distributions are equal (KS test). In every alternate specification, the p-values are large, which increases our confidence that the emissions trajectory for the UK and its synthetic control (prior to treatment) are the same, and thus that the algorithm generates an appropriate counterfactual to estimate the causal effect of the CCP.

Donor country	Spec 1	Spec 2	Spec 3	Spec 4	Spec 5	Spec 6	Spec 7	Spec 8	Spec 9	Spec 10	Spec 11
Argentina	0.0050	0.0087	2e-04	-	-	3e-04	-	-	4e-04	-	-
Australia	0.0018	3e-04	0.0867	0.001	7e-04	2e-04	2e-04	5e-05	2e-04	3e-04	2e-05
Austria	0.0044	1e-06	0.0033	0.2742	9e-04	0.0011	0.2256	6e-05	6e-04	0.2409	1e-04
Bahamas	0.1811	NA	1e-05	0.2055	-	0.1493	0.1732	-	0.1419	0.155	-
Bahrain	0.0005	NA	4e-06	-	-	5e-05	-	-	6e-05	-	-
Barbados	0.0009	NA	1e-04	-	-	8e-05	-	-	1e-04	-	-
Belgium	0.1641	0.1394	2e-04	0.1628	0.1232	0.1325	0.0966	0.1557	0.1298	0.0442	0.1406
Botswana	0.0013	0.0024	2e-04	-	-	9e-05	-	-	1e-04	-	-
Brazil	0.0020	7e-04	2e-04	-	-	1e-04	-	-	1e-04	-	-
Brunei	0.0143	NA	0.0015	6e-04	-	0.0303	0.0133	-	0.046	0.0248	-
Canada	0.0026	3e-04	5e-05	0.0016	0.0013	2e-04	3e-04	5e-05	2e-04	5e-04	2e-05
Chile	0.0017	2e-04	1e-04	-	-	8e-05	-	-	9e-05	-	-
Cyprus	0.0022	4e-04	0.0107	5e-04	-	1e-04	2e-04	-	1e-04	2e-04	-
France	0.0011	0.2373	2e-04	0.0072	0.3528	0.0381	0.0019	0.192	0.0169	0.0042	0.1726
Gabon	0.0030	8e-04	6e-05	-	-	2e-04	-	-	1e-04	-	-
Germany	0.0030	0.2073	0.2156	0.0015	0.0991	0.0501	4e-09	0.174	0.0188	0.0011	0.155
Greece	0.0026	3e-04	0.0016	8e-04	0.0012	2e-04	4e-04	4e-05	3e-04	5e-04	2e-05
Hong Kong	0.0014	NA	0.0834	3e-04	-	8e-05	1e-04	-	9e-05	2e-04	-
Hungary	0.0020	NA	1e-04	2e-04	3e-05	7e-06	4e-05	1e-07	4e-07	3e-04	2e-08
Iceland	0.0014	NA	6e-05	4e-04	3e-05	1e-04	2e-04	5e-05	1e-04	1e-04	2e-05
Ireland	0.0023	2e-04	0.001	8e-04	9e-04	2e-04	3e-04	4e-05	2e-04	3e-04	2e-05
Israel	0.0025	4e-04	4e-04	0.001	-	1e-04	2e-04	-	1e-04	3e-04	-
Italy	0.0060	0.1025	0.0023	0.0056	0.0181	0.0012	0.0139	0.0672	0.0021	0.0081	0.1313
Japan	0.0092	0.1785	1e-04	0.1988	0.3293	8e-04	0.181	0.2193	0.0014	0.2908	0.2424
Kuwait	0.0006	NA	1e-05	5e-04	-	5e-05	7e-05	-	1e-05	9e-05	-
Lebanon	0.0016	4e-04	1e-04	-	-	5e-05	-	-	6e-05	-	-

Donor country	Spec 1	Spec 2	Spec 3	Spec 4	Spec 5	Spec 6	Spec 7	Spec 8	Spec 9	Spec 10	Spec 11
Libya	0.1874	NA	0.2342	-	-	0.1394	-	-	0.2024	-	-
Luxembourg	0.0419	0.0354	0.0479	0.0678	0.0659	0.1735	0.1902	0.1896	0.1452	0.1589	0.1577
Macao	0.0017	NA	0.1777	5e-05	-	1e-04	1e-04	-	1e-04	1e-04	-
Malaysia	0.0052	0.0044	3e-04	-	-	8e-05	-	-	3e-07	-	-
Malta	0.0008	NA	9e-05	-	-	6e-05	-	-	4e-05	-	-
Mauritius	0.0021	8e-04	3e-04	-	-	1e-04	-	-	1e-04	-	-
Mexico	0.0021	8e-04	0.0014	3e-05	1e-05	2e-04	2e-04	4e-05	2e-04	3e-04	2e-05
New Zealand	0.0020	2e-04	2e-06	6e-04	4e-04	1e-04	2e-04	2e-05	2e-04	3e-04	1e-05
Oman	0.0022	NA	5e-05	-	-	1e-04	-	-	1e-04	-	-
Panama	0.0016	2e-04	2e-04	-	-	7e-05	-	-	1e-04	-	-
Poland	0.1917	0.0026	0.0967	0.04	1e-04	0.207	0.094	0.0015	0.1742	0.0629	2e-05
Portugal	0.0034	0.0059	0.0016	7e-04	0.0021	2e-04	3e-04	1e-04	2e-04	6e-04	2e-05
Qatar	0.0013	NA	8e-06	8e-04	-	3e-05	6e-05	-	8e-05	5e-04	-
Saudi Arabia	0.0096	0.039	0.023	-	-	2e-04	-	-	8e-04	-	-
Singapore	0.0012	2e-04	3e-05	2e-05	-	1e-04	1e-04	-	1e-04	4e-05	-
South Africa	0.0029	0.0201	7e-05	-	-	2e-04	-	-	2e-04	-	-
South Korea	0.0034	2e-04	9e-05	0.0215	8e-05	1e-04	3e-04	1e-06	8e-05	6e-06	2e-05
Spain	0.0070	8e-04	3e-04	0.0015	0.0027	0.0137	0.0053	4e-06	0.0119	0.0025	4e-05
Switzerland	0.0043	3e-04	3e-04	7e-05	2e-05	9e-04	5e-04	7e-05	0.001	4e-04	1e-05
Trinidad & Tobago	0.0577	3e-04	0.0067	-	-	0.0454	-	-	0.0801	-	-
Turkey	0.0025	3e-04	5e-04	1e-04	4e-09	2e-04	3e-04	2e-05	2e-04	8e-04	2e-05
Utd. Arab Emirates	0.0006	NA	5e-08	0.0014	-	9e-05	3e-04	-	1e-04	3e-04	-
United States	0.0025	4e-04	9e-05	0.0021	0.0013	2e-04	4e-04	6e-05	2e-04	5e-04	3e-05
Uruguay	0.0452	0.0019	3e-04	-	-	0.0121	-	-	0.0224	-	-
Venezuela	0.0009	0.0065	7e-05	-	-	6e-05	-	-	5e-05	-	-

Table A.3: Weights applied to countries in the donor pool in all specifications.

A.7.2 Specification 2

Outcome variable: CO_2 emissions per capita

Donor pool: OECD, high, and upper middle income countries in 2001, $n = 37$

Covariates: Yes

Optimization period: 1990-2001

Predictor	Treated UK	Synthetic UK	Sample Mean	Weight
GDP per capita (constant 2010 US$)	31560.917	35379.844	22297.98	0
Renewable energy consumption (% of total)	0.894	5.613	15.96	0
Fossil fuel energy consumption (% of total)	88.158	77.424	80.924	0
Energy use (kg of oil equivalent per capita)	3745.566	4134.615	3109.721	0
1990 emissions per capita	9.711	9.722	7.66	0.095
1991 emissions per capita	9.871	9.842	7.794	0.101
1992 emissions per capita	9.661	9.641	7.907	0.092
1993 emissions per capita	9.455	9.517	7.983	0.106
1994 emissions per capita	9.448	9.433	8.011	0.089
1995 emissions per capita	9.275	9.184	7.698	0.075
1996 emissions per capita	9.48	9.47	8.01	0.087
1997 emissions per capita	9.043	9.121	8.018	0.075
1998 emissions per capita	9.094	9.109	7.944	0.069
1999 emissions per capita	9.048	9.026	8.079	0.081
2000 emissions per capita	9.2	9.213	8.263	0.129

Table A.4: Pre-treatment values of the predictor variables in the UK (column 2), its synthetic control (column 3), and in the unweighted sample (column 4); and weights applied to those predictor variables (column 5).

Balance statistic	
p-value two sample t-test	0.9668475
p-value Kolmogorov Smirnov test	0.9984853
Mean difference in QQ plots	0.0486111
Median difference in QQ plots	0.0833333
Maximum difference in QQ plots	0.1666667

Table A.5: Balance statistics between pre-treatment values of the dependent variable in the UK and its synthetic counterpart. p-values are reported for a two sample t-test for a difference in means and for a Kolmogorov-Smirnov test for whether the samples come from different distributions. QQ statistics are reported for the empirical CDF of both samples.

Treatment effect:

- 94 Mt CO_2 abated between 2002-2005
- 0.39 tons of CO_2 per capita abated between 2002-2005
- -4.4% in 2005 compared to what emissions would have been *without* the CCP

Statistical significance:

- Two-sided test: $2/38 \approx 0.053$
- One-sided test: $1/26 \approx 0.038$

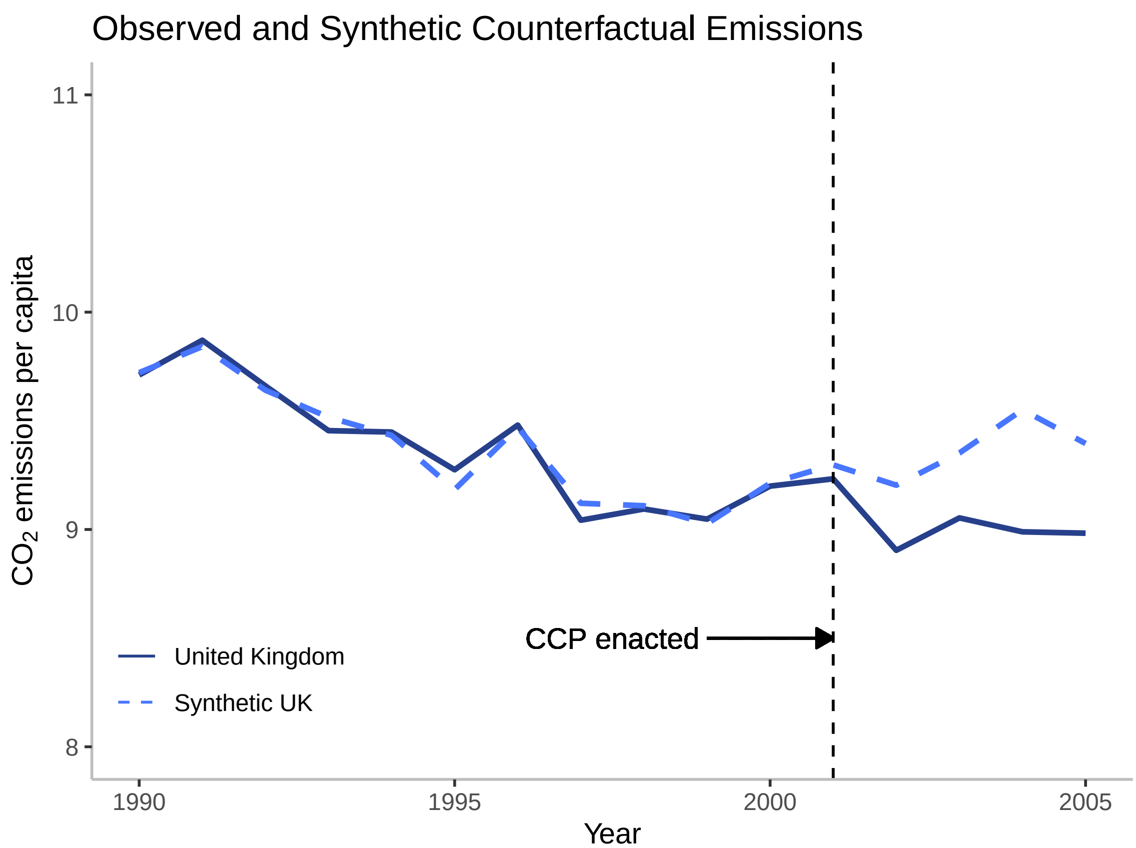


Figure A.61: Observed and synthetic counterfactual emissions for the UK. The dashed line represents the emissions trajectory of a synthetic UK as estimated by Specification 2.

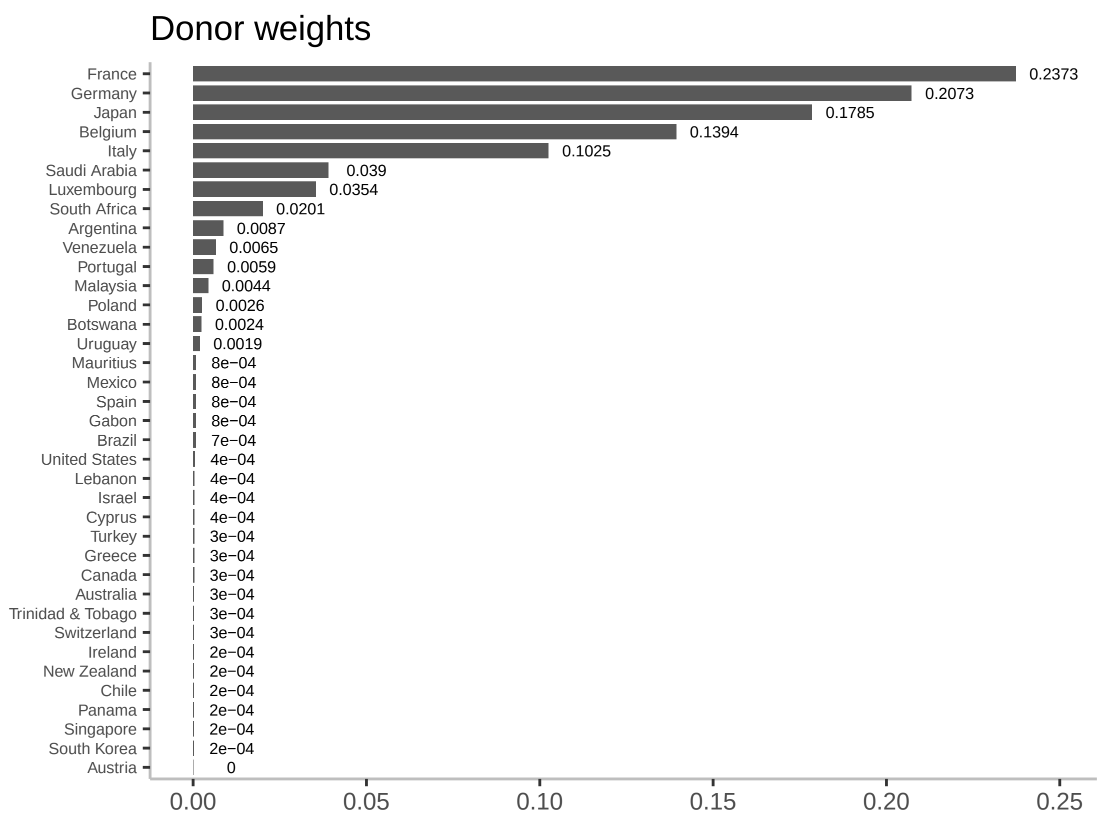


Figure A.62: Weights applied to donor countries in Specification 2.

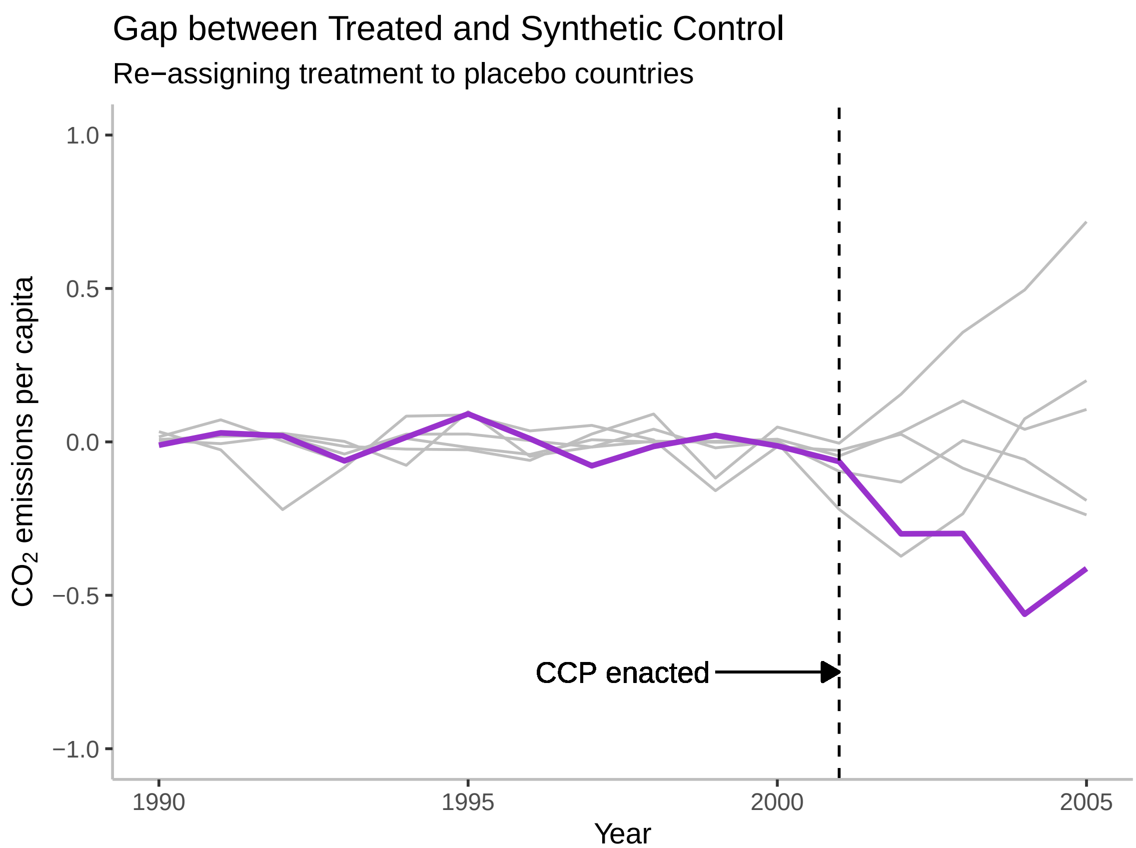


Figure A.63: Gaps in emissions per capita between the treated unit and its synthetic counterpart as estimated by Specification 2. The grey lines represent the gaps in emissions for placebo countries. Countries with a pre-treatment MSPE greater than 5 times that of the UK have been excluded.

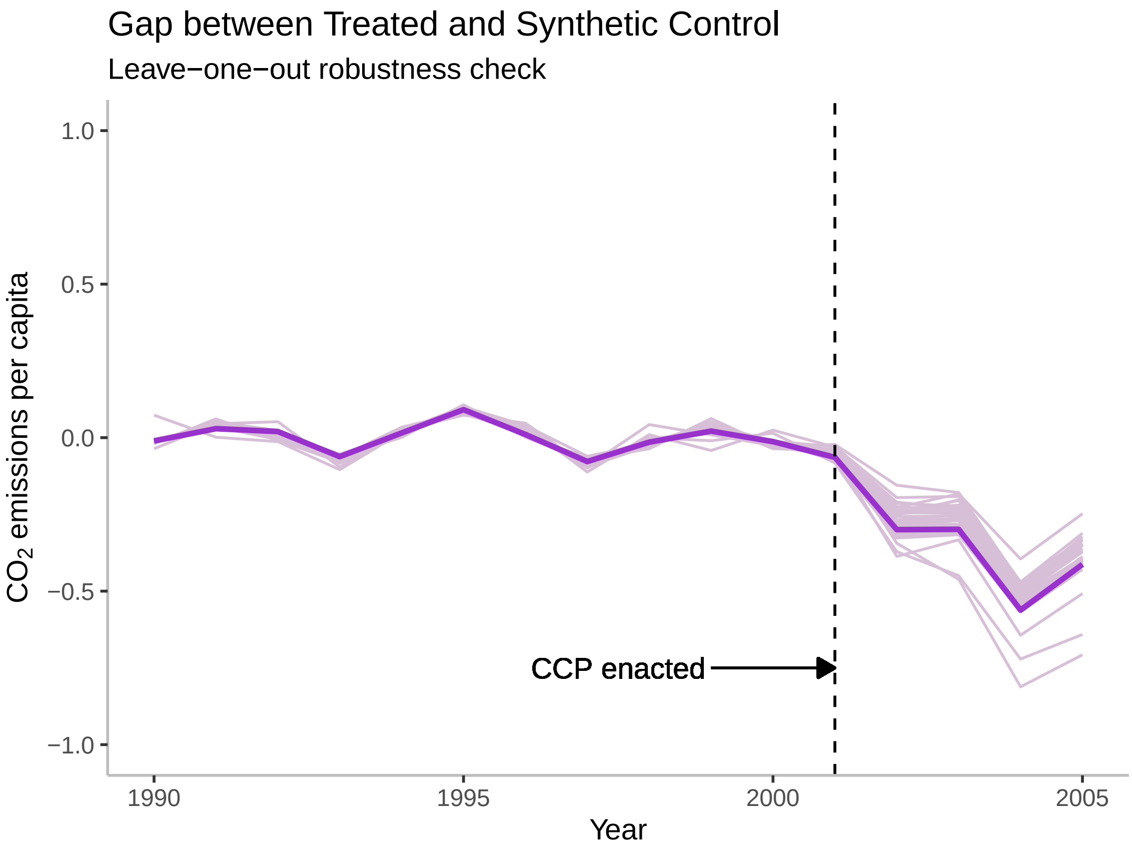


Figure A.64: Gaps between the UK and the synthetic UK in Specification 2. The thick purple line represents the gaps when the synthetic UK is constructed using all countries in the donor pool (37 countries). Each thin purple line represents the gaps when one country is dropped from the donor pool.

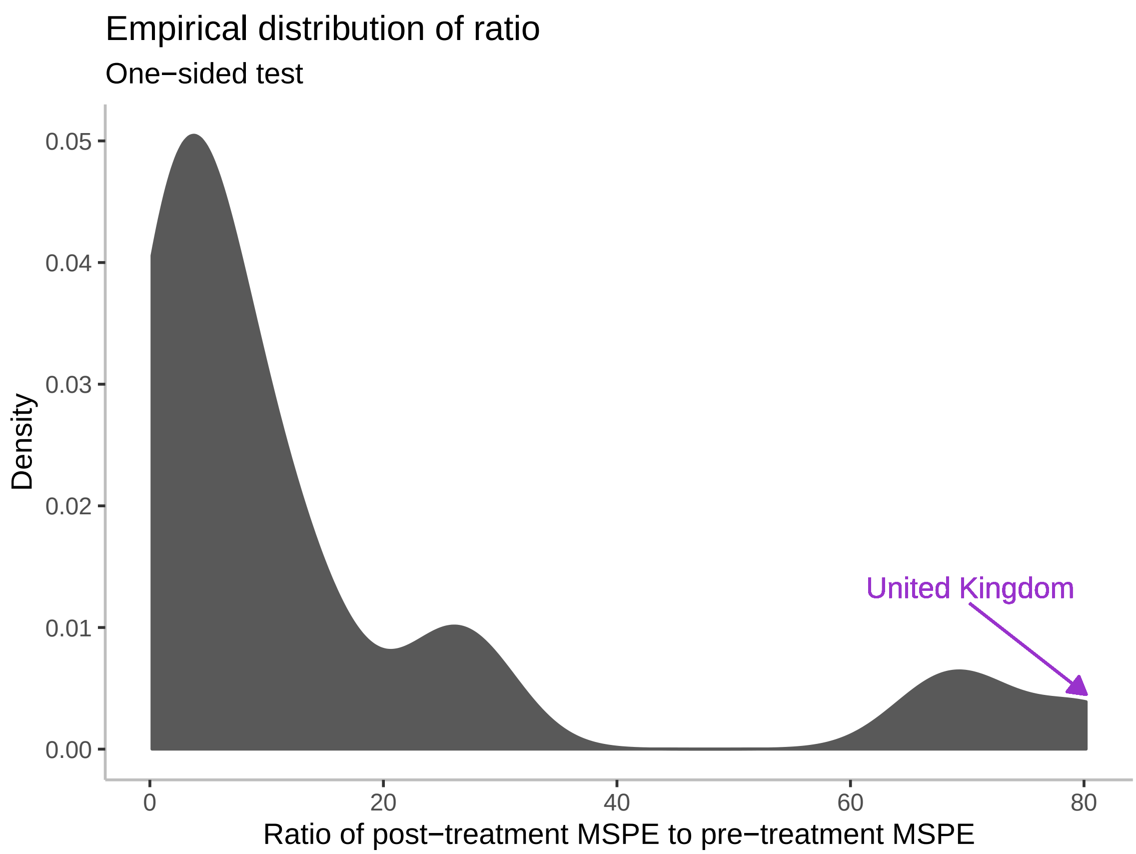


Figure A.65: Non-parametric null distribution for a one-sided test. The density represents the empirical distribution of the ratio statistic (computed as the ratio of post- to pre-treatment Mean Square Prediction Error) for all countries in the sample where the effect of the treatment is estimated to reduce emissions.

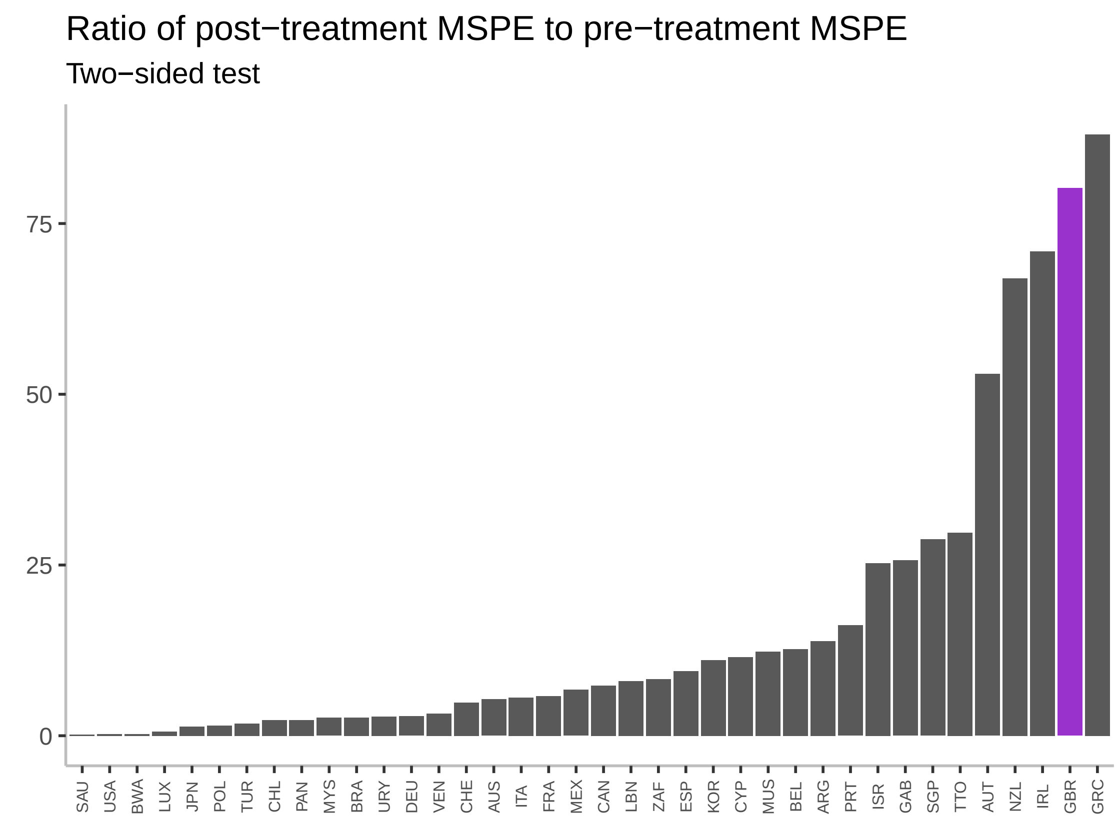


Figure A.66: Ratio of post-treatment Mean Square Prediction Error (MSPE) to pre-treatment MSPE for all countries in the sample.

A.7.3 Specification 3

Outcome variable: CO_2 emissions per capita

Donor pool: OECD, high, and upper middle income countries in 2001, $n = 51$

Covariates: No

Optimization period: 1980-2001

Predictor	Treated UK	Synthetic UK	Sample Mean	Weight
1980 emissions per capita	10.287	10.375	10.654	0.059
1981 emissions per capita	9.955	9.919	9.158	0.055
1982 emissions per capita	9.74	9.743	8.829	0.063
1983 emissions per capita	9.688	9.599	8.411	0.038
1984 emissions per capita	9.382	9.563	8.656	0.044
1985 emissions per capita	9.9	9.874	8.775	0.046
1986 emissions per capita	10.035	10.063	8.746	0.054
1987 emissions per capita	10.068	9.962	8.642	0.045
1988 emissions per capita	10.021	10.077	8.904	0.057
1989 emissions per capita	10.192	10.12	9.257	0.052
1990 emissions per capita	9.711	9.706	9.101	0.05
1991 emissions per capita	9.871	9.871	8.94	0.049
1992 emissions per capita	9.661	9.459	9.353	0.033
1993 emissions per capita	9.455	9.557	9.905	0.06
1994 emissions per capita	9.448	9.527	9.949	0.02
1995 emissions per capita	9.275	9.216	9.897	0.046
1996 emissions per capita	9.48	9.328	9.818	0.028
1997 emissions per capita	9.043	9.145	10.041	0.051
1998 emissions per capita	9.094	9.135	10.022	0.04
1999 emissions per capita	9.048	9.038	9.91	0.045
2000 emissions per capita	9.2	9.238	10.297	0.067

Table A.6: Pre-treatment values of the predictor variables in the UK (column 2), its synthetic control (column 3), and in the unweighted sample (column 4); and weights applied to those predictor variables (column 5).

Balance statistic	
p-value two sample t-test	0.9915562
p-value Kolmogorov Smirnov test	0.999998
Mean difference in QQ plots	0.0309917
Median difference in QQ plots	0.0454545
Maximum difference in QQ plots	0.0909091

Table A.7: Balance statistics between pre-treatment values of the dependent variable in the UK and its synthetic counterpart. p-values are reported for a two sample t-test for a difference in means and for a Kolmogorov-Smirnov test for whether the samples come from different distributions. QQ statistics are reported for the empirical CDF of both samples.

Treatment effect:

- 111 Mt CO_2 abated between 2002-2005
- 0.46 tons of CO_2 per capita abated between 2002-2005
- -6.8% in 2005 compared to what emissions would have been *without* the CCP

Statistical significance:

- Two-sided test: $3/52 \approx 0.058$
- One-sided test: $1/27 \approx 0.037$

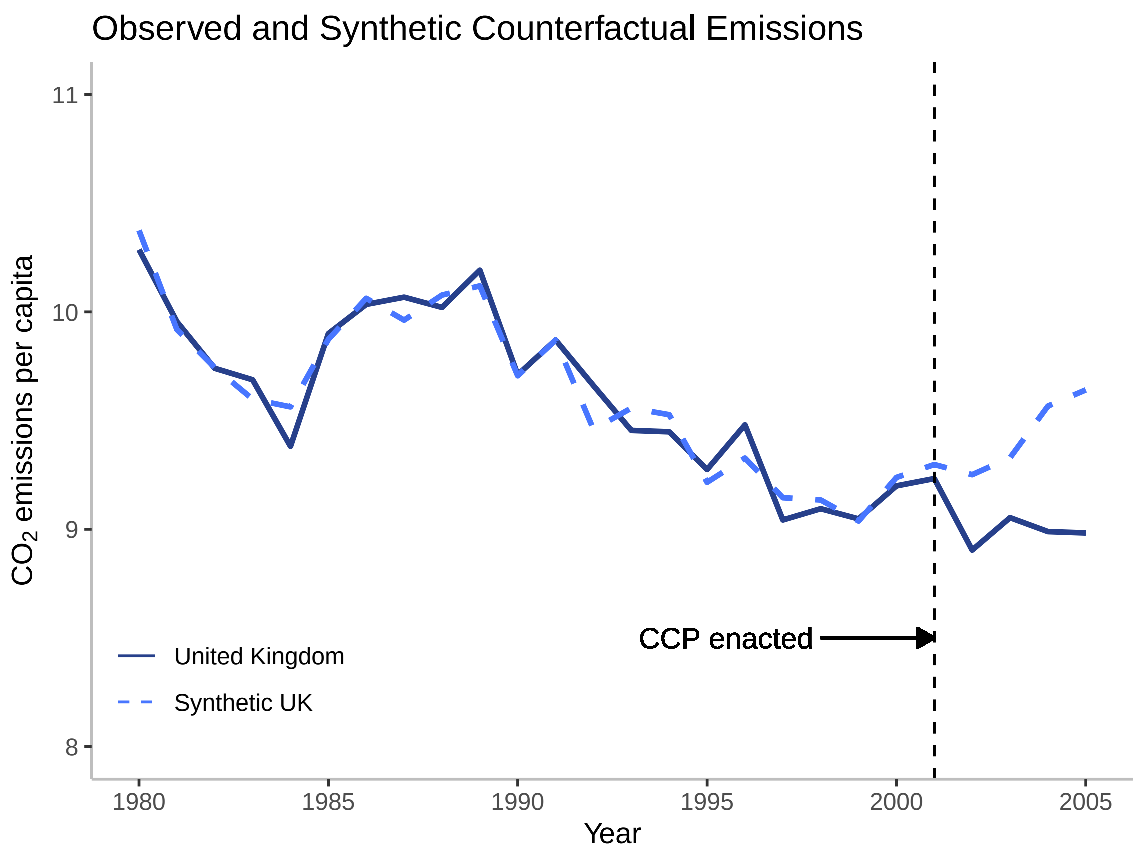


Figure A.67: Observed and synthetic counterfactual emissions for the UK. The dashed line represents the emissions trajectory of a synthetic UK as estimated by Specification 3.

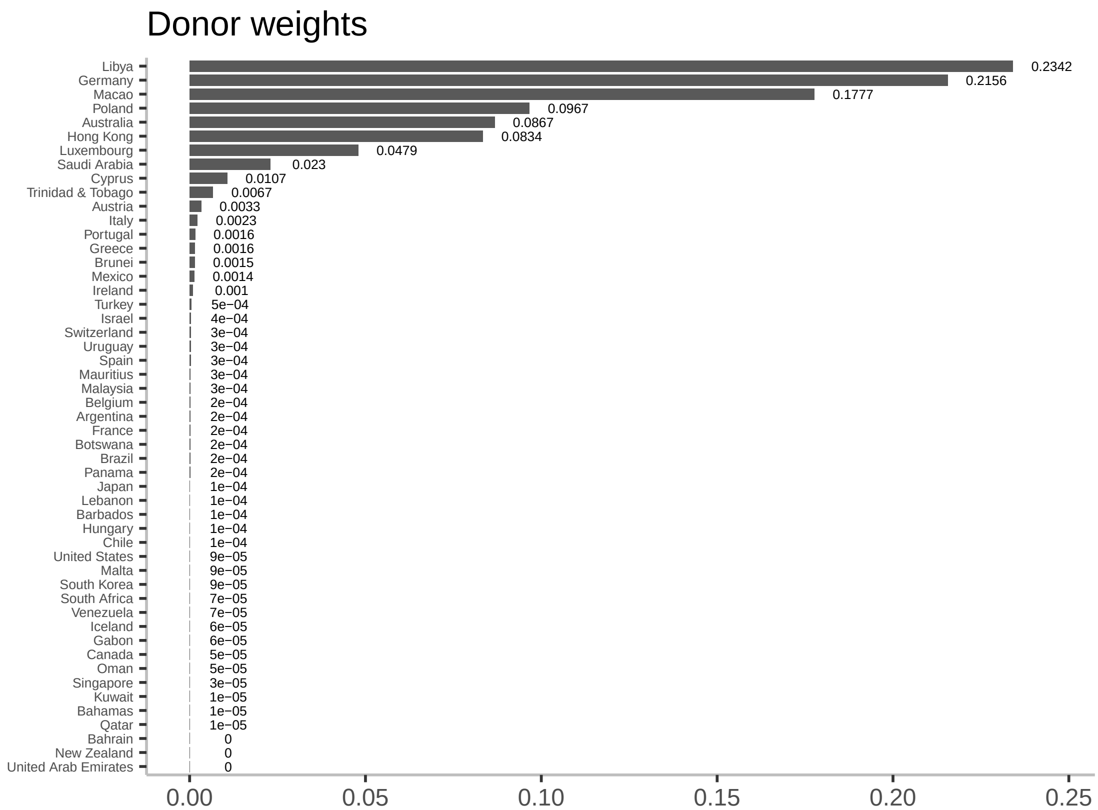


Figure A.68: Weights applied to donor countries in Specification 3.

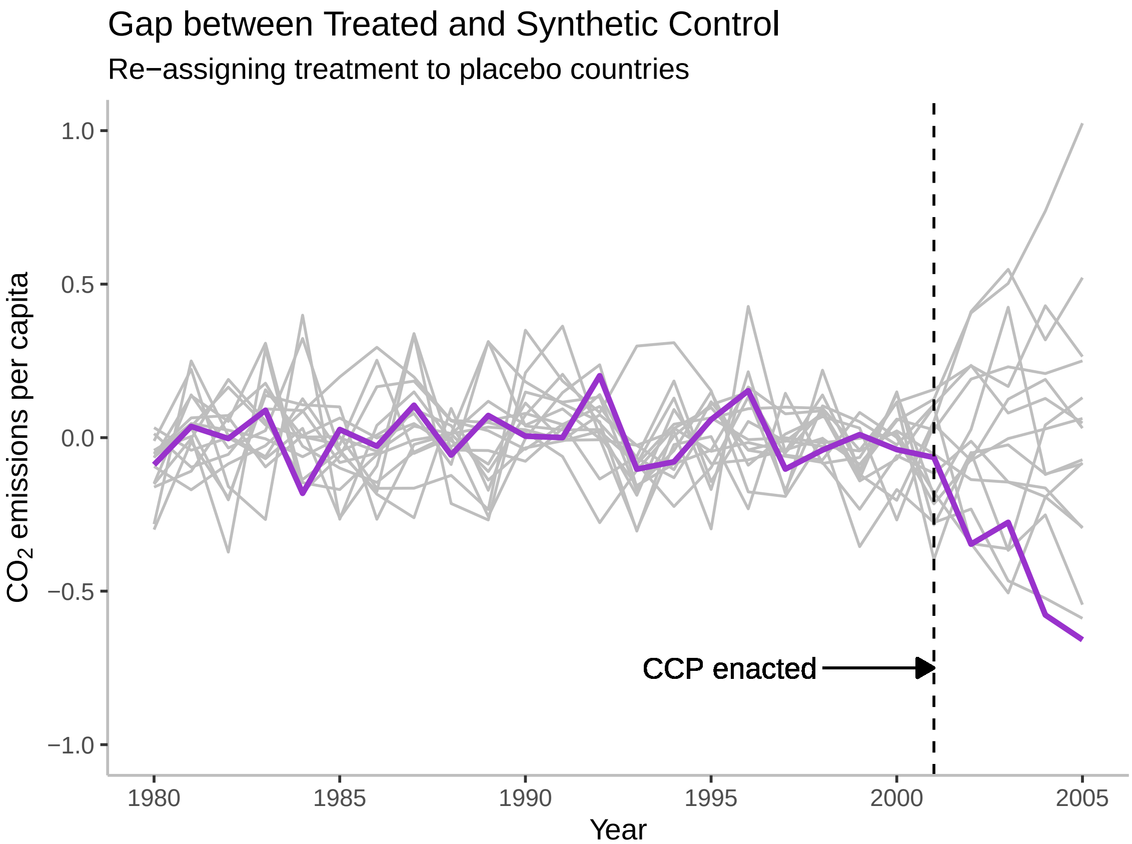


Figure A.69: Gaps in emissions per capita between the treated unit and its synthetic counterpart as estimated by Specification 3. The grey lines represent the gaps in emissions for placebo countries. Countries with a pre-treatment MSPE greater than 5 times that of the UK have been excluded.

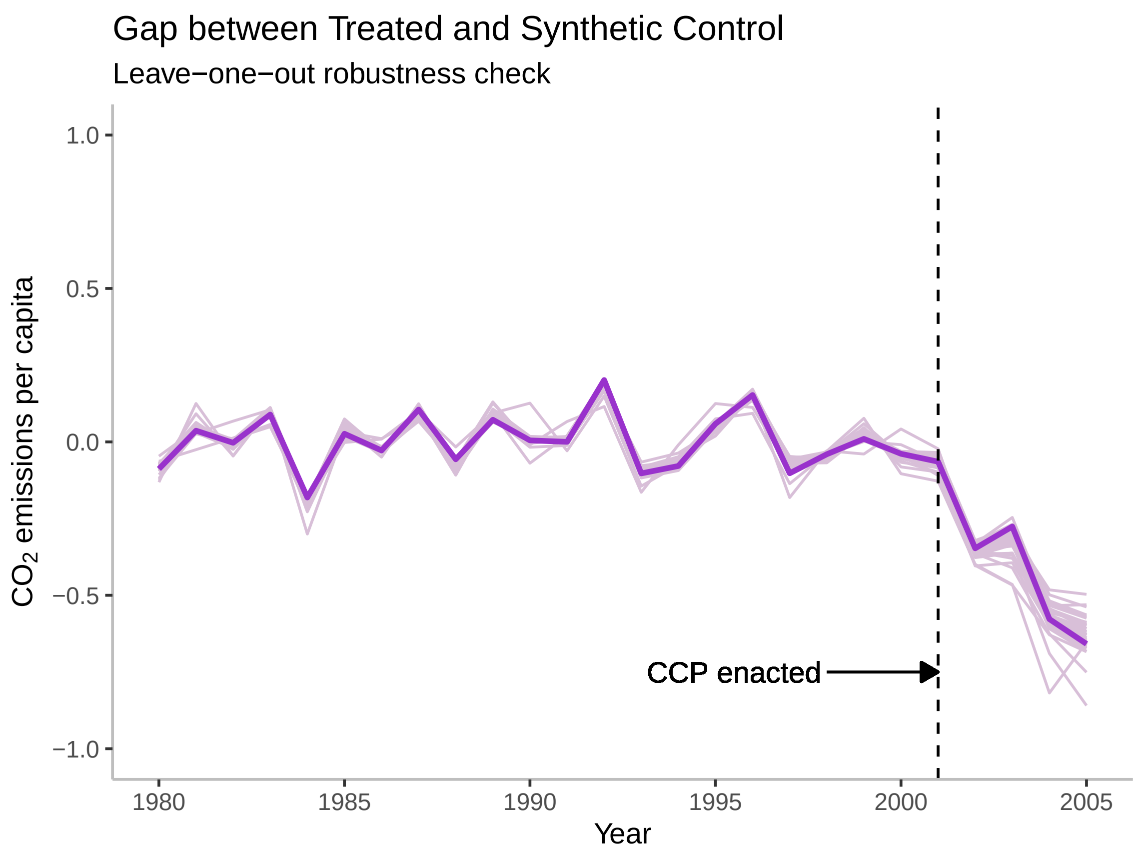


Figure A.70: Gaps between the UK and the synthetic UK in Specification 3. The thick purple line represents the gaps when the synthetic UK is constructed using all countries in the donor pool (51 countries). Each thin purple line represents the gaps when one country is dropped from the donor pool.

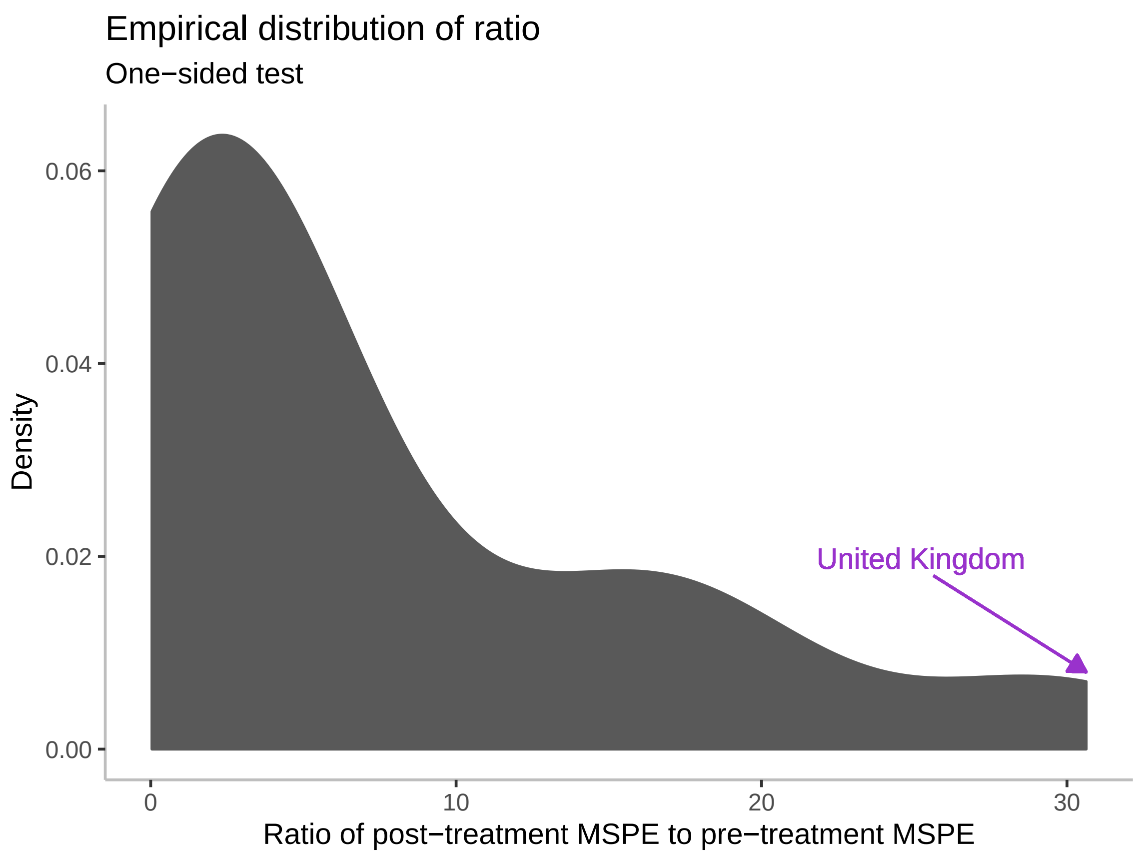


Figure A.71: Non-parametric null distribution for a one-sided test. The density represents the empirical distribution of the ratio statistic (computed as the ratio of post- to pre-treatment Mean Square Prediction Error) for all countries in the sample where the effect of the treatment is estimated to reduce emissions.

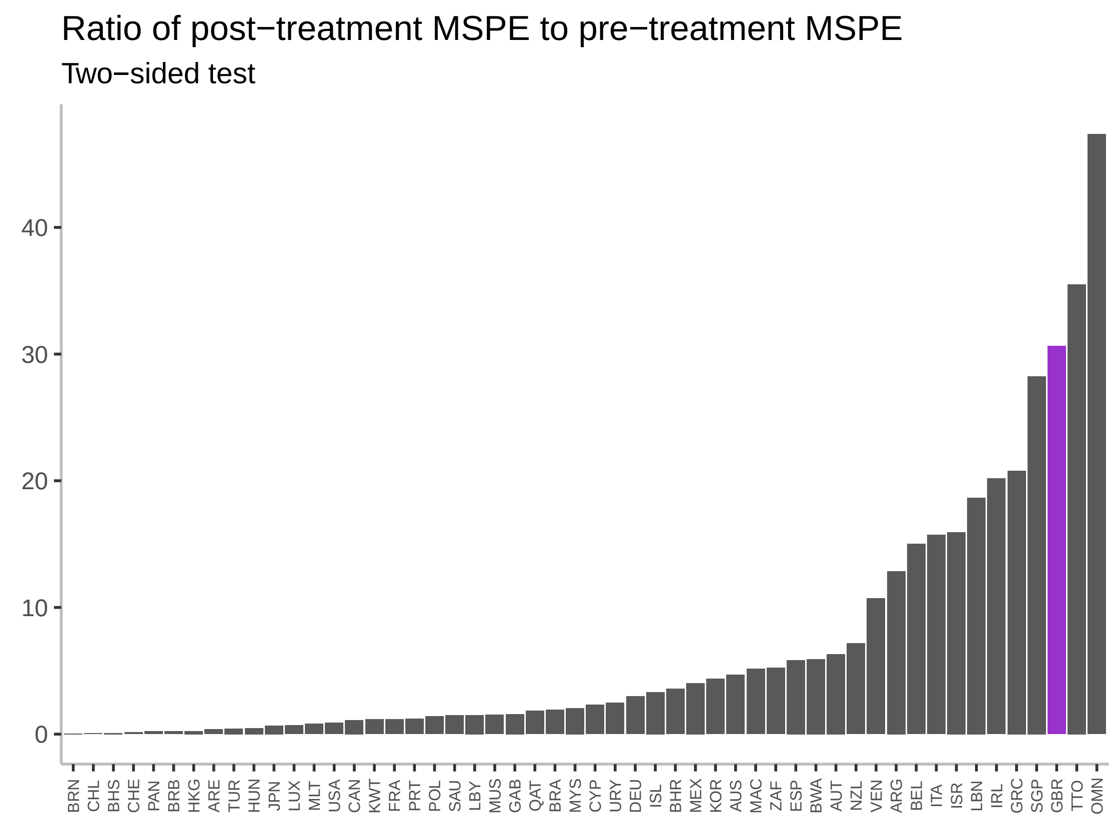


Figure A.72: Ratio of post-treatment Mean Square Prediction Error (MSPE) to pre-treatment MSPE for all countries in the sample.

A.7.4 Specification 4

Outcome variable: CO_2 emissions per capita

Donor pool: OECD and high income countries in 2001, $n = 32$

Covariates: No

Optimization period: 1990-2001

Predictor	Treated UK	Synthetic UK	Sample Mean	Weight
1990 emissions per capita	9.711	9.727	10.952	0.049
1991 emissions per capita	9.871	9.836	10.566	0.045
1992 emissions per capita	9.661	9.647	11.294	0.094
1993 emissions per capita	9.455	9.485	11.822	0.104
1994 emissions per capita	9.448	9.471	11.963	0.09
1995 emissions per capita	9.275	9.241	11.986	0.134
1996 emissions per capita	9.48	9.456	11.717	0.094
1997 emissions per capita	9.043	9.058	12.043	0.122
1998 emissions per capita	9.094	9.123	11.938	0.092
1999 emissions per capita	9.048	9.06	11.643	0.065
2000 emissions per capita	9.2	9.187	12.183	0.111

Table A.8: Pre-treatment values of the predictor variables in the UK (column 2), its synthetic control (column 3), and in the unweighted sample (column 4); and weights applied to those predictor variables (column 5).

Balance statistic	
p-value two sample t-test	0.990644
p-value Kolmogorov Smirnov test	0.9984853
Mean difference in QQ plots	0.0555556
Median difference in QQ plots	0.0833333
Maximum difference in QQ plots	0.1666667

Table A.9: Balance statistics between pre-treatment values of the dependent variable in the UK and its synthetic counterpart. p-values are reported for a two sample t-test for a difference in means and for a Kolmogorov-Smirnov test for whether the samples come from different distributions. QQ statistics are reported for the empirical CDF of both samples.

Treatment effect:

- 139 Mt CO_2 abated between 2002-2005
- 0.58 tons of CO_2 per capita abated between 2002-2005
- -7.7% in 2005 compared to what emissions would have been *without* the CCP

Statistical significance:

- Two-sided test: $1/33 \approx 0.030$
- One-sided test: $1/19 \approx 0.053$

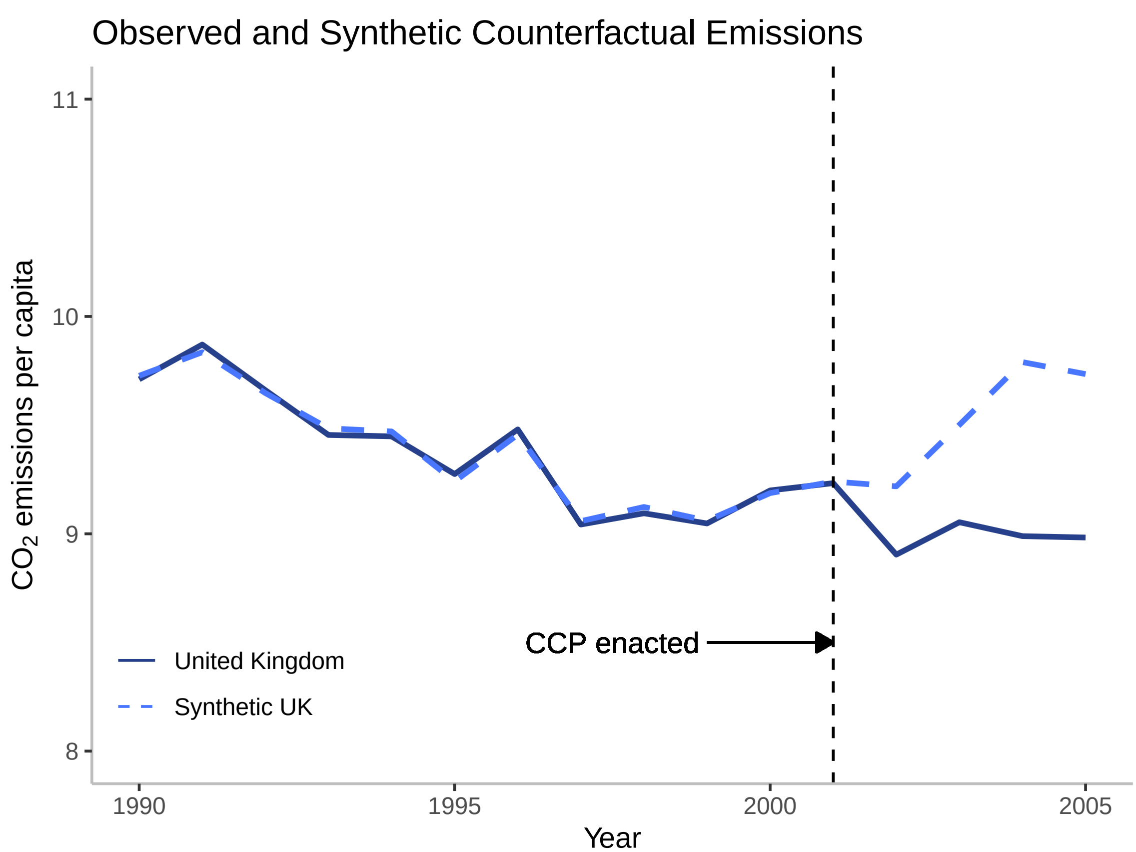


Figure A.73: Observed and synthetic counterfactual emissions for the UK. The dashed line represents the emissions trajectory of a synthetic UK as estimated by Specification 4.

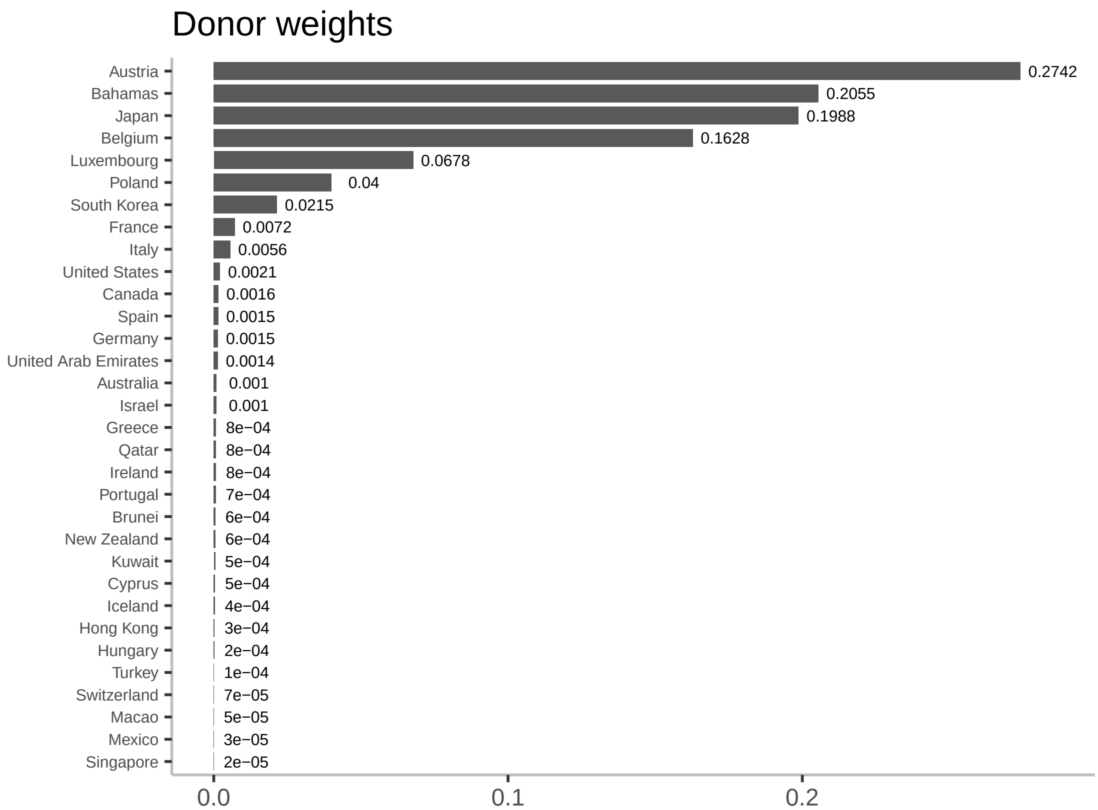


Figure A.74: Weights applied to donor countries in Specification 4.

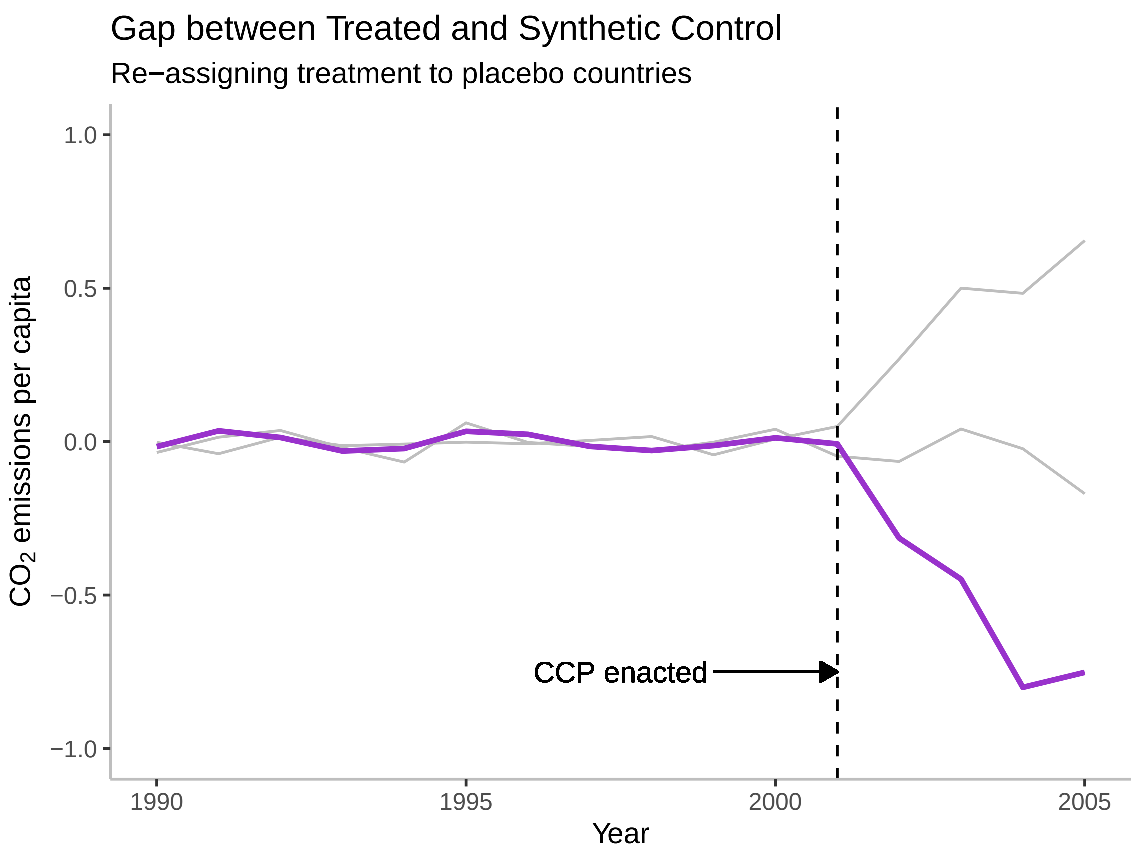


Figure A.75: Gaps in emissions per capita between the treated unit and its synthetic counterpart as estimated by Specification 4. The grey lines represent the gaps in emissions for placebo countries. Countries with a pre-treatment MSPE greater than 5 times that of the UK have been excluded.

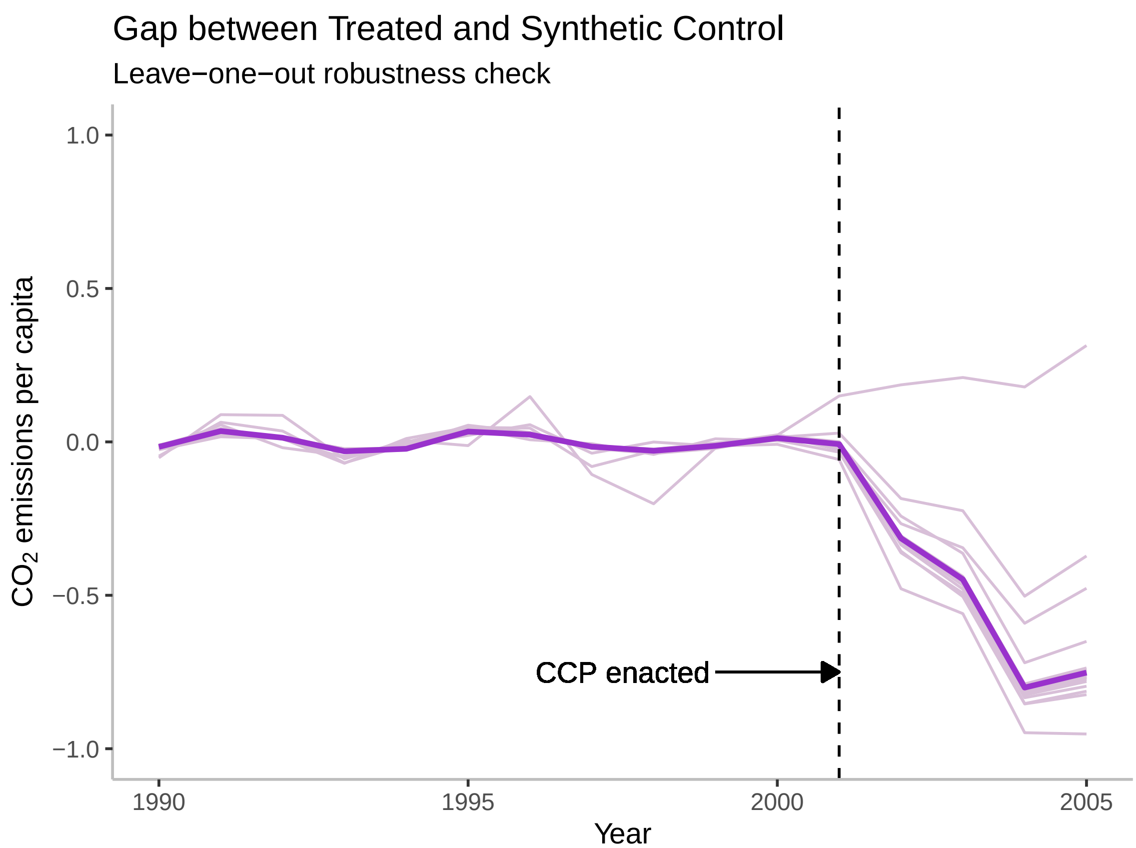


Figure A.76: Gaps between the UK and the synthetic UK in Specification 4. The thick purple line represents the gaps when the synthetic UK is constructed using all countries in the donor pool (32 countries). Each thin purple line represents the gaps when one country is dropped from the donor pool.

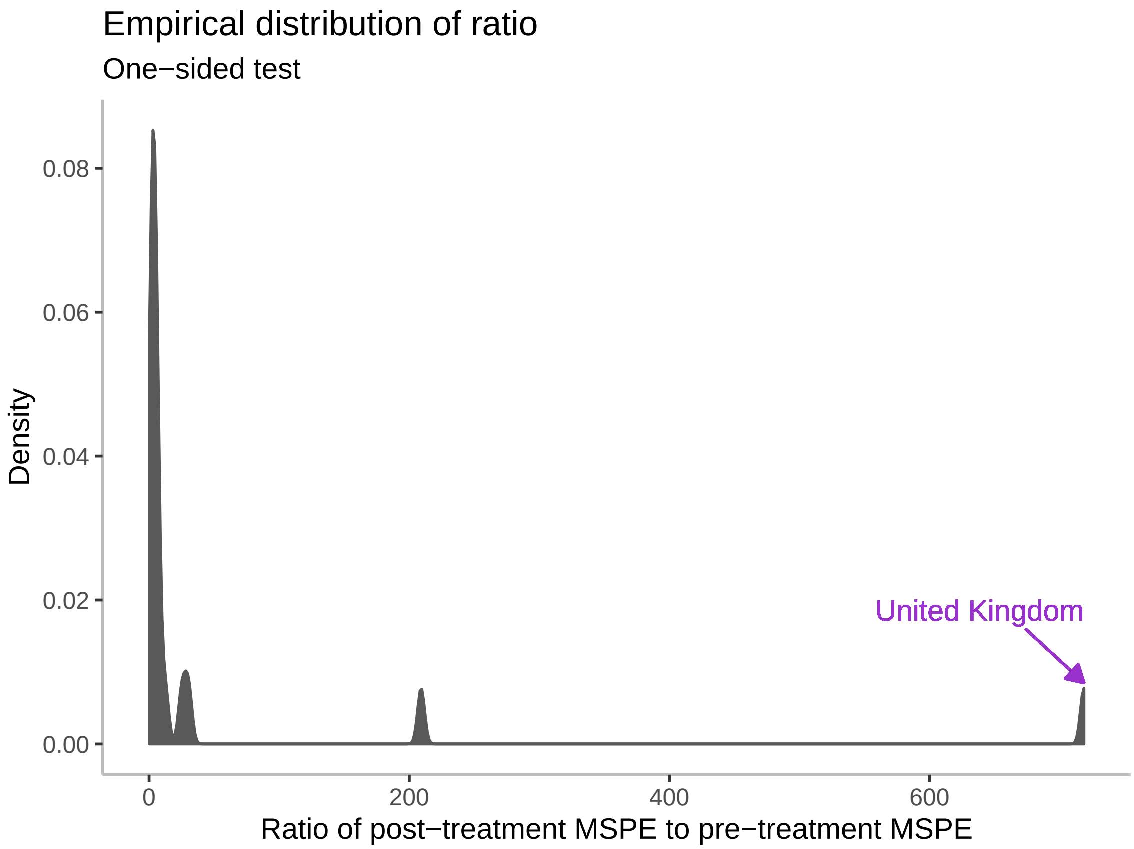


Figure A.77: Non-parametric null distribution for a one-sided test. The density represents the empirical distribution of the ratio statistic (computed as the ratio of post- to pre-treatment Mean Square Prediction Error) for all countries in the sample where the effect of the treatment is estimated to reduce emissions.

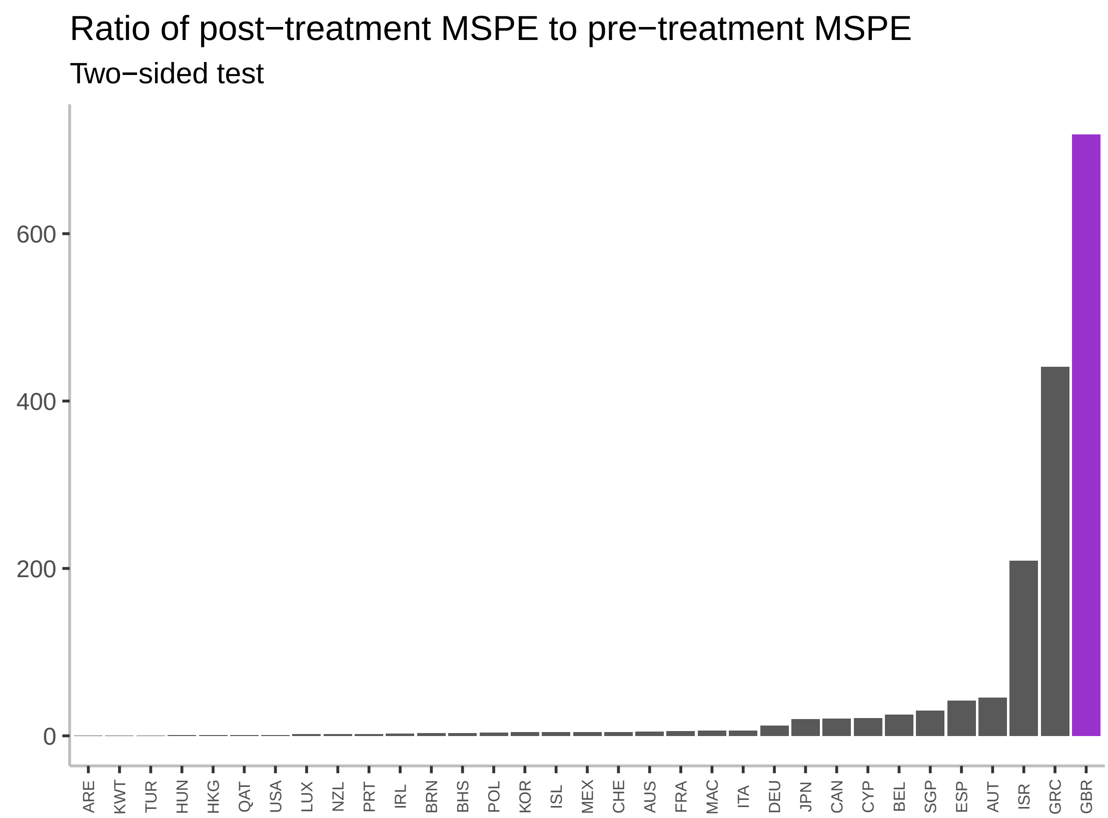


Figure A.78: Ratio of post-treatment Mean Square Prediction Error (MSPE) to pre-treatment MSPE for all countries in the sample.

A.7.5 Specification 5

Outcome variable: CO_2 emissions per capita

Donor pool: OECD in 2001, $n = 22$

Covariates: No

Optimization period: 1990-2001

Predictor	Treated UK	Synthetic UK	Sample Mean	Weight
1990 emissions per capita	9.711	9.728	9.367	0.068
1991 emissions per capita	9.871	9.842	9.326	0.103
1992 emissions per capita	9.661	9.66	9.28	0.087
1993 emissions per capita	9.455	9.496	9.305	0.145
1994 emissions per capita	9.448	9.429	9.26	0.095
1995 emissions per capita	9.275	9.178	9.13	0.087
1996 emissions per capita	9.48	9.468	9.398	0.118
1997 emissions per capita	9.043	9.142	9.402	0.074
1998 emissions per capita	9.094	9.09	9.327	0.104
1999 emissions per capita	9.048	9.07	9.417	0.096
2000 emissions per capita	9.2	9.161	9.577	0.025

Table A.10: Pre-treatment values of the predictor variables in the UK (column 2), its synthetic control (column 3), and in the unweighted sample (column 4); and weights applied to those predictor variables (column 5).

Balance statistic	
p-value two sample t-test	0.99347
p-value Kolmogorov Smirnov test	0.9984853
Mean difference in QQ plots	0.0555556
Median difference in QQ plots	0.0833333
Maximum difference in QQ plots	0.1666667

Table A.11: Balance statistics between pre-treatment values of the dependent variable in the UK and its synthetic counterpart. p-values are reported for a two sample t-test for a difference in means and for a Kolmogorov-Smirnov test for whether the samples come from different distributions. QQ statistics are reported for the empirical CDF of both samples.

Treatment effect:

- 108 Mt CO_2 abated between 2002-2005
- 0.45 tons of CO_2 per capita abated between 2002-2005
- -5.3% in 2005 compared to what emissions would have been *without* the CCP

Statistical significance:

- Two-sided test: $1/23 \approx 0.043$
- One-sided test: $1/15 \approx 0.067$

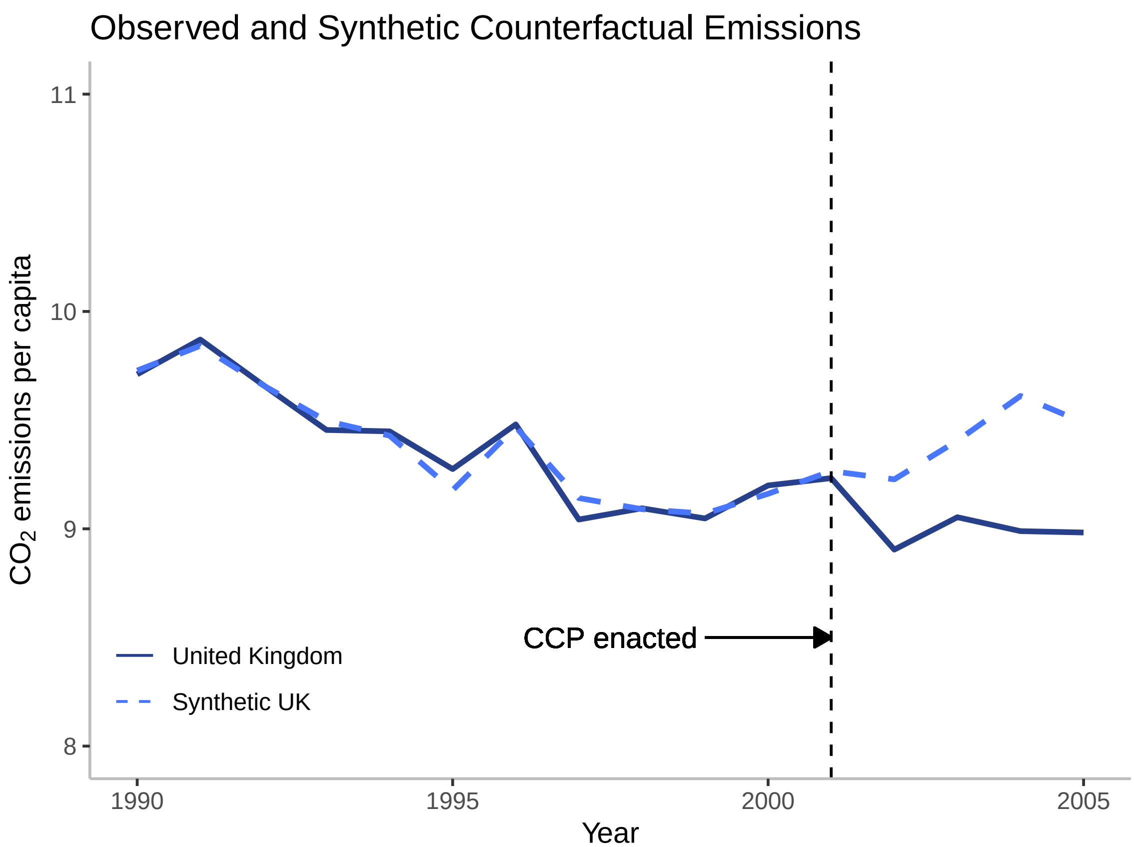


Figure A.79: Observed and synthetic counterfactual emissions for the UK. The dashed line represents the emissions trajectory of a synthetic UK as estimated by Specification 5.

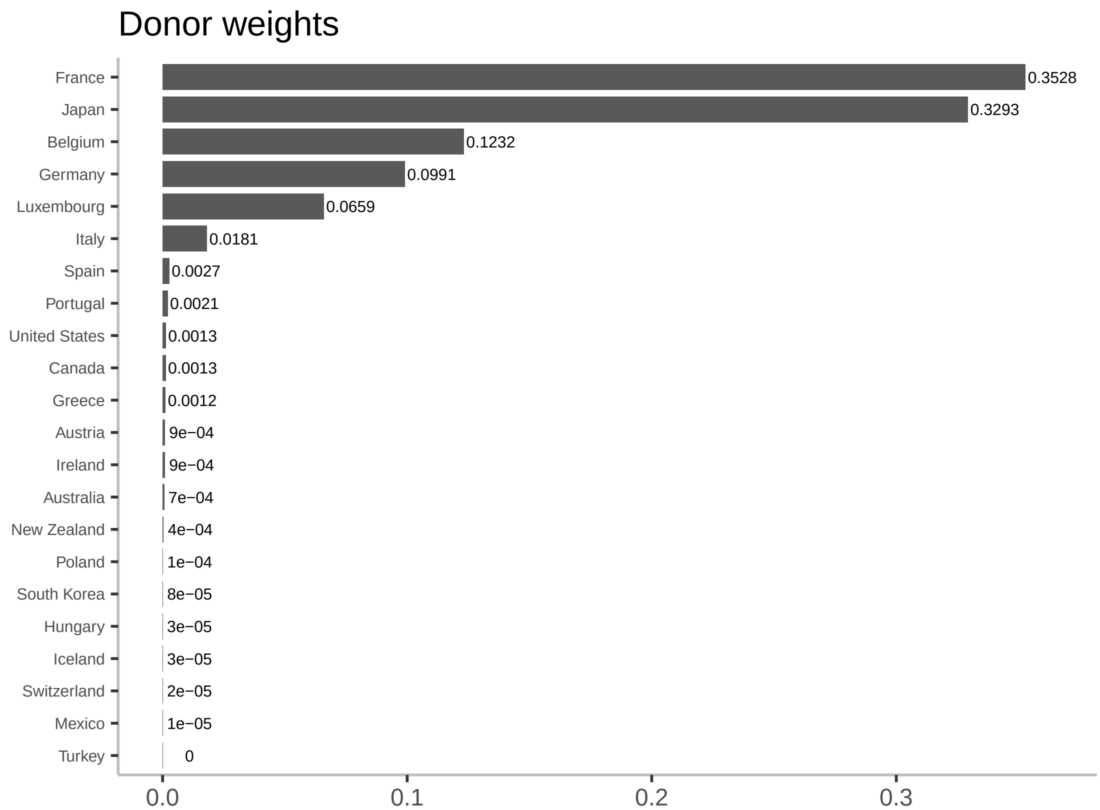


Figure A.80: Weights applied to donor countries in Specification 5.

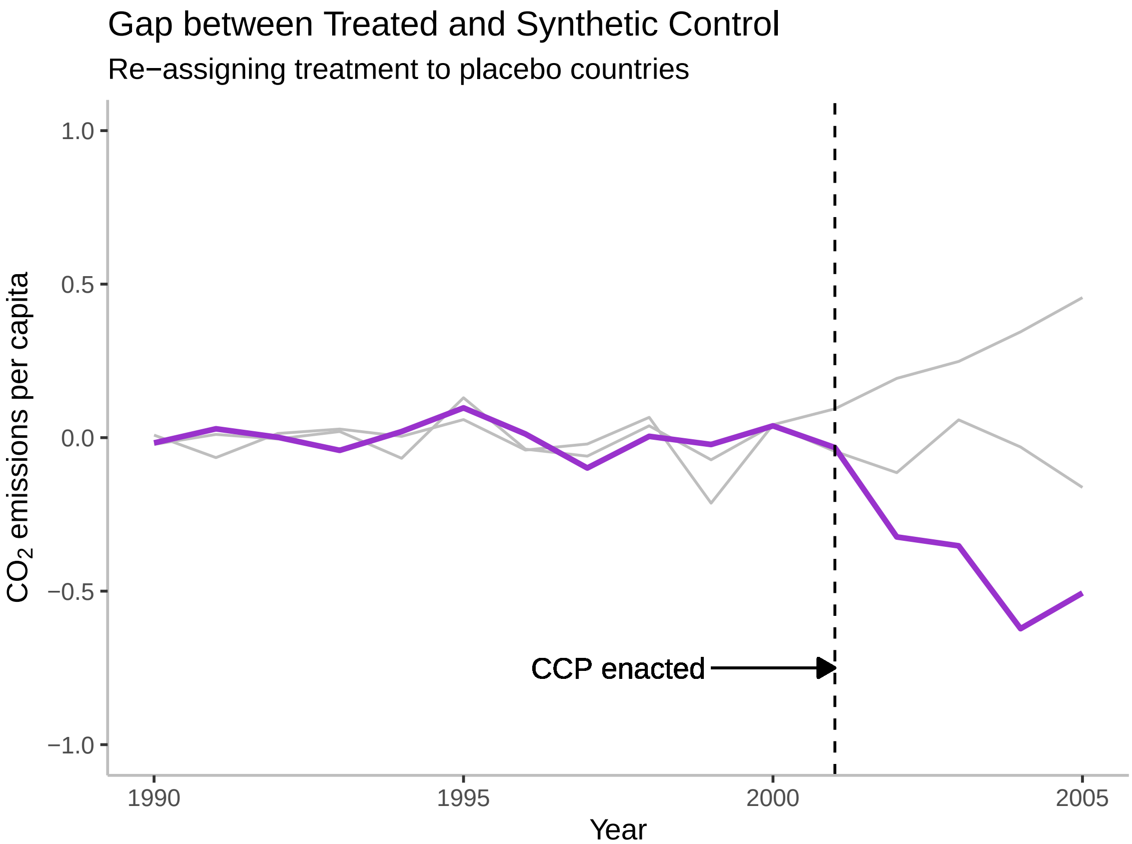


Figure A.81: Gaps in emissions per capita between the treated unit and its synthetic counterpart as estimated by Specification 5. The grey lines represent the gaps in emissions for placebo countries. Countries with a pre-treatment MSPE greater than 5 times that of the UK have been excluded.

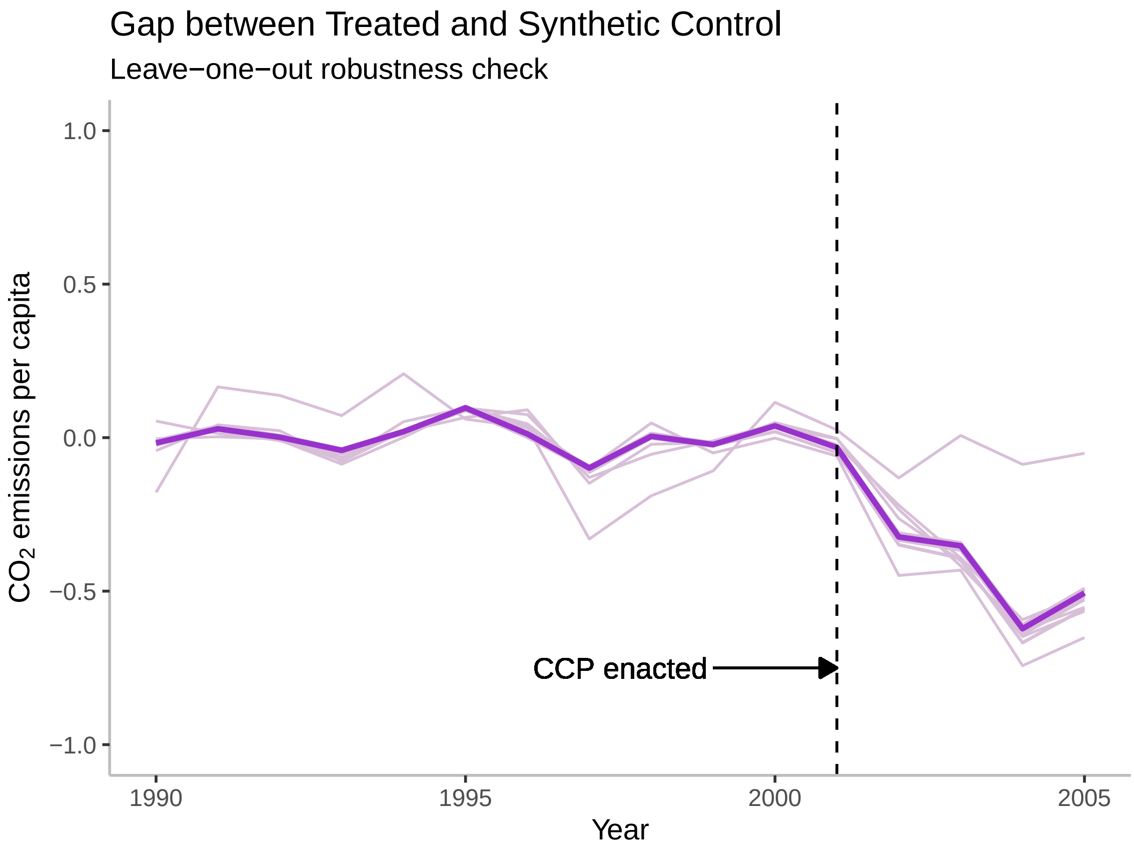


Figure A.82: Gaps between the UK and the synthetic UK in Specification 5. The thick purple line represents the gaps when the synthetic UK is constructed using all countries in the donor pool (22 countries). Each thin purple line represents the gaps when one country is dropped from the donor pool.

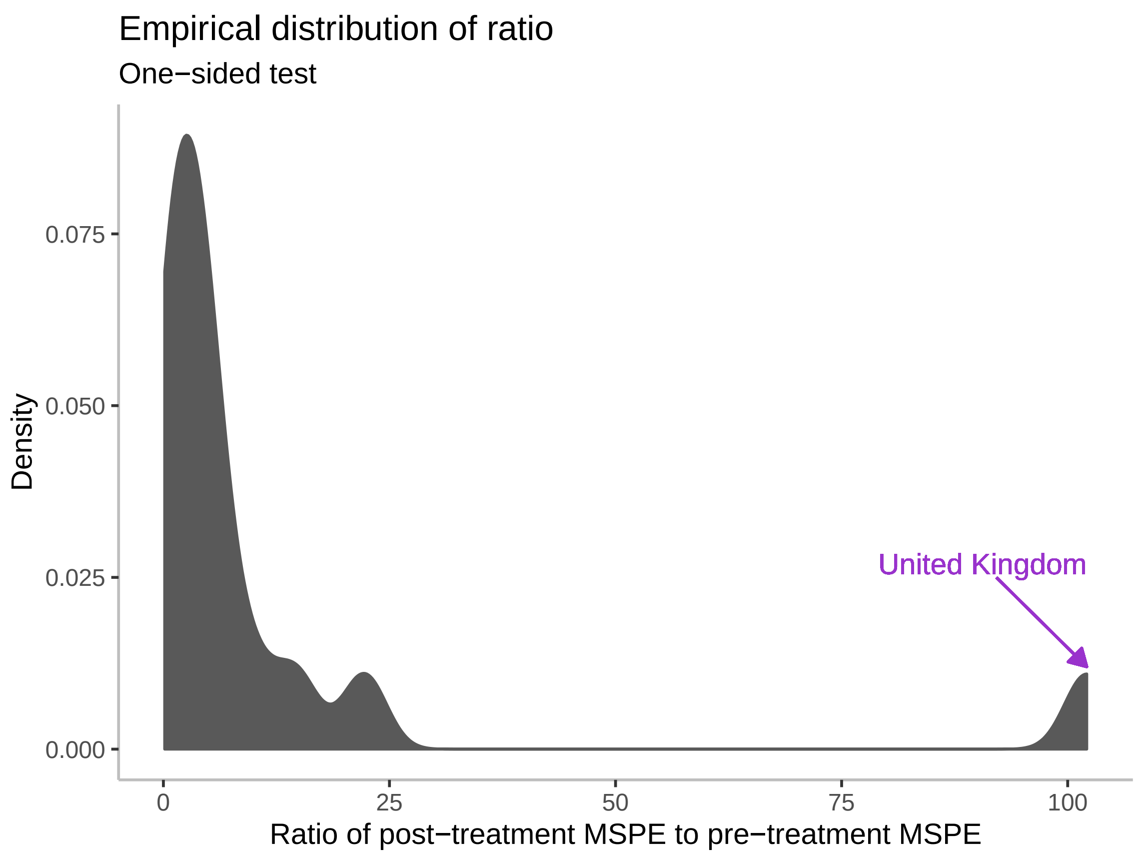


Figure A.83: Non-parametric null distribution for a one-sided test. The density represents the empirical distribution of the ratio statistic (computed as the ratio of post- to pre-treatment Mean Square Prediction Error) for all countries in the sample where the effect of the treatment is estimated to reduce emissions.

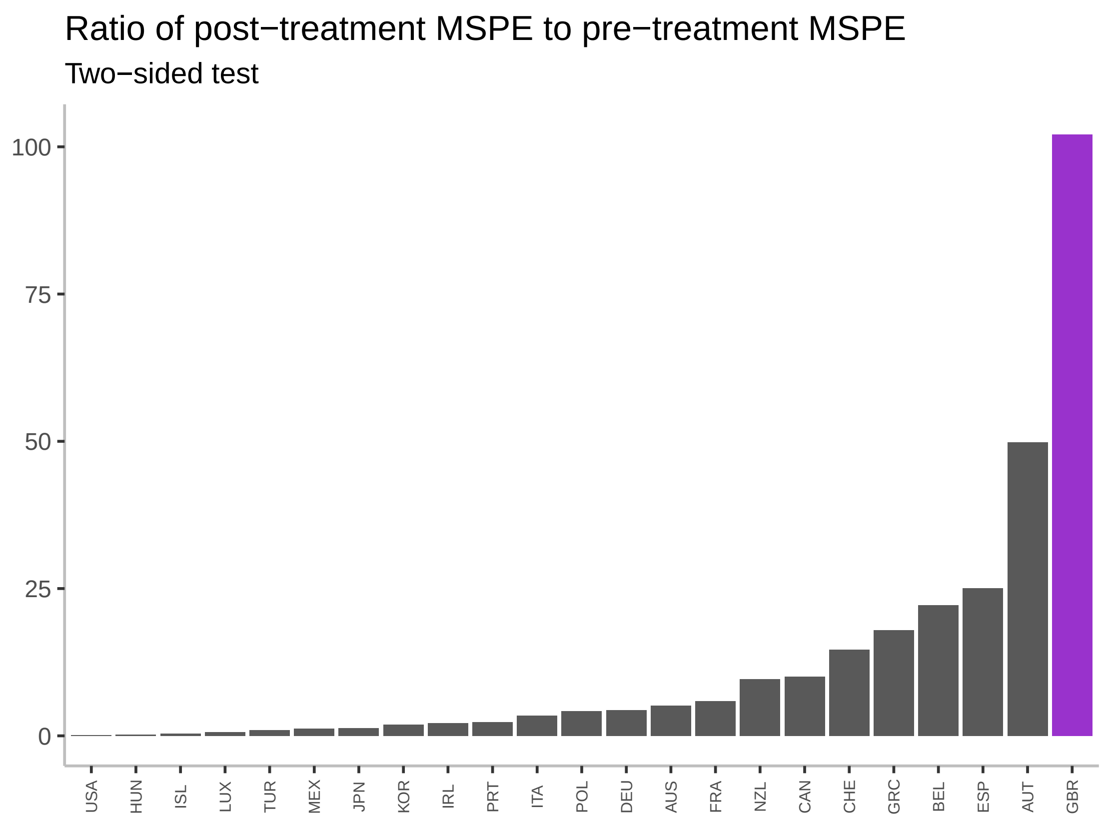


Figure A.84: Ratio of post-treatment Mean Square Prediction Error (MSPE) to pre-treatment MSPE for all countries in the sample.

A.7.6 Specification 6

Outcome variable: Emissions rescaled to 1990 baseline

Donor pool: OECD, high, and upper middle income countries in 2001, $n = 51$

Covariates: No

Optimization period: 1990-2001

Predictor	Treated UK	Synthetic UK	Sample Mean	Weight
1991 emissions rescaled to 1990	1.02	1.019	1.024	0.091
1992 emissions rescaled to 1990	1.001	0.998	1.073	0.115
1993 emissions rescaled to 1990	0.982	0.987	1.129	0.118
1994 emissions rescaled to 1990	0.983	0.984	1.143	0.101
1995 emissions rescaled to 1990	0.968	0.967	1.155	0.122
1996 emissions rescaled to 1990	0.992	0.99	1.201	0.092
1997 emissions rescaled to 1990	0.949	0.948	1.249	0.082
1998 emissions rescaled to 1990	0.957	0.96	1.286	0.11
1999 emissions rescaled to 1990	0.955	0.957	1.312	0.105
2000 emissions rescaled to 1990	0.975	0.971	1.36	0.064

Table A.12: Pre-treatment values of the predictor variables in the UK (column 2), its synthetic control (column 3), and in the unweighted sample (column 4); and weights applied to those predictor variables (column 5).

Balance statistic	
p-value two sample t-test	0.9749674
p-value Kolmogorov Smirnov test	0.8689817
Mean difference in QQ plots	0.0625
Median difference in QQ plots	0.0833333
Maximum difference in QQ plots	0.25

Table A.13: Balance statistics between pre-treatment values of the dependent variable in the UK and its synthetic counterpart. p-values are reported for a two sample t-test for a difference in means and for a Kolmogorov-Smirnov test for whether the samples come from different distributions. QQ statistics are reported for the empirical CDF of both samples.

Treatment effect:

- Emissions in 2005 were 11.7% lower relative to a 1990 baseline compared to emissions *without* the CCP

Statistical significance:

- Two-sided test: $2/52 \approx 0.038$
- One-sided test: $1/34 \approx 0.029$

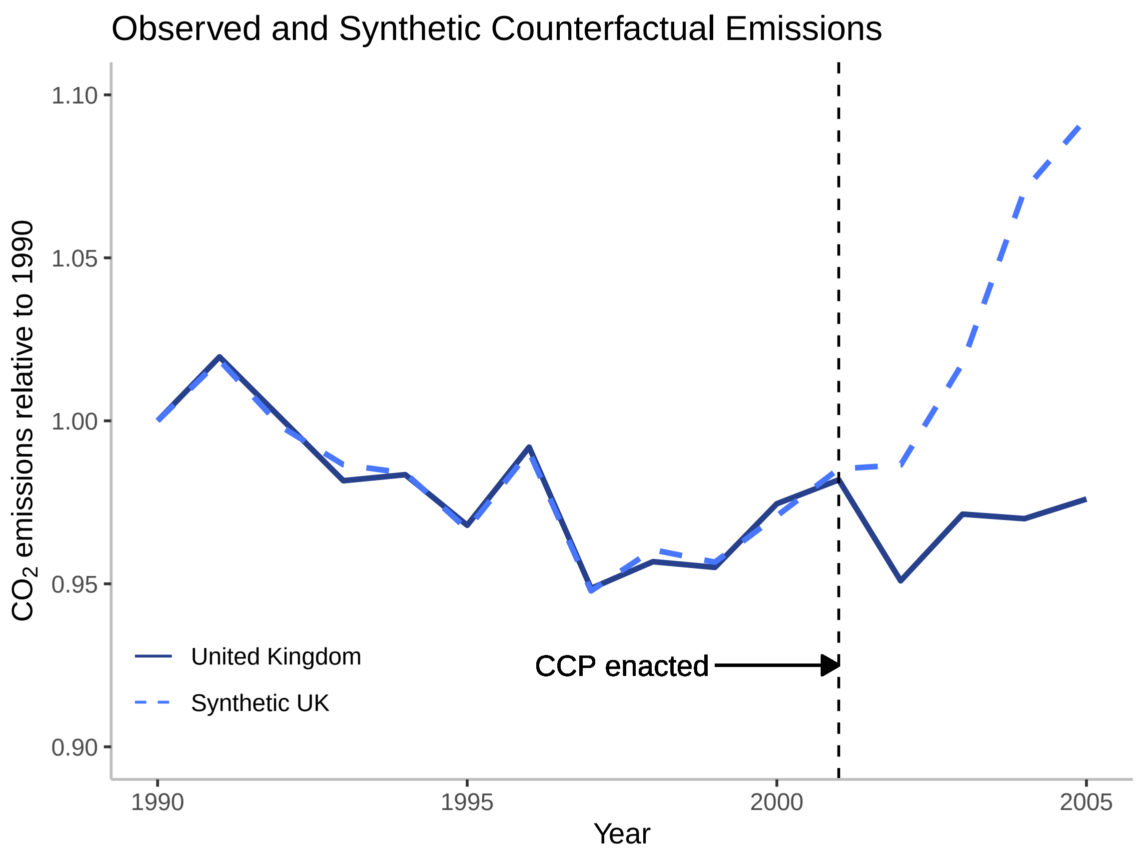


Figure A.85: Observed and synthetic counterfactual emissions for the UK. The dashed line represents the emissions trajectory of a synthetic UK as estimated by Specification 6.

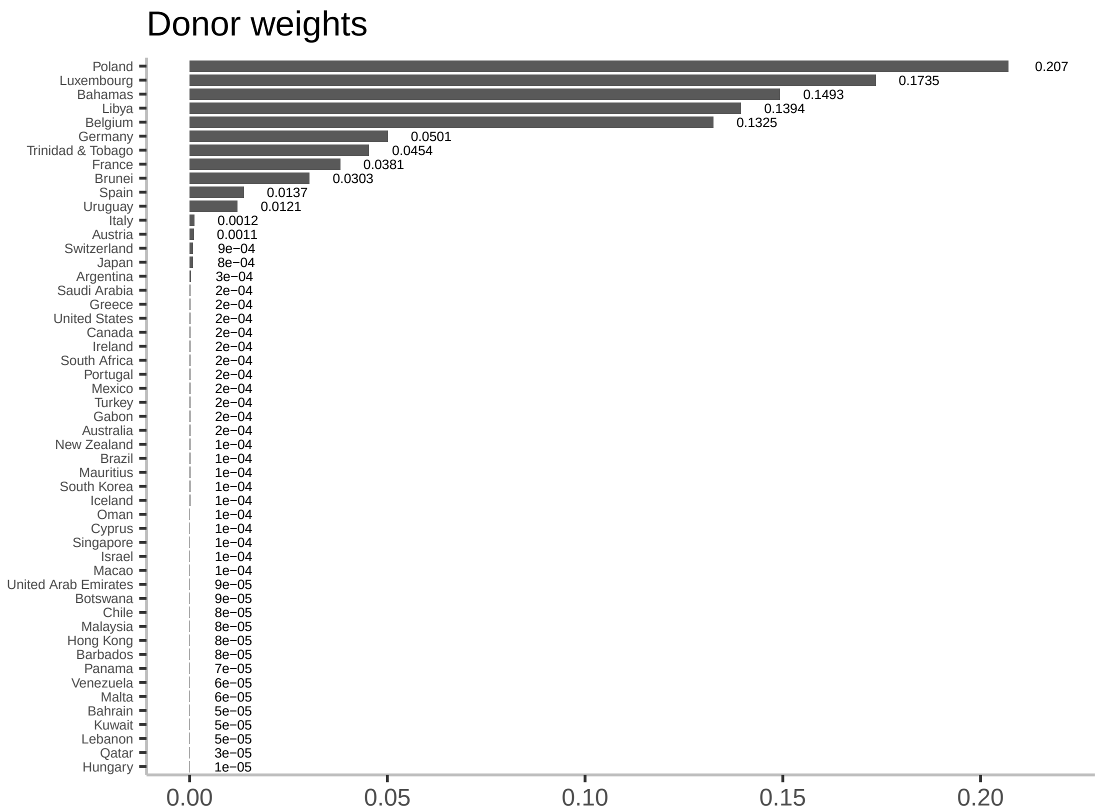


Figure A.86: Weights applied to donor countries in Specification 6.

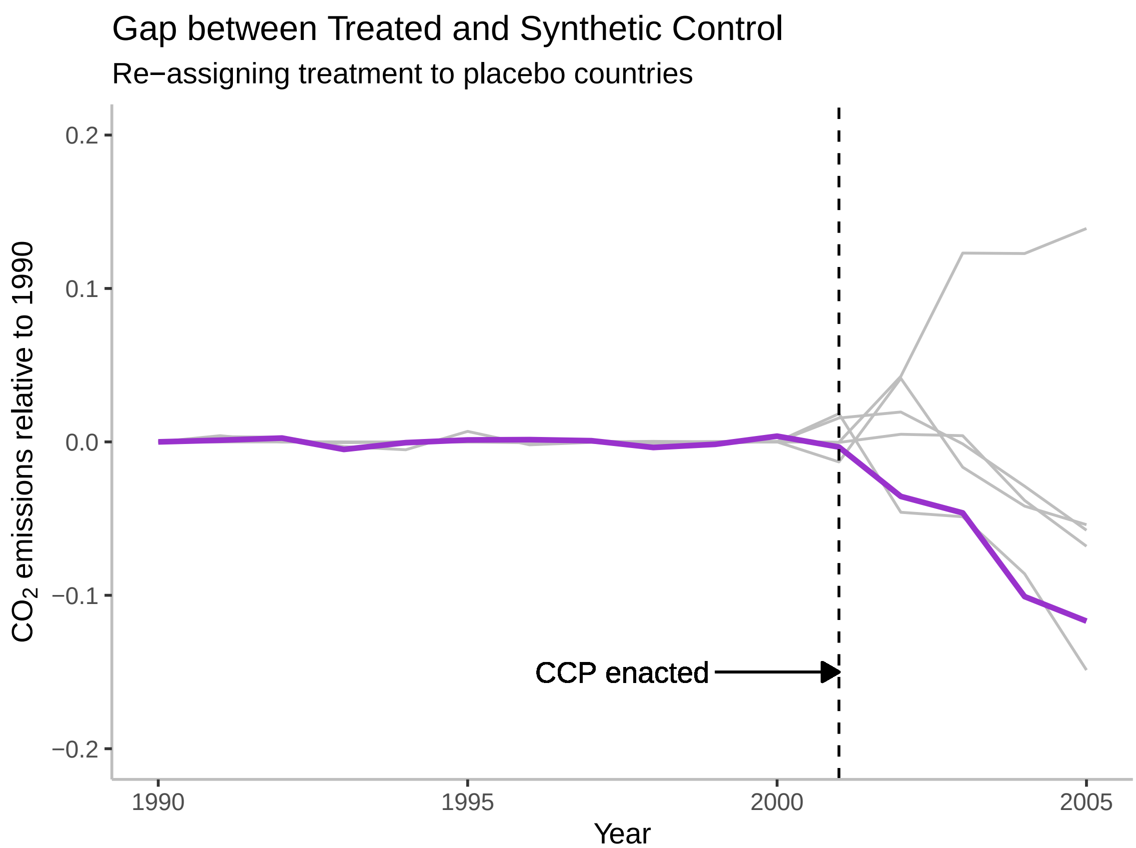


Figure A.87: Gaps in emissions (rescaled to 1990 baseline) between the treated unit and its synthetic counterpart as estimated by Specification 6. The grey lines represent the gaps in emissions for placebo countries. Countries with a pre-treatment MSPE greater than 5 times that of the UK have been excluded.

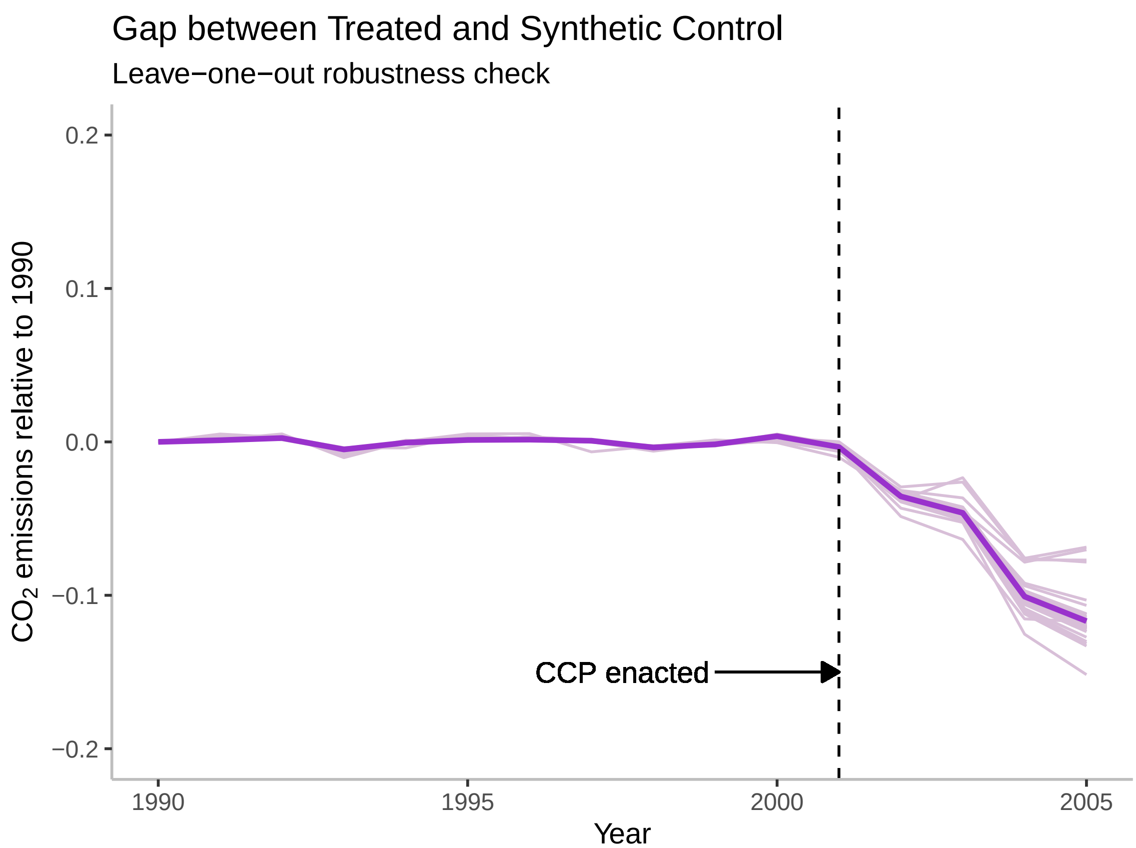


Figure A.88: Gaps between the UK and the synthetic UK in Specification 6. The thick purple line represents the gaps when the synthetic UK is constructed using all countries in the donor pool (51 countries). Each thin purple line represents the gaps when one country is dropped from the donor pool.

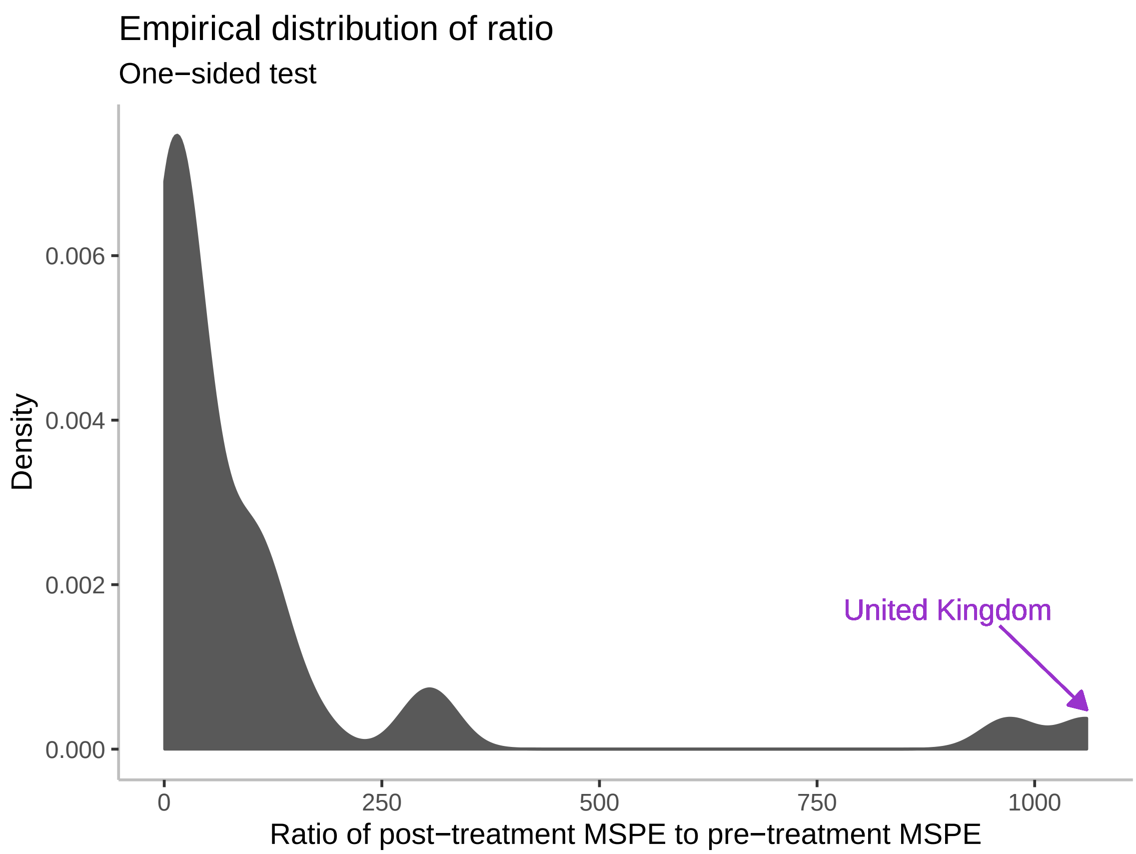


Figure A.89: Non-parametric null distribution for a one-sided test. The density represents the empirical distribution of the ratio statistic (computed as the ratio of post- to pre-treatment Mean Square Prediction Error) for all countries in the sample where the effect of the treatment is estimated to reduce emissions.

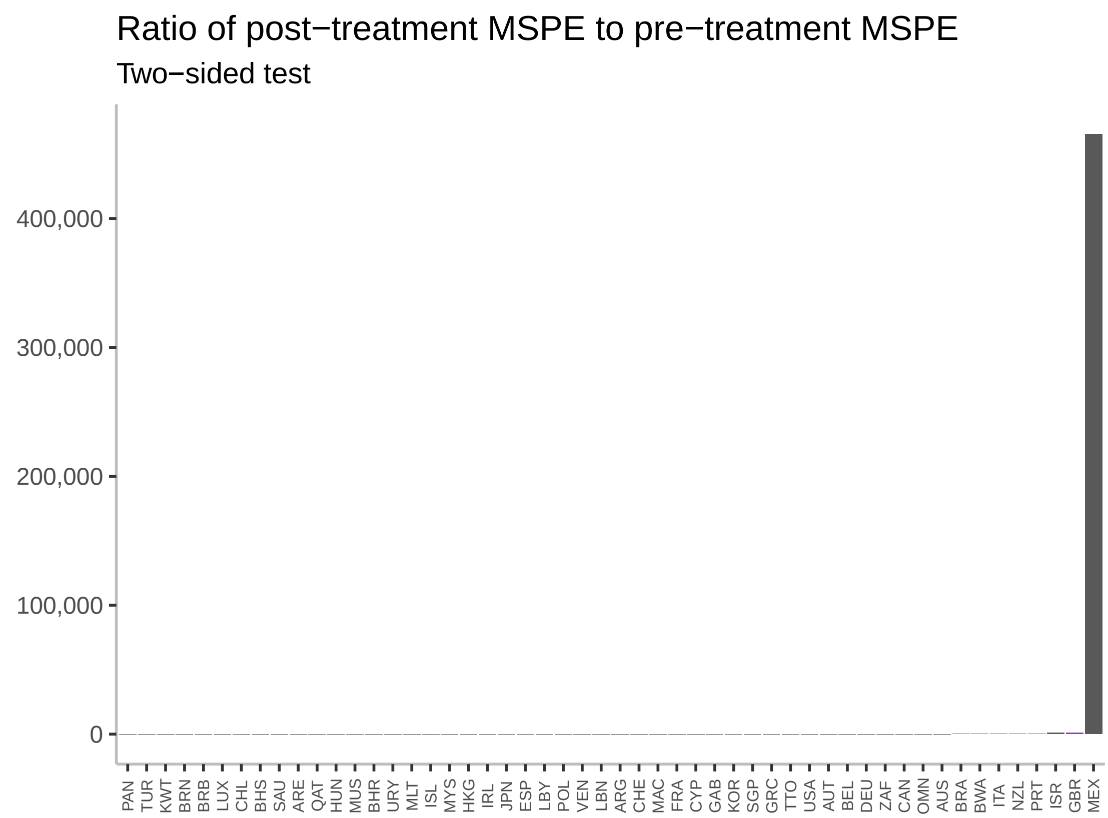


Figure A.90: Ratio of post-treatment Mean Square Prediction Error (MSPE) to pre-treatment MSPE for all countries in the sample.

A.7.7 Specification 7

Outcome variable: Emissions rescaled to 1990 baseline

Donor pool: OECD and high income countries in 2001, $n = 32$

Covariates: No

Optimization period: 1990-2001

Predictor	Treated UK	Synthetic UK	Sample Mean	Weight
1991 emissions rescaled to 1990 baseline	1.02	1.014	0.998	0.019
1992 emissions rescaled to 1990 baseline	1.001	0.998	1.05	0.073
1993 emissions rescaled to 1990 baseline	0.982	0.988	1.088	0.082
1994 emissions rescaled to 1990 baseline	0.983	0.985	1.109	0.071
1995 emissions rescaled to 1990 baseline	0.968	0.965	1.118	0.108
1996 emissions rescaled to 1990 baseline	0.992	0.991	1.14	0.122
1997 emissions rescaled to 1990 baseline	0.949	0.949	1.18	0.161
1998 emissions rescaled to 1990 baseline	0.957	0.961	1.206	0.094
1999 emissions rescaled to 1990 baseline	0.955	0.957	1.214	0.117
2000 emissions rescaled to 1990 baseline	0.975	0.972	1.273	0.154

Table A.14: Pre-treatment values of the predictor variables in the UK (column 2), its synthetic control (column 3), and in the unweighted sample (column 4); and weights applied to those predictor variables (column 5).

Balance statistic	
p-value two sample t-test	0.9845835
p-value Kolmogorov Smirnov test	0.8689817
Mean difference in QQ plots	0.0694444
Median difference in QQ plots	0.0833333
Maximum difference in QQ plots	0.25

Table A.15: Balance statistics between pre-treatment values of the dependent variable in the UK and its synthetic counterpart. p-values are reported for a two sample t-test for a difference in means and for a Kolmogorov-Smirnov test for whether the samples come from different distributions. QQ statistics are reported for the empirical CDF of both samples.

Treatment effect:

- Emissions in 2005 were 11.3% lower relative to a 1990 baseline compared to emissions *without* the CCP

Statistical significance:

- Two-sided test: $1/33 \approx 0.030$

- One-sided test: $1/18 \approx 0.056$

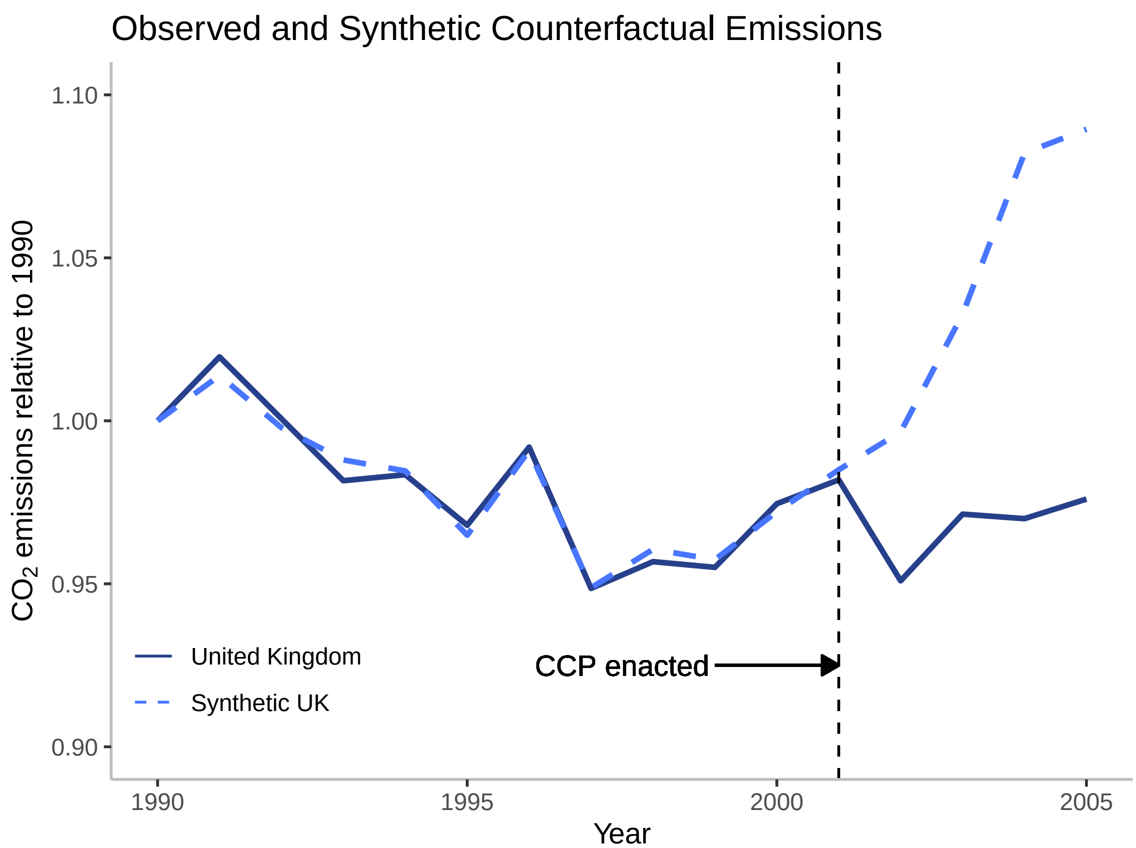


Figure A.91: Observed and synthetic counterfactual emissions for the UK. The dashed line represents the emissions trajectory of a synthetic UK as estimated by Specification 7.

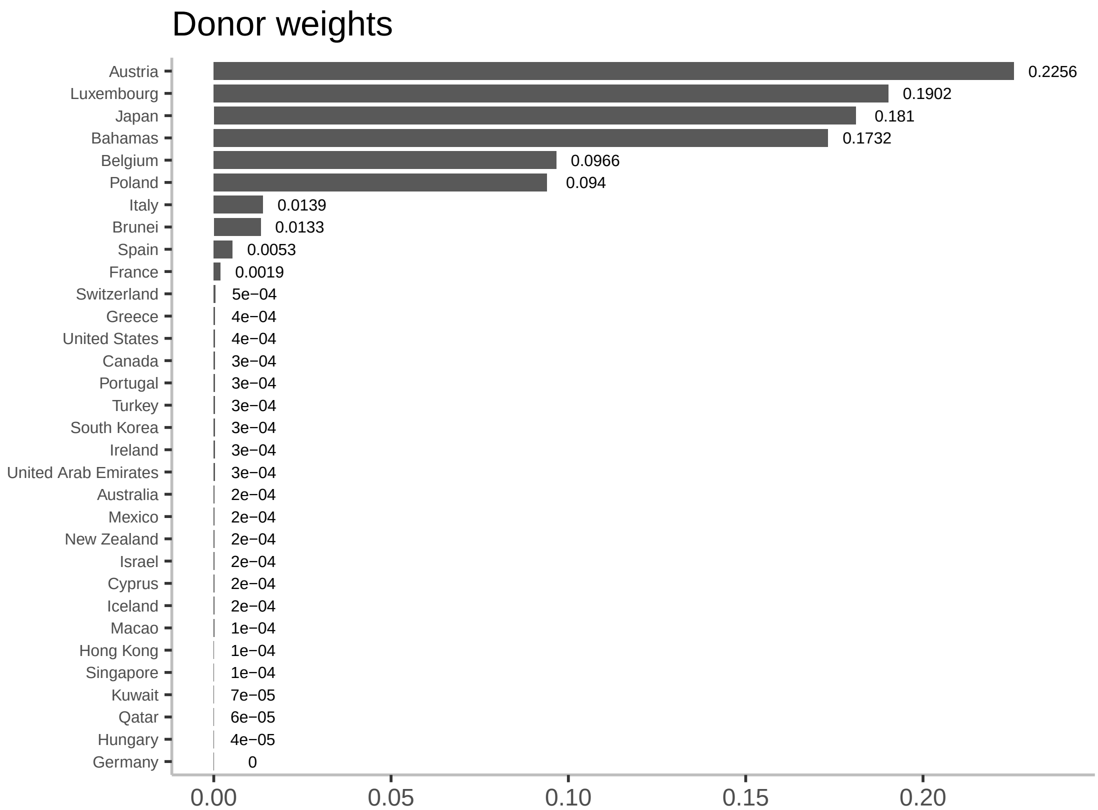


Figure A.92: Weights applied to donor countries in Specification 7.

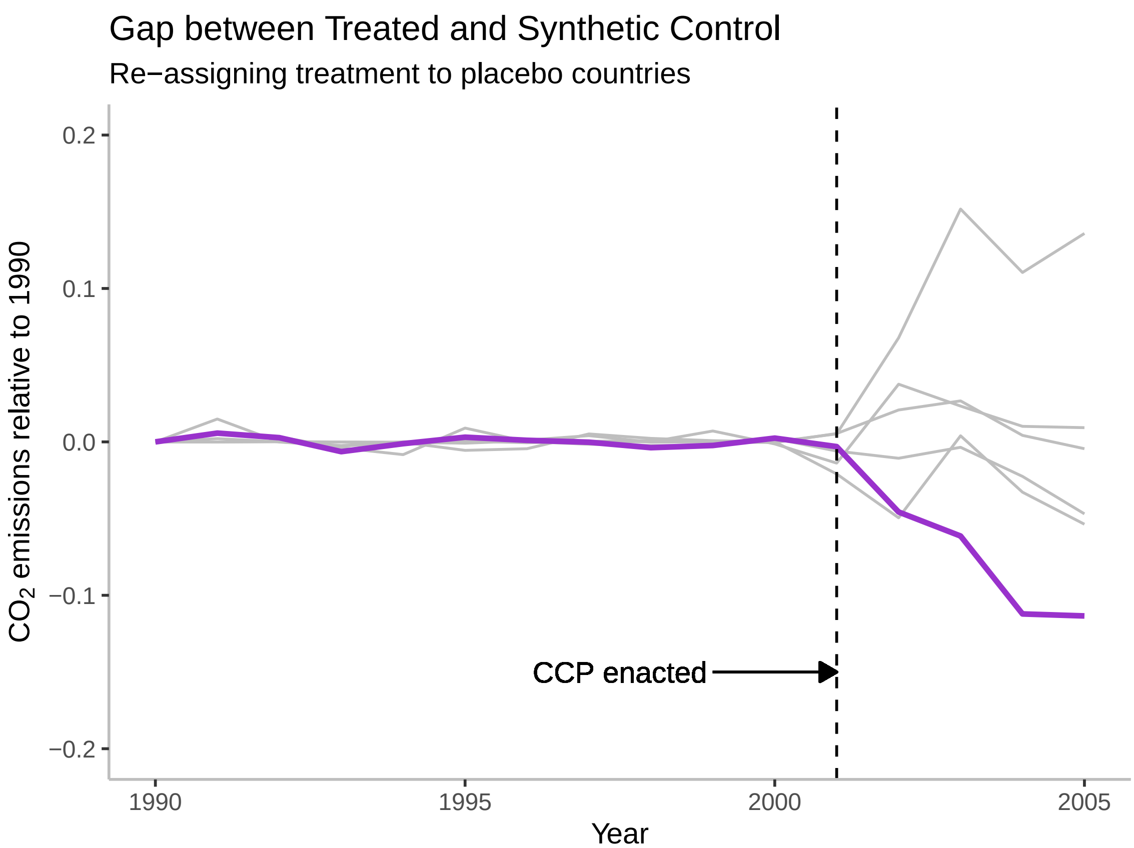


Figure A.93: Gaps in emissions (rescaled to 1990 baseline) between the treated unit and its synthetic counterpart as estimated by Specification 7. The grey lines represent the gaps in emissions for placebo countries. Countries with a pre-treatment MSPE greater than 5 times that of the UK have been excluded.

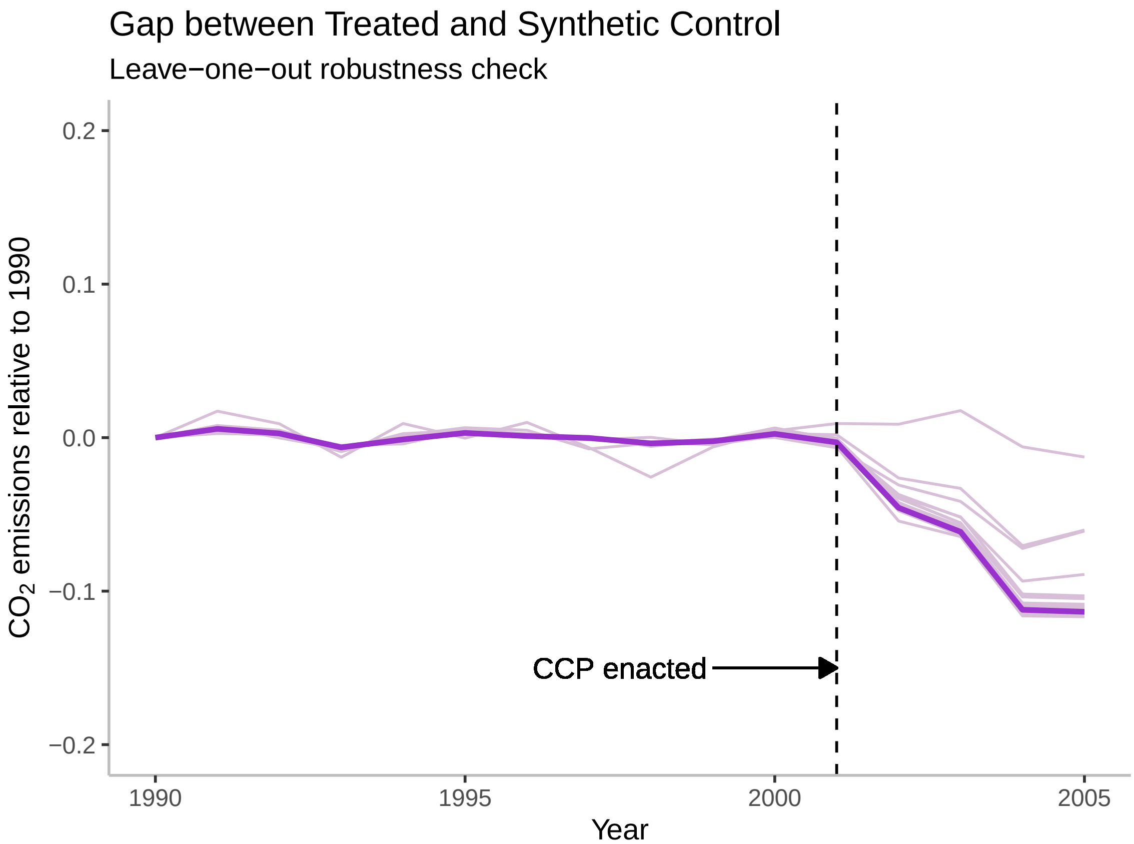


Figure A.94: Gaps between the UK and the synthetic UK in Specification 7. The thick purple line represents the gaps when the synthetic UK is constructed using all countries in the donor pool (32 countries). Each thin purple line represents the gaps when one country is dropped from the donor pool.

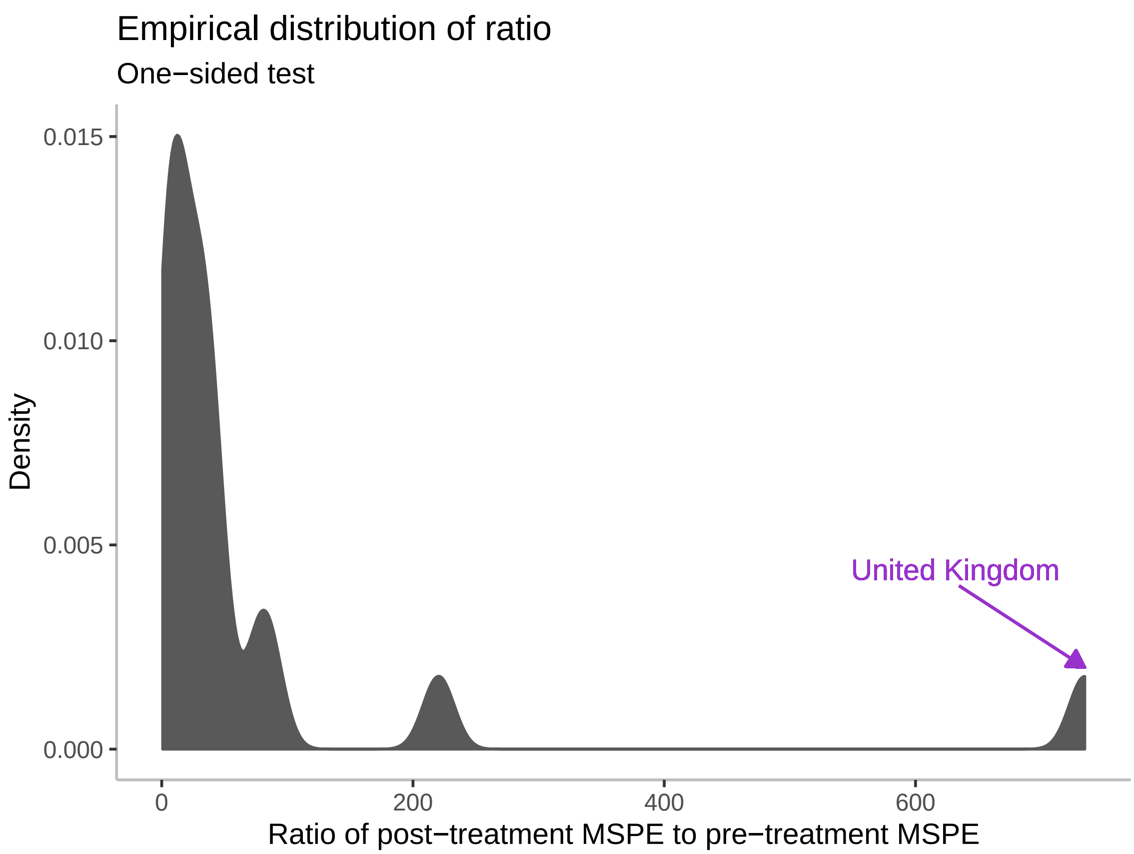


Figure A.95: Non-parametric null distribution for a one-sided test. The density represents the empirical distribution of the ratio statistic (computed as the ratio of post- to pre-treatment Mean Square Prediction Error) for all countries in the sample where the effect of the treatment is estimated to reduce emissions.

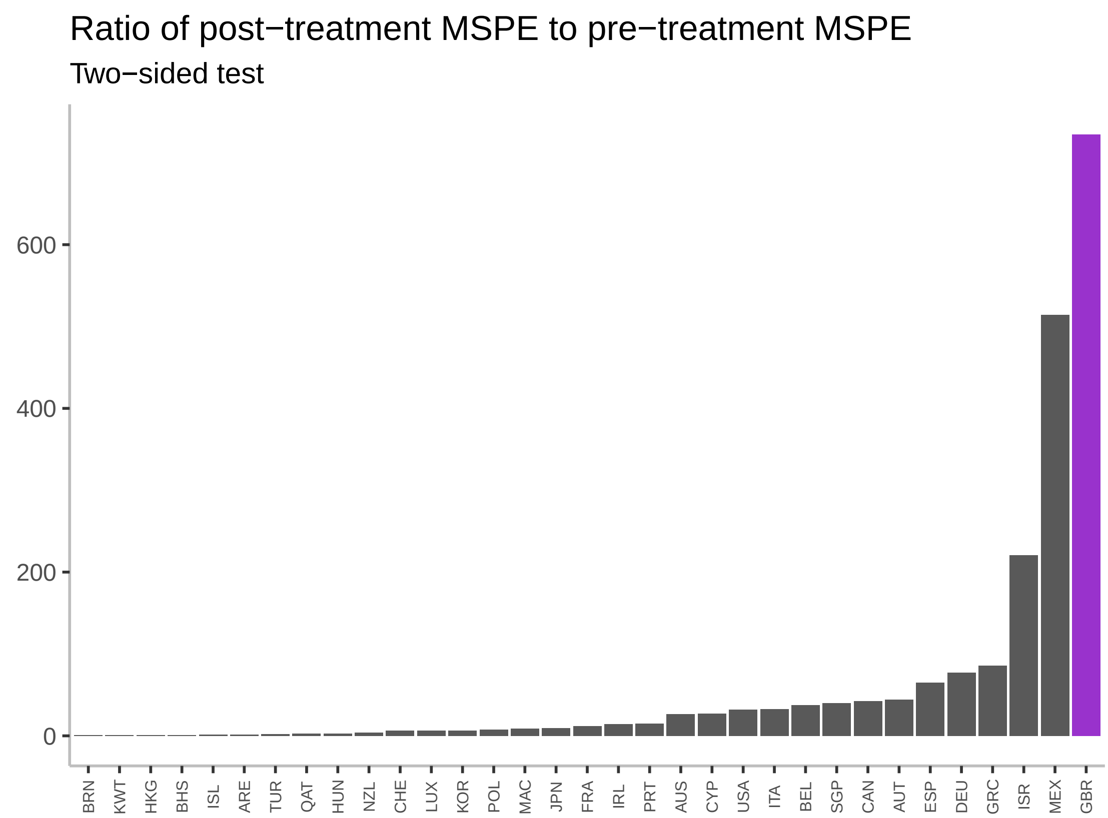


Figure A.96: Ratio of post-treatment Mean Square Prediction Error (MSPE) to pre-treatment MSPE for all countries in the sample.

A.7.8 Specification 8

Outcome variable: Emissions rescaled to 1990 baseline

Donor pool: OECD income countries in 2001, $n = 22$

Covariates: No

Optimization period: 1990-2001

Predictor	Treated UK	Synthetic UK	Sample Mean	Weight
1991 emissions rescaled to 1990 baseline	1.02	1.014	1.007	0.003
1992 emissions rescaled to 1990 baseline	1.001	1	1.014	0.046
1993 emissions rescaled to 1990 baseline	0.982	0.988	1.025	0.034
1994 emissions rescaled to 1990 baseline	0.983	0.982	1.034	0.04
1995 emissions rescaled to 1990 baseline	0.968	0.958	1.048	0.071
1996 emissions rescaled to 1990 baseline	0.992	0.99	1.09	0.119
1997 emissions rescaled to 1990 baseline	0.949	0.959	1.11	0.146
1998 emissions rescaled to 1990 baseline	0.957	0.958	1.113	0.15
1999 emissions rescaled to 1990 baseline	0.955	0.956	1.135	0.284
2000 emissions rescaled to 1990 baseline	0.975	0.972	1.162	0.108

Table A.16: Pre-treatment values of the predictor variables in the UK (column 2), its synthetic control (column 3), and in the unweighted sample (column 4); and weights applied to those predictor variables (column 5).

Balance statistic	
p-value two sample t-test	0.9886529
p-value Kolmogorov Smirnov test	0.9984853
Mean difference in QQ plots	0.0694444
Median difference in QQ plots	0.0833333
Maximum difference in QQ plots	0.1666667

Table A.17: Balance statistics between pre-treatment values of the dependent variable in the UK and its synthetic counterpart. p-values are reported for a two sample t-test for a difference in means and for a Kolmogorov-Smirnov test for whether the samples come from different distributions. QQ statistics are reported for the empirical CDF of both samples.

Treatment effect:

- Emissions in 2005 were 6.2% lower relative to a 1990 baseline compared to emissions *without* the CCP

Statistical significance:

- Two-sided test: $1/23 \approx 0.043$
- One-sided test: $1/14 \approx 0.071$

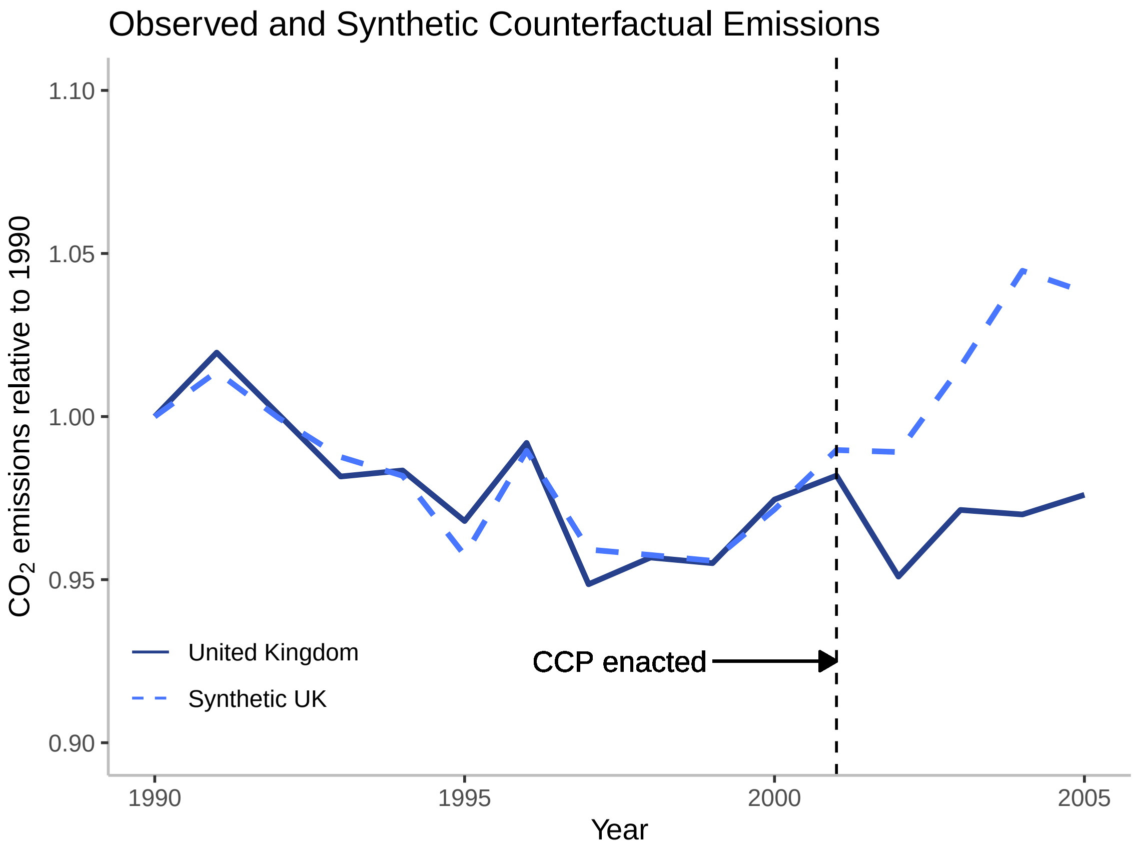


Figure A.97: Observed and synthetic counterfactual emissions for the UK. The dashed line represents the emissions trajectory of a synthetic UK as estimated by Specification 8.

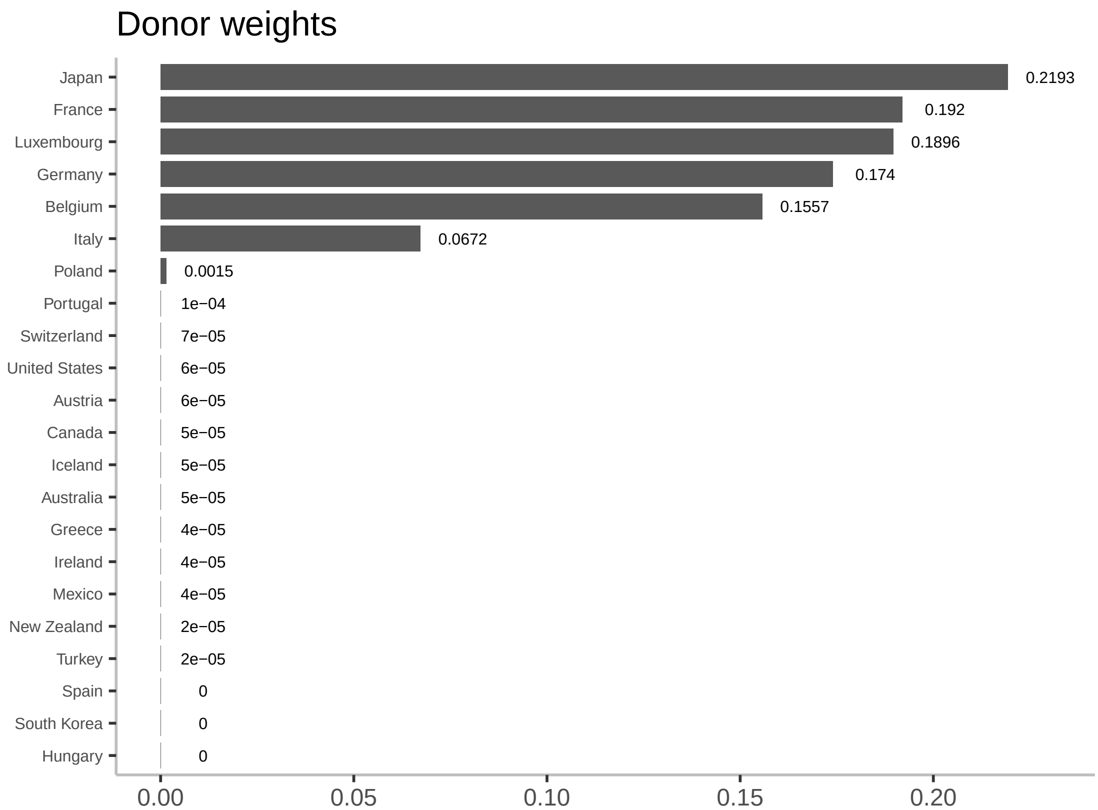


Figure A.98: Weights applied to donor countries in Specification 8.

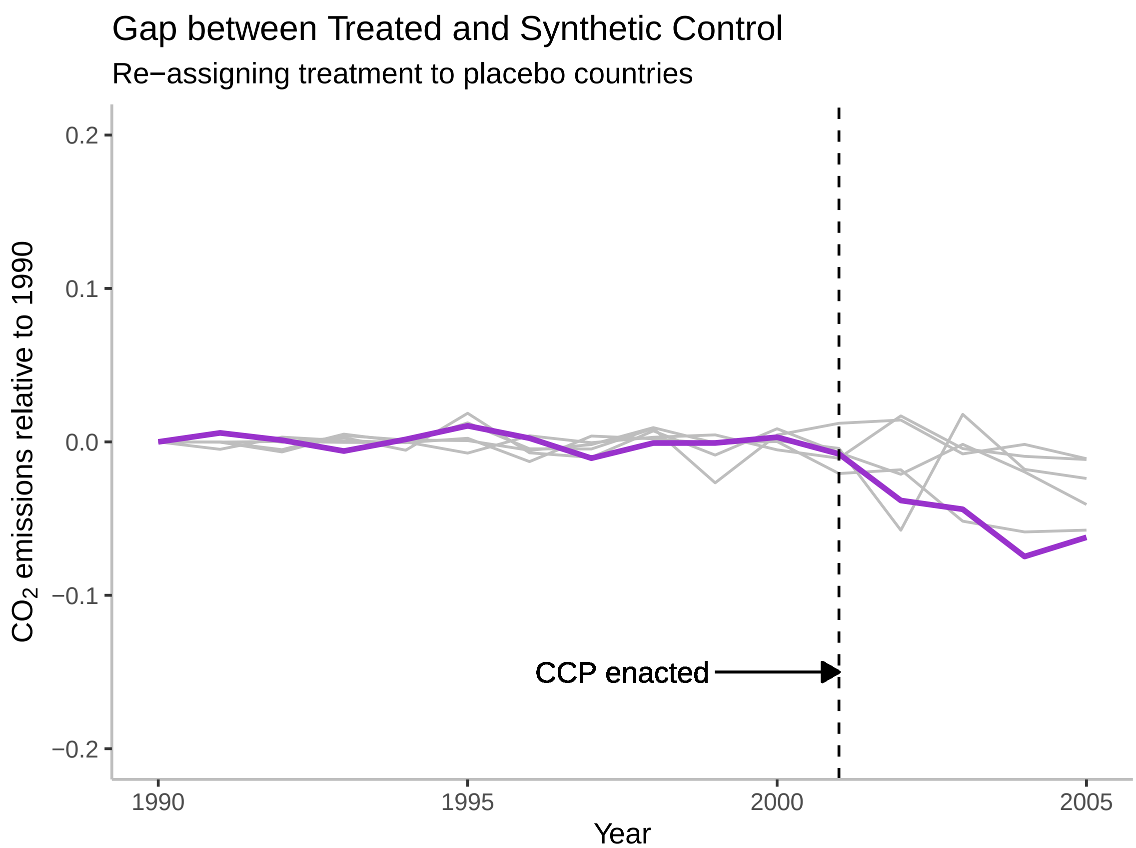


Figure A.99: Gaps in emissions (rescaled to 1990 baseline) between the treated unit and its synthetic counterpart as estimated by Specification 8. The grey lines represent the gaps in emissions for placebo countries. Countries with a pre-treatment MSPE greater than 5 times that of the UK have been excluded.

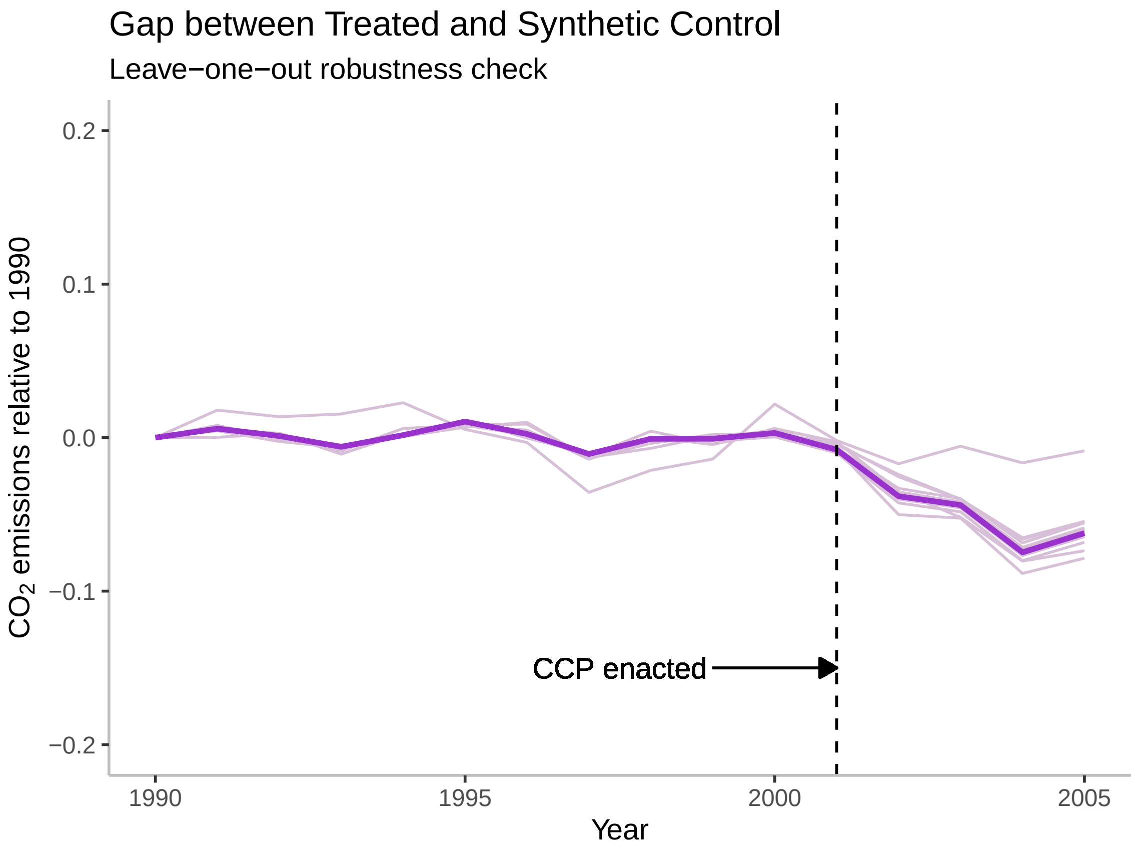


Figure A.100: Gaps between the UK and the synthetic UK in Specification 8. The thick purple line represents the gaps when the synthetic UK is constructed using all countries in the donor pool (22 countries). Each thin purple line represents the gaps when one country is dropped from the donor pool.

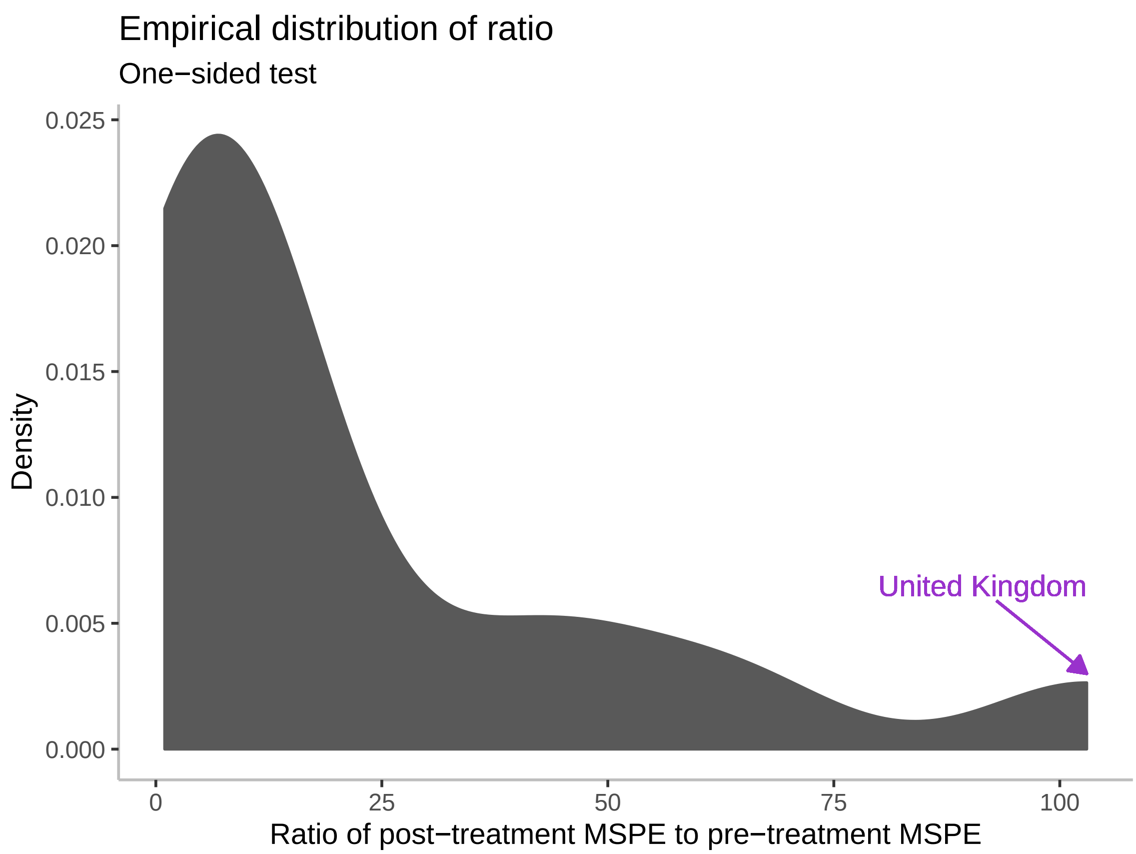


Figure A.101: Non-parametric null distribution for a one-sided test. The density represents the empirical distribution of the ratio statistic (computed as the ratio of post- to pre-treatment Mean Square Prediction Error) for all countries in the sample where the effect of the treatment is estimated to reduce emissions.

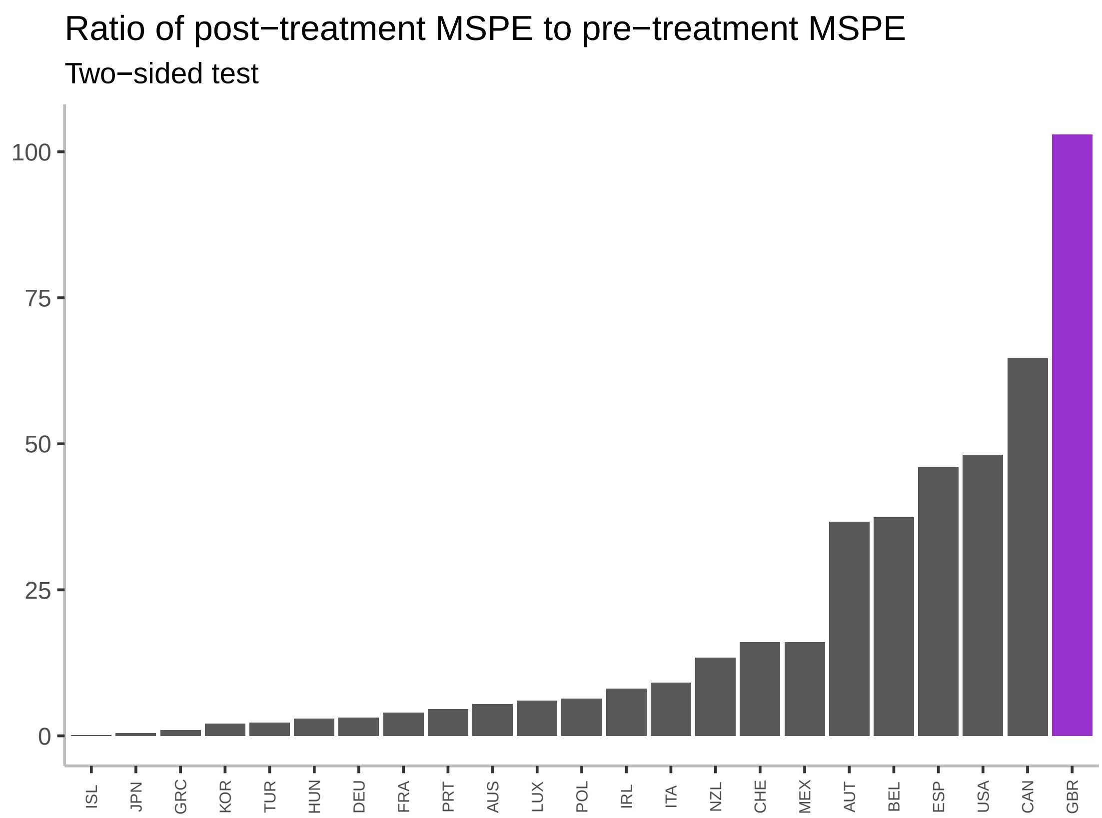


Figure A.102: Ratio of post-treatment Mean Square Prediction Error (MSPE) to pre-treatment MSPE for all countries in the sample.

A.7.9 Specification 9

Outcome variable: Emissions rescaled to 2000 baseline

Donor pool: OECD, high, and upper middle income countries in 2001, $n = 51$

Covariates: No

Optimization period: 1990-2001

Predictor	Treated UK	Synthetic UK	Sample Mean	Weight
1990 emissions rescaled to 2000 baseline	1.026	1.027	0.801	0.208
1991 emissions rescaled to 2000 baseline	1.046	1.044	0.804	0.184
1992 emissions rescaled to 2000 baseline	1.027	1.025	0.827	0.169
1993 emissions rescaled to 2000 baseline	1.007	1.012	0.863	0.106
1994 emissions rescaled to 2000 baseline	1.009	1.009	0.867	0.11
1995 emissions rescaled to 2000 baseline	0.993	0.993	0.873	0.093
1996 emissions rescaled to 2000 baseline	1.018	1.017	0.906	0.046
1997 emissions rescaled to 2000 baseline	0.973	0.973	0.934	0.064
1998 emissions rescaled to 2000 baseline	0.982	0.988	0.959	0.009
1999 emissions rescaled to 2000 baseline	0.98	0.982	0.975	0.011

Table A.18: Pre-treatment values of the predictor variables in the UK (column 2), its synthetic control (column 3), and in the unweighted sample (column 4); and weights applied to those predictor variables (column 5).

Balance statistic	
p-value two sample t-test	0.9041184
p-value Kolmogorov Smirnov test	0.8689817
Mean difference in QQ plots	0.0694444
Median difference in QQ plots	0.0833333
Maximum difference in QQ plots	0.25

Table A.19: Balance statistics between pre-treatment values of the dependent variable in the UK and its synthetic counterpart. p-values are reported for a two sample t-test for a difference in means and for a Kolmogorov-Smirnov test for whether the samples come from different distributions. QQ statistics are reported for the empirical CDF of both samples.

Treatment effect:

- Emissions in 2005 were 13.6% lower relative to a 2000 baseline compared to emissions *without* the CCP

Statistical significance:

- Two-sided test: $2/52 \approx 0.038$
- One-sided test: $1/33 \approx 0.030$

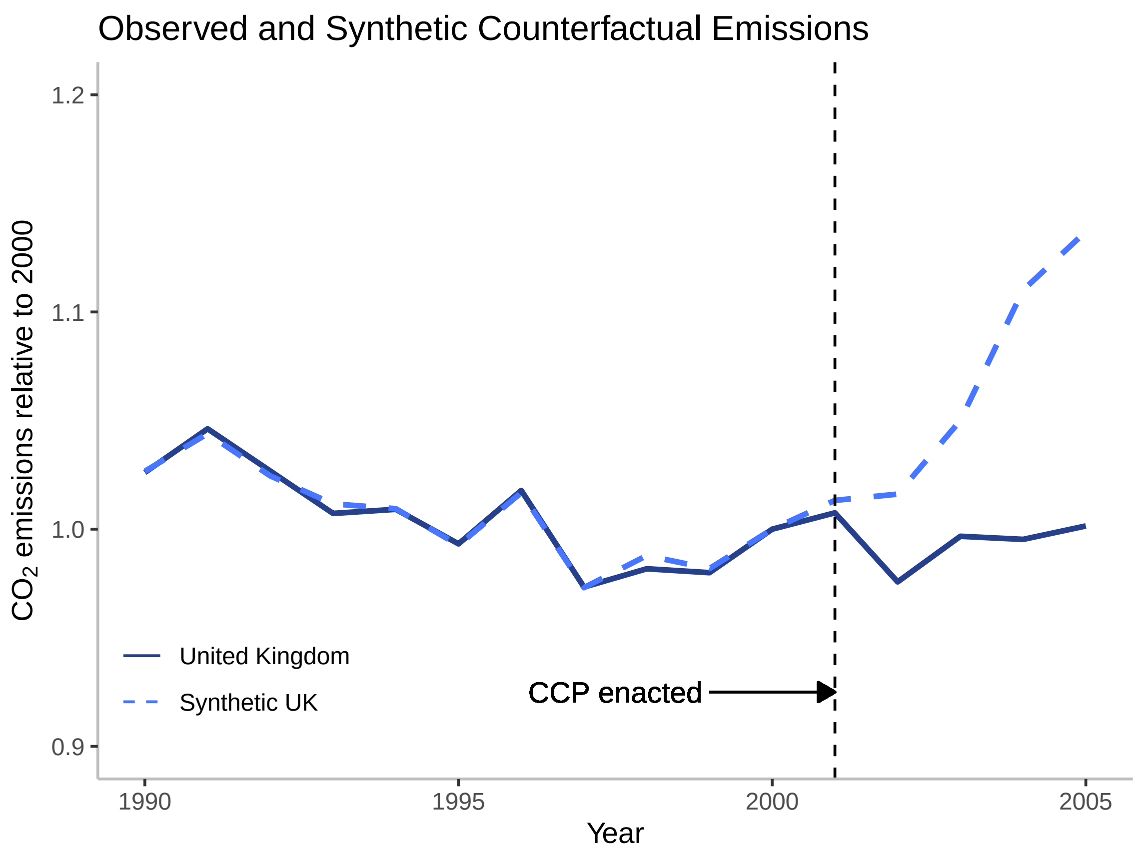


Figure A.103: Observed and synthetic counterfactual emissions for the UK. The dashed line represents the emissions trajectory of a synthetic UK as estimated by Specification 9.

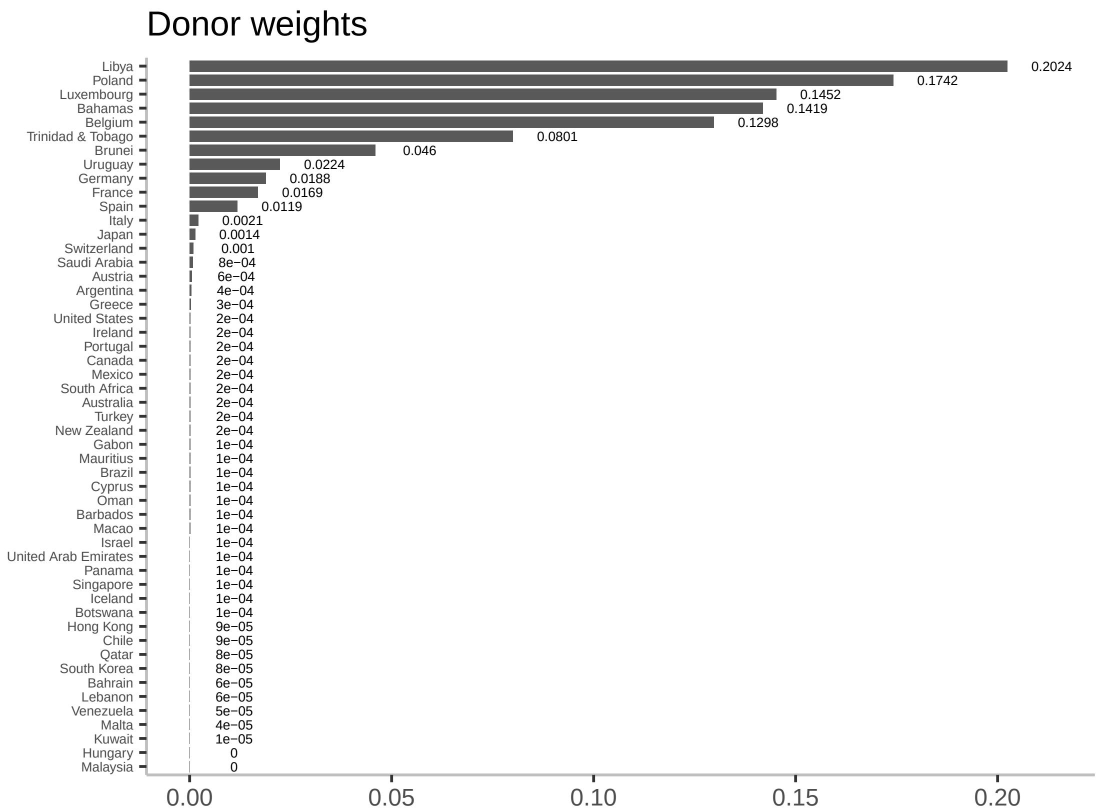


Figure A.104: Weights applied to donor countries in Specification 9.

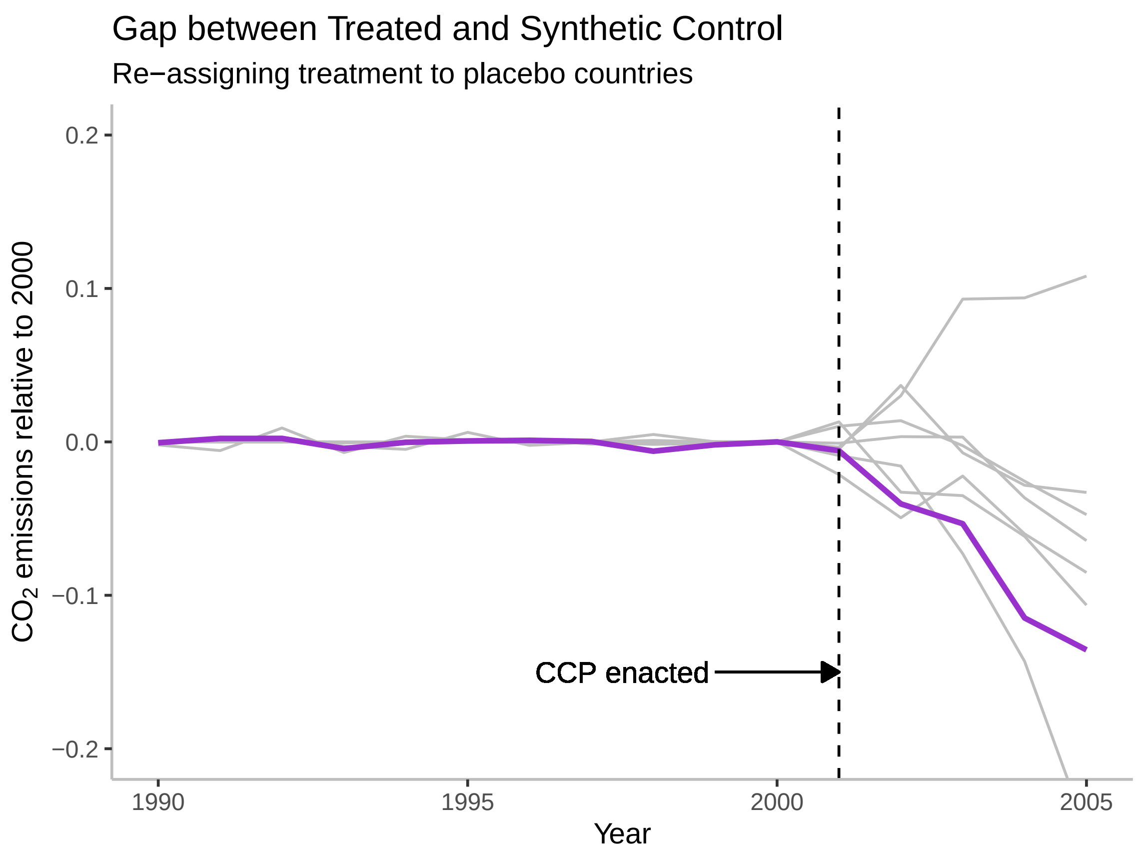


Figure A.105: Gaps in emissions (rescaled to 2000 baseline) between the treated unit and its synthetic counterpart as estimated by Specification 9. The grey lines represent the gaps in emissions for placebo countries. Countries with a pre-treatment MSPE greater than 5 times that of the UK have been excluded.

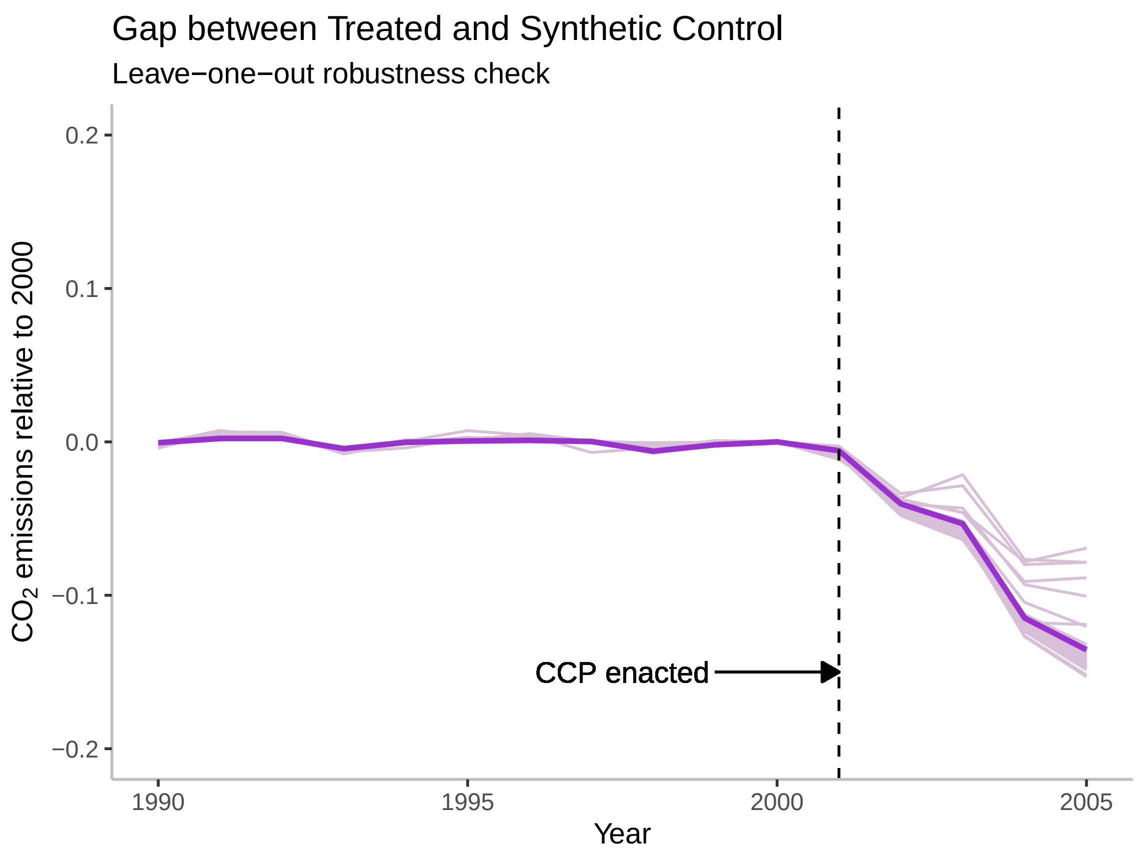


Figure A.106: Gaps between the UK and the synthetic UK in Specification 9. The thick purple line represents the gaps when the synthetic UK is constructed using all countries in the donor pool (51 countries). Each thin purple line represents the gaps when one country is dropped from the donor pool.

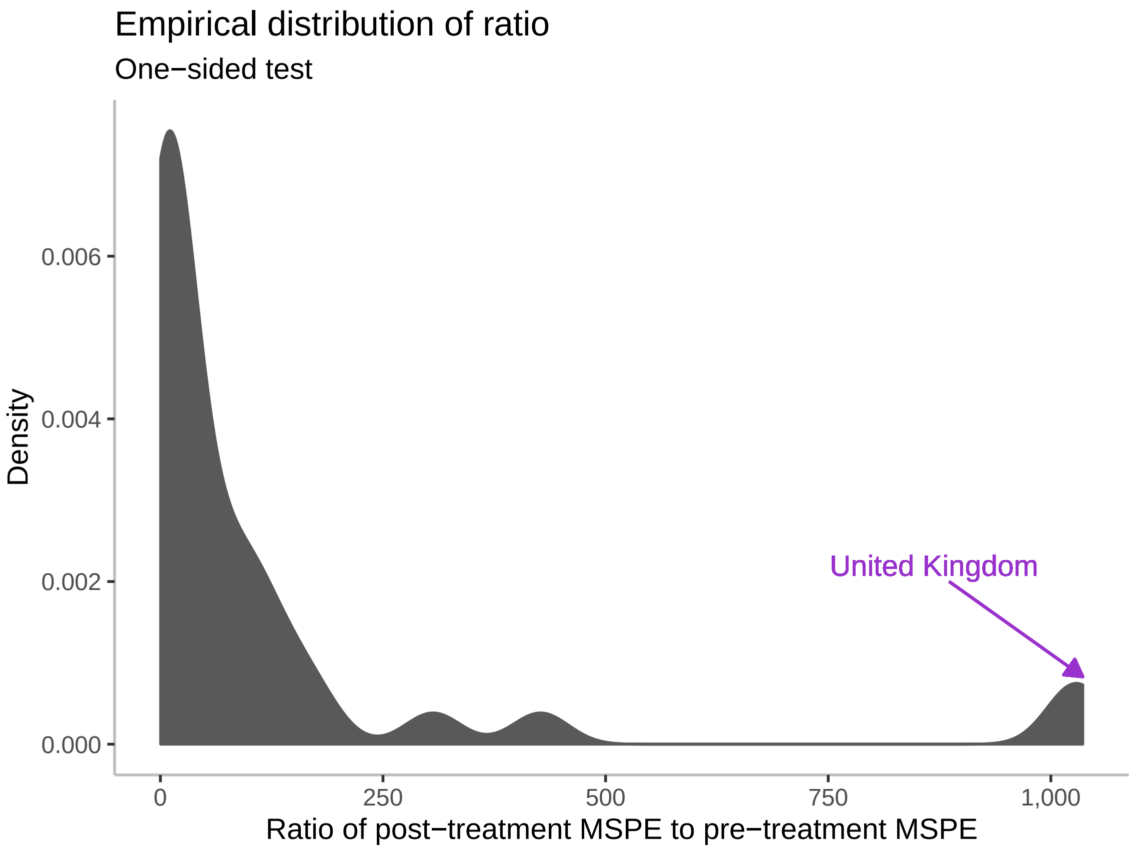


Figure A.107: Non-parametric null distribution for a one-sided test. The density represents the empirical distribution of the ratio statistic (computed as the ratio of post- to pre-treatment Mean Square Prediction Error) for all countries in the sample where the effect of the treatment is estimated to reduce emissions.

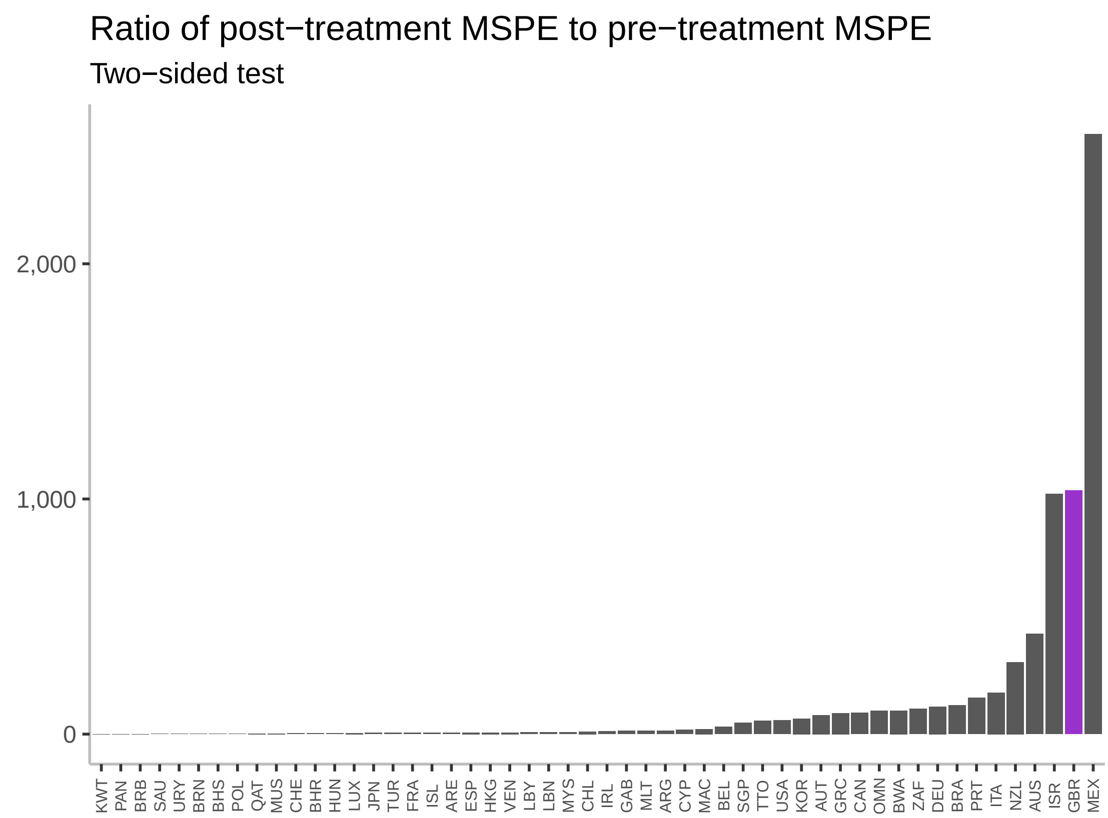


Figure A.108: Ratio of post-treatment Mean Square Prediction Error (MSPE) to pre-treatment MSPE for all countries in the sample.

A.7.10 Specification 10

Outcome variable: Emissions rescaled to 2000 baseline

Donor pool: OECD and high income countries in 2001, $n = 32$

Covariates: No

Optimization period: 1990-2001

Predictor	Treated UK	Synthetic UK	Sample Mean	Weight
1990 emissions rescaled to 2000 baseline	1.026	1.029	0.86	0.154
1991 emissions rescaled to 2000 baseline	1.046	1.037	0.84	0.091
1992 emissions rescaled to 2000 baseline	1.027	1.023	0.865	0.116
1993 emissions rescaled to 2000 baseline	1.007	1.013	0.887	0.087
1994 emissions rescaled to 2000 baseline	1.009	1.011	0.9	0.071
1995 emissions rescaled to 2000 baseline	0.993	0.991	0.902	0.101
1996 emissions rescaled to 2000 baseline	1.018	1.016	0.926	0.1
1997 emissions rescaled to 2000 baseline	0.973	0.974	0.946	0.174
1998 emissions rescaled to 2000 baseline	0.982	0.983	0.964	0.042
1999 emissions rescaled to 2000 baseline	0.98	0.981	0.965	0.064

Table A.20: Pre-treatment values of the predictor variables in the UK (column 2), its synthetic control (column 3), and in the unweighted sample (column 4); and weights applied to those predictor variables (column 5).

Balance statistic	
p-value two sample t-test	0.9931237
p-value Kolmogorov Smirnov test	0.8689817
Mean difference in QQ plots	0.0625
Median difference in QQ plots	0.0833333
Maximum difference in QQ plots	0.25

Table A.21: Balance statistics between pre-treatment values of the dependent variable in the UK and its synthetic counterpart. p-values are reported for a two sample t-test for a difference in means and for a Kolmogorov-Smirnov test for whether the samples come from different distributions. QQ statistics are reported for the empirical CDF of both samples.

Treatment effect:

- Emissions in 2005 were 12.1% lower relative to a 2000 baseline compared to emissions *without* the CCP

Statistical significance:

- Two-sided test: $1/33 \approx 0.030$
- One-sided test: $1/18 \approx 0.056$

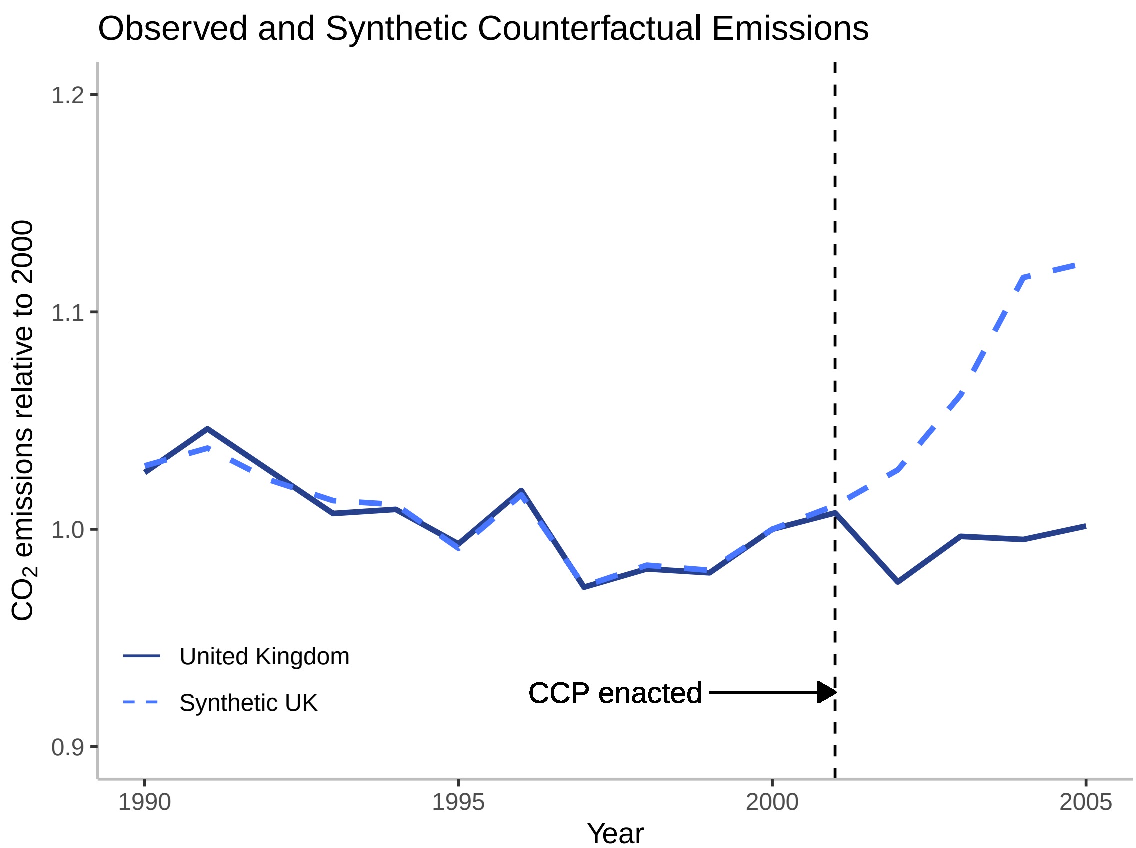


Figure A.109: Observed and synthetic counterfactual emissions for the UK. The dashed line represents the emissions trajectory of a synthetic UK as estimated by Specification 10.

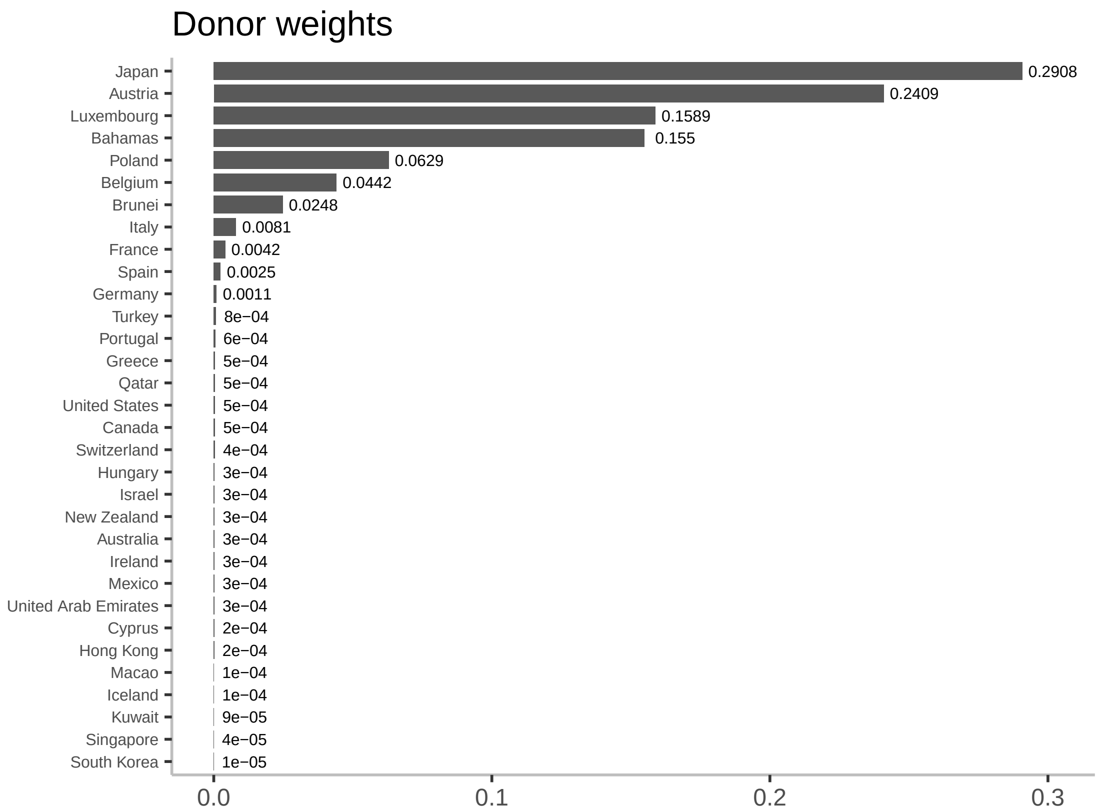


Figure A.110: Weights applied to donor countries in Specification 10.

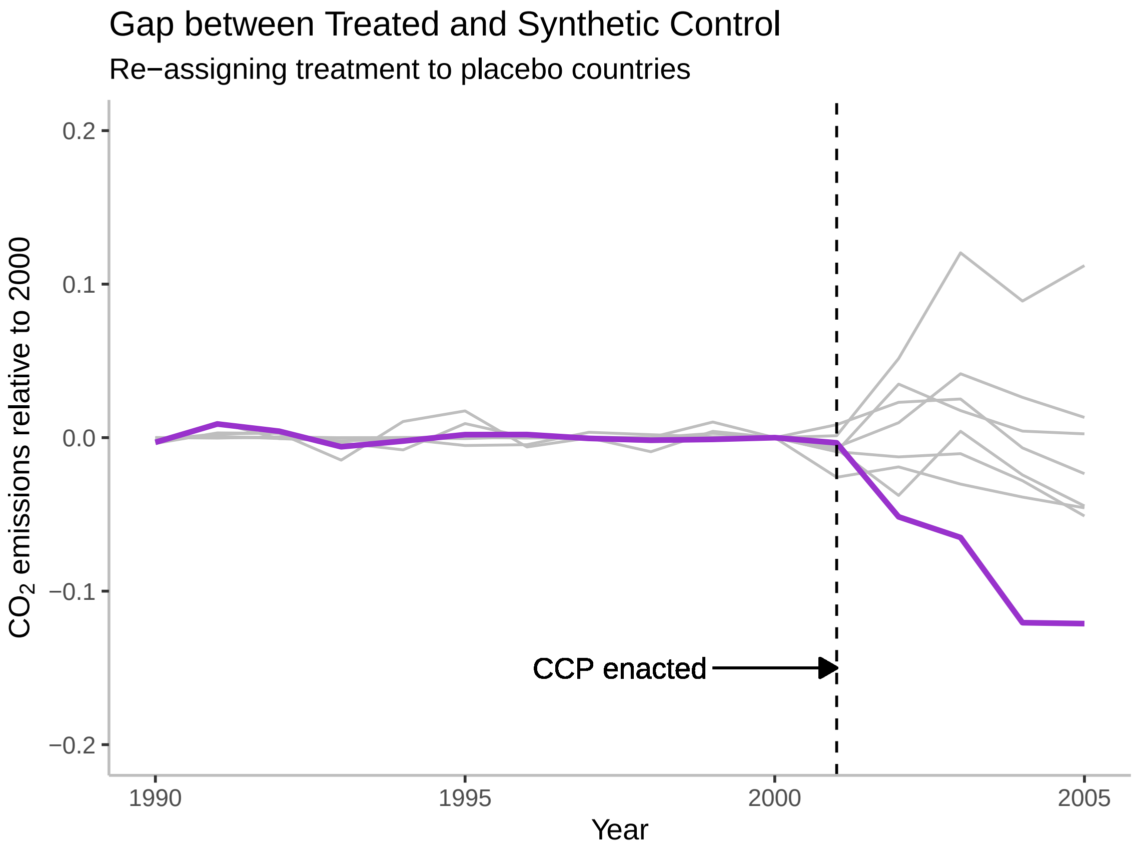


Figure A.111: Gaps in emissions (rescaled to 2000 baseline) between the treated unit and its synthetic counterpart as estimated by Specification 10. The grey lines represent the gaps in emissions for placebo countries. Countries with a pre-treatment MSPE greater than 5 times that of the UK have been excluded.

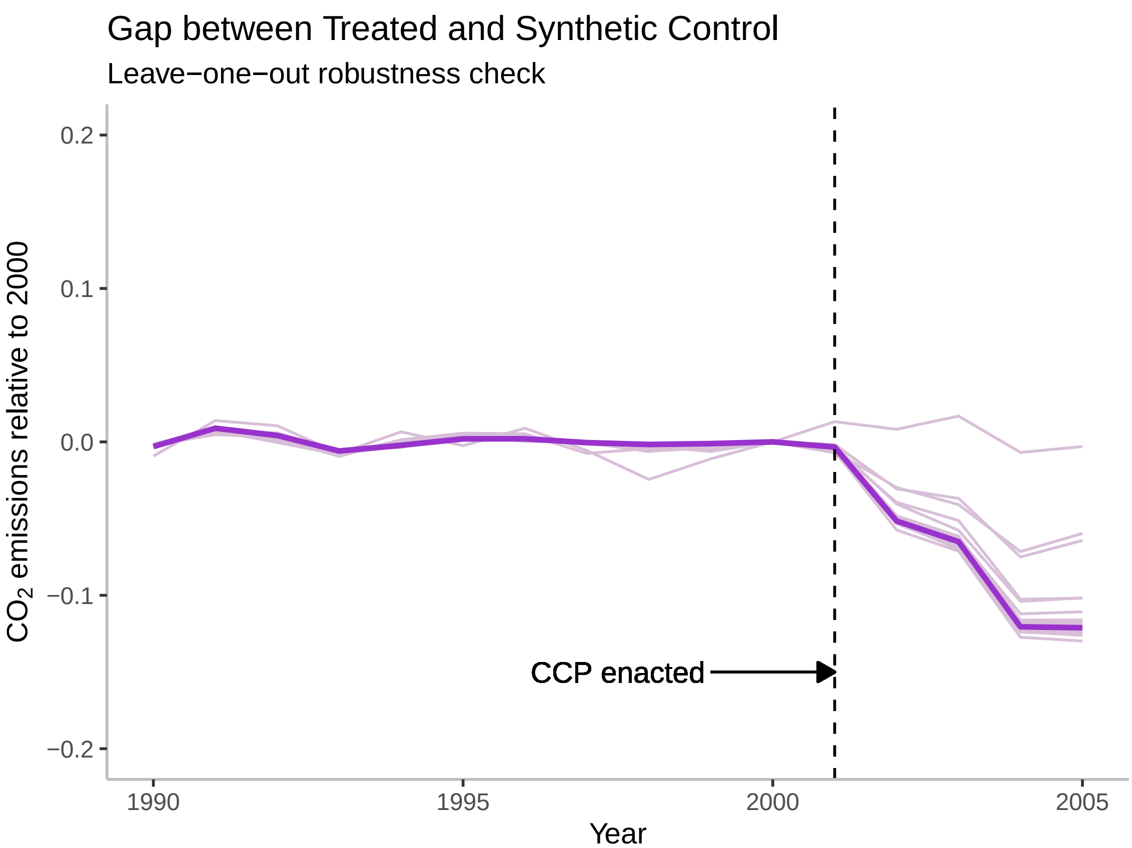


Figure A.112: Gaps between the UK and the synthetic UK in Specification 10. The thick purple line represents the gaps when the synthetic UK is constructed using all countries in the donor pool (32 countries). Each thin purple line represents the gaps when one country is dropped from the donor pool.

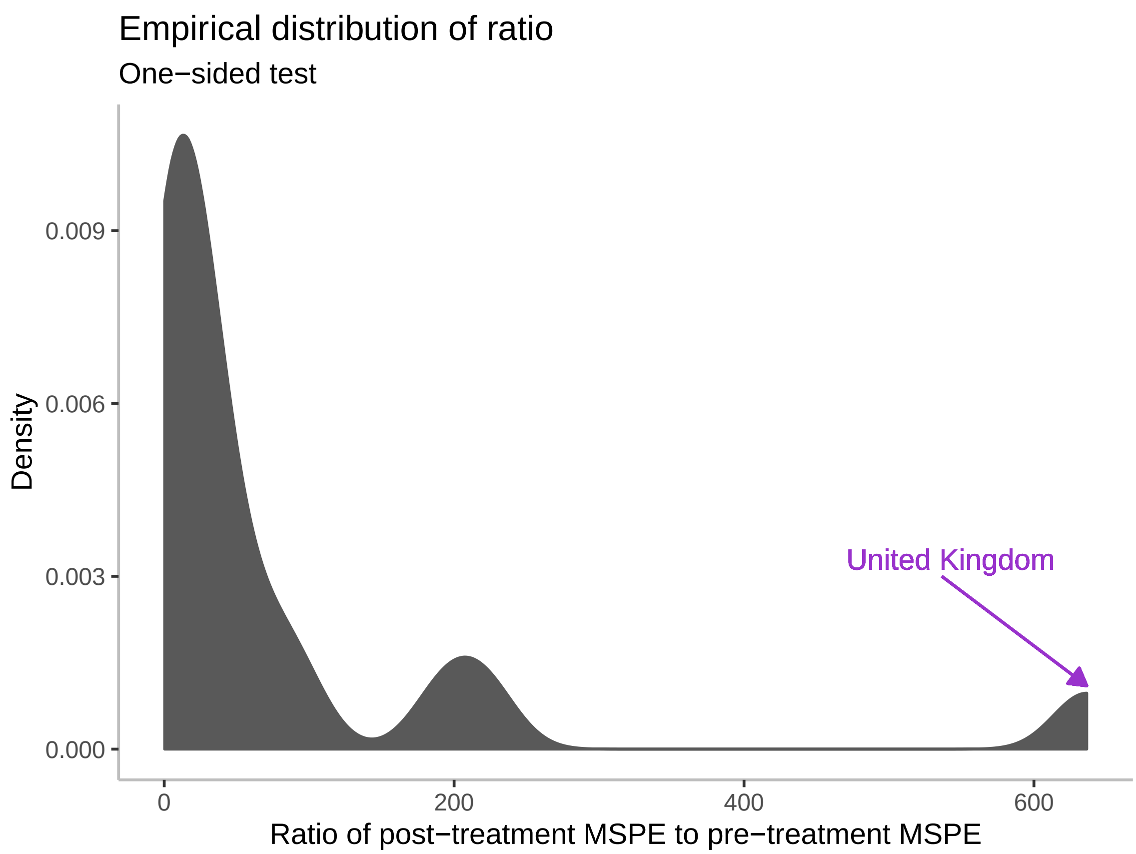


Figure A.113: Non-parametric null distribution for a one-sided test. The density represents the empirical distribution of the ratio statistic (computed as the ratio of post- to pre-treatment Mean Square Prediction Error) for all countries in the sample where the effect of the treatment is estimated to reduce emissions.

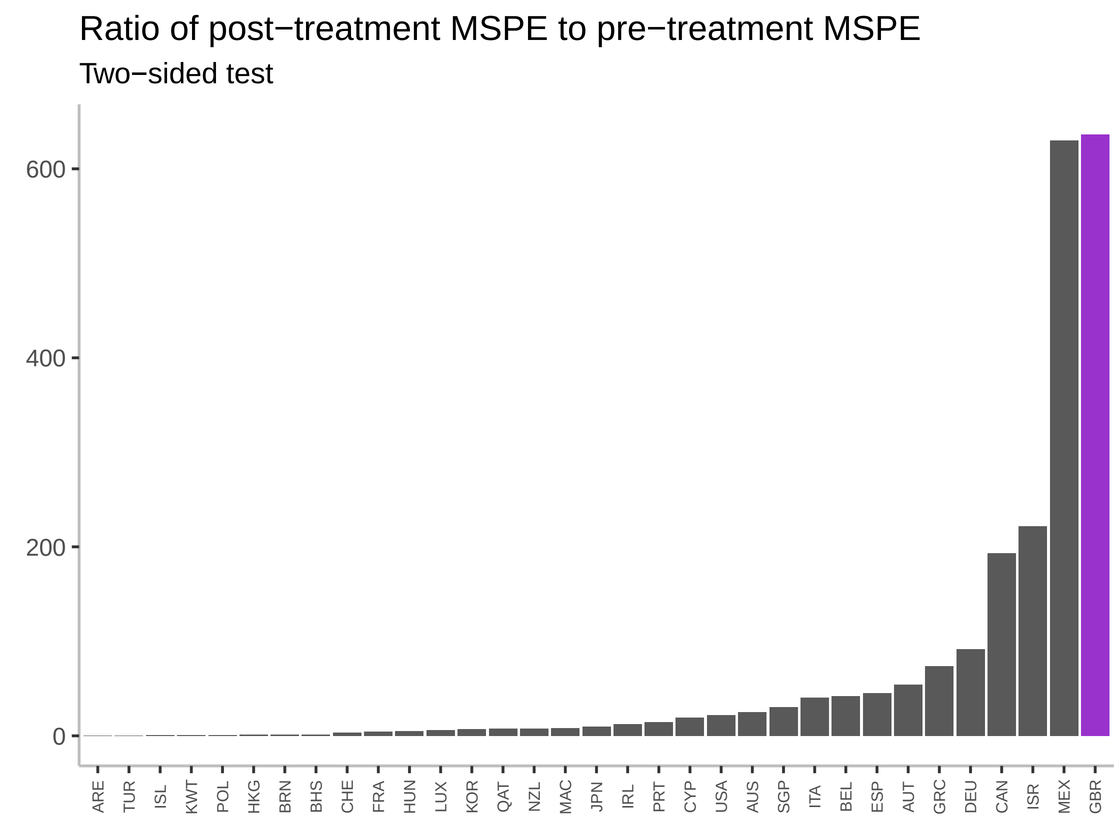


Figure A.114: Ratio of post-treatment Mean Square Prediction Error (MSPE) to pre-treatment MSPE for all countries in the sample.

A.7.11 Specification 11

Outcome variable: Emissions rescaled to 2000 baseline

Donor pool: OECD countries in 2001, $n = 22$

Covariates: No

Optimization period: 1990-2001

Predictor	Treated UK	Synthetic UK	Sample Mean	Weight
1990 emissions rescaled to 2000 baseline	1.026	1.029	0.901	0.189
1991 emissions rescaled to 2000 baseline	1.046	1.041	0.905	0.215
1992 emissions rescaled to 2000 baseline	1.027	1.026	0.904	0.13
1993 emissions rescaled to 2000 baseline	1.007	1.014	0.911	0.163
1994 emissions rescaled to 2000 baseline	1.009	1.007	0.914	0.092
1995 emissions rescaled to 2000 baseline	0.993	0.985	0.92	0.103
1996 emissions rescaled to 2000 baseline	1.018	1.014	0.955	0.051
1997 emissions rescaled to 2000 baseline	0.973	0.986	0.966	0.033
1998 emissions rescaled to 2000 baseline	0.982	0.984	0.969	0.016
1999 emissions rescaled to 2000 baseline	0.98	0.983	0.982	0.007

Table A.22: Pre-treatment values of the predictor variables in the UK (column 2), its synthetic control (column 3), and in the unweighted sample (column 4); and weights applied to those predictor variables (column 5).

Balance statistic	
p-value two sample t-test	0.863521
p-value Kolmogorov Smirnov test	0.8689817
Mean difference in QQ plots	0.0694444
Median difference in QQ plots	0.0833333
Maximum difference in QQ plots	0.25

Table A.23: Balance statistics between pre-treatment values of the dependent variable in the UK and its synthetic counterpart. p-values are reported for a two sample t-test for a difference in means and for a Kolmogorov-Smirnov test for whether the samples come from different distributions. QQ statistics are reported for the empirical CDF of both samples.

Treatment effect:

- Emissions in 2005 were 6.9% lower relative to a 2000 baseline compared to emissions *without* the CCP

Statistical significance:

- Two-sided test: $1/23 \approx 0.043$
- One-sided test: $1/15 \approx 0.067$

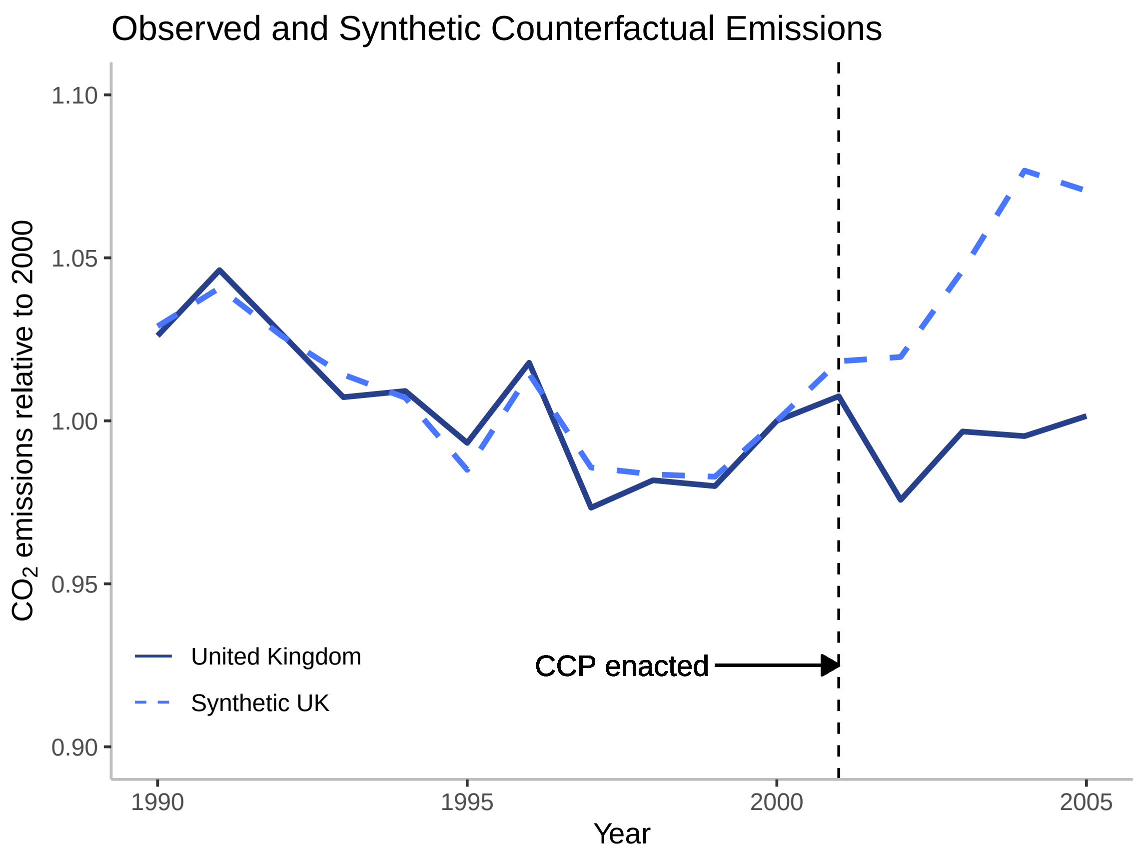


Figure A.115: Observed and synthetic counterfactual emissions for the UK. The dashed line represents the emissions trajectory of a synthetic UK as estimated by Specification 11.

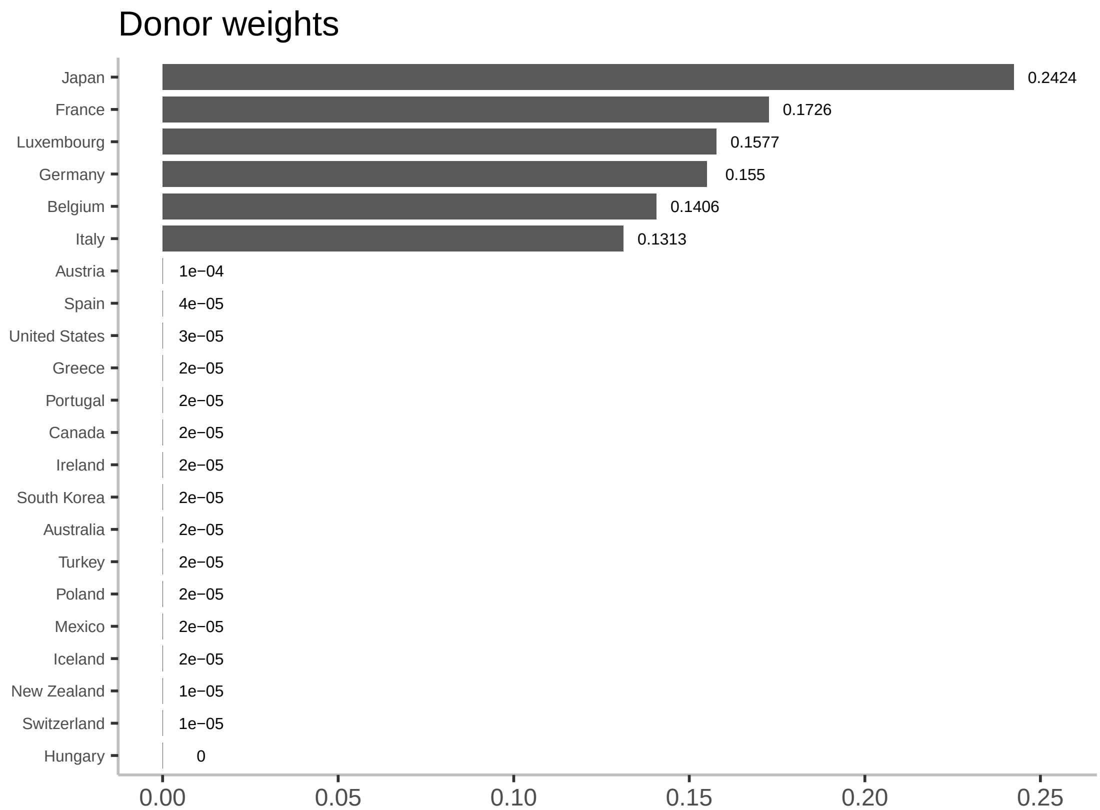


Figure A.116: Weights applied to donor countries in Specification 11.

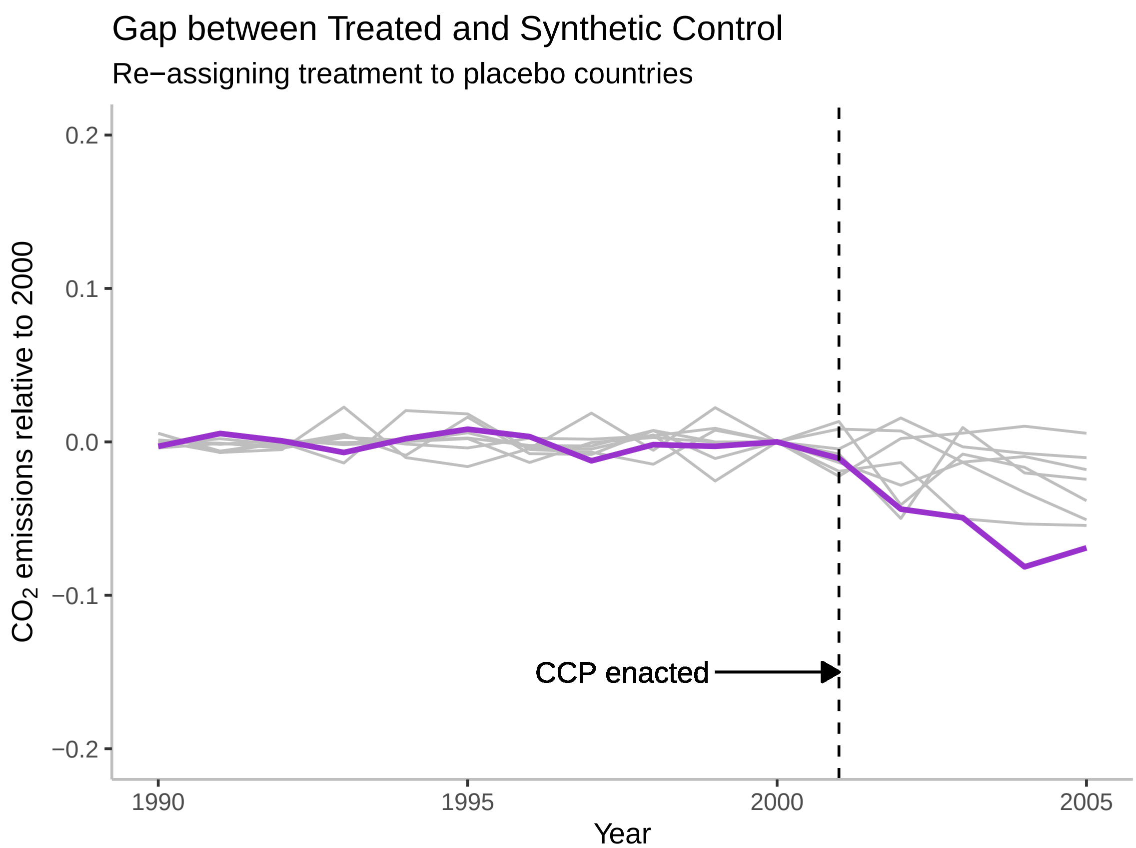


Figure A.117: Gaps in emissions (rescaled to 2000 baseline) between the treated unit and its synthetic counterpart as estimated by Specification 11. The grey lines represent the gaps in emissions for placebo countries. Countries with a pre-treatment MSPE greater than 5 times that of the UK have been excluded.

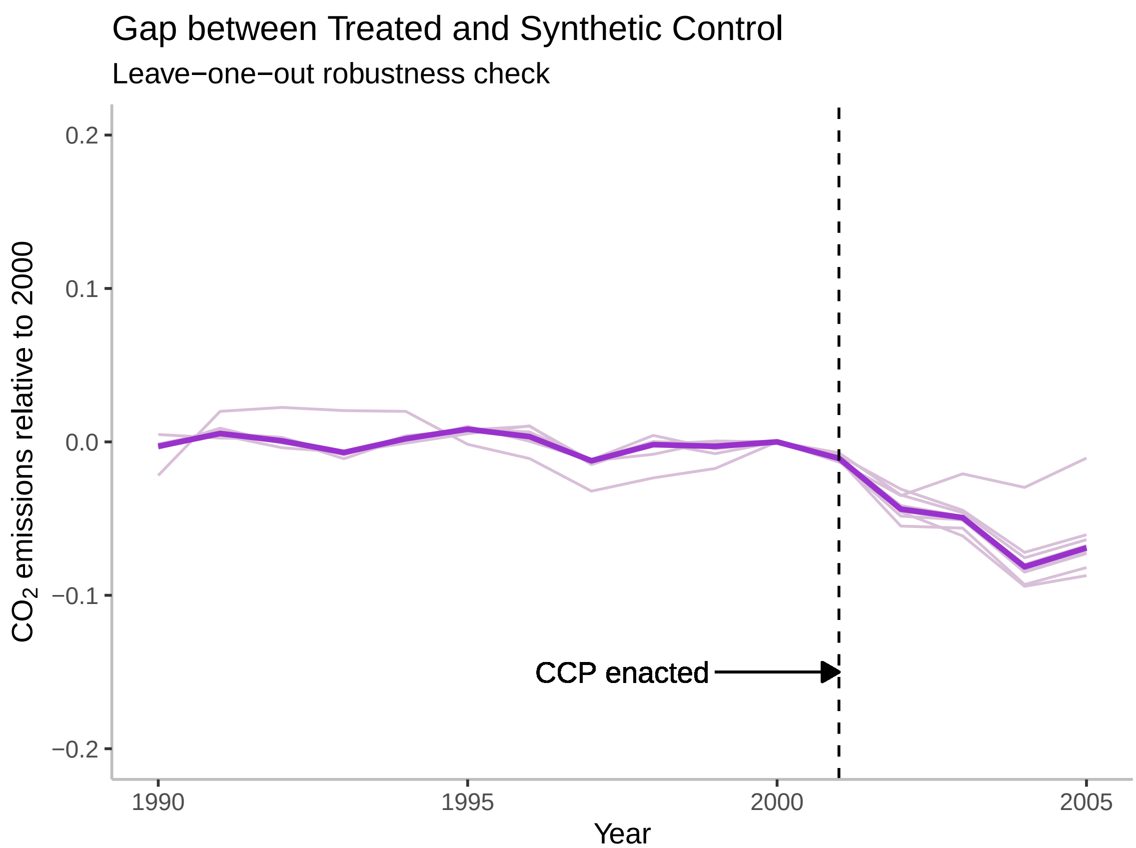


Figure A.118: Gaps between the UK and the synthetic UK in Specification 11. The thick purple line represents the gaps when the synthetic UK is constructed using all countries in the donor pool (22 countries). Each thin purple line represents the gaps when one country is dropped from the donor pool.

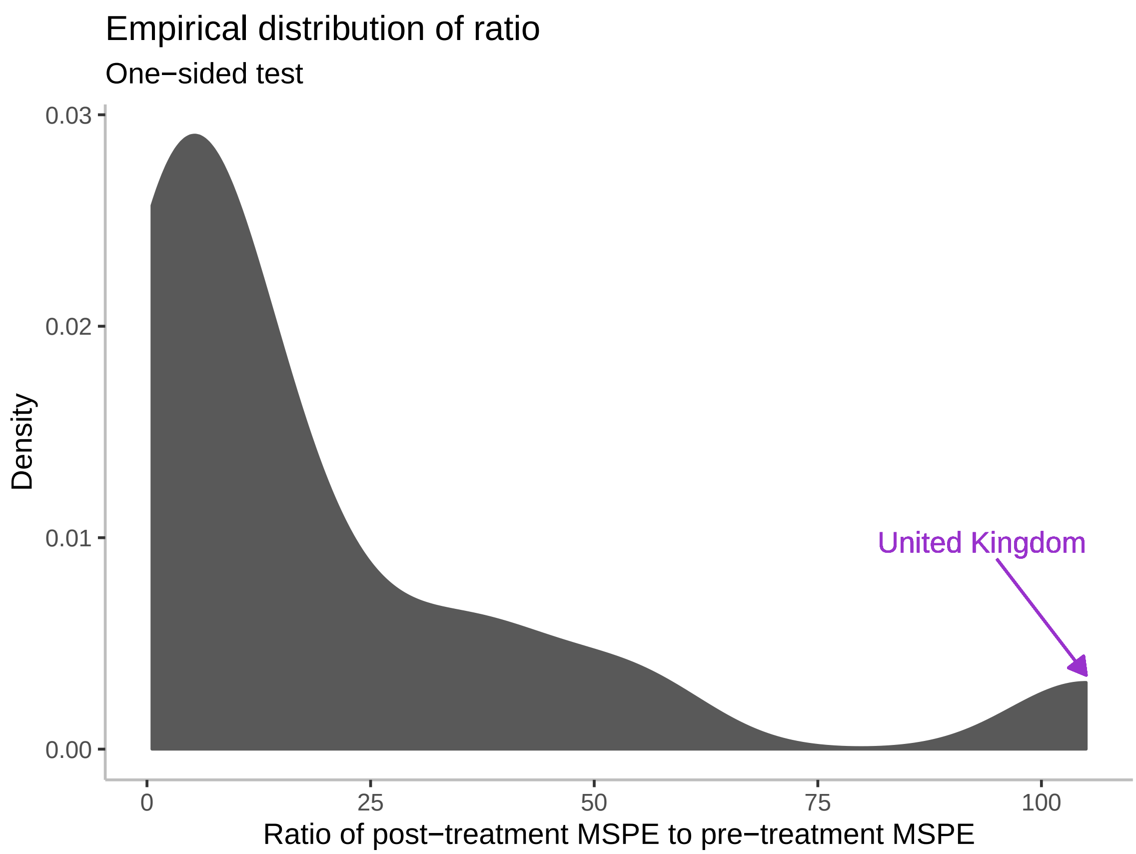


Figure A.119: Non-parametric null distribution for a one-sided test. The density represents the empirical distribution of the ratio statistic (computed as the ratio of post- to pre-treatment Mean Square Prediction Error) for all countries in the sample where the effect of the treatment is estimated to reduce emissions.

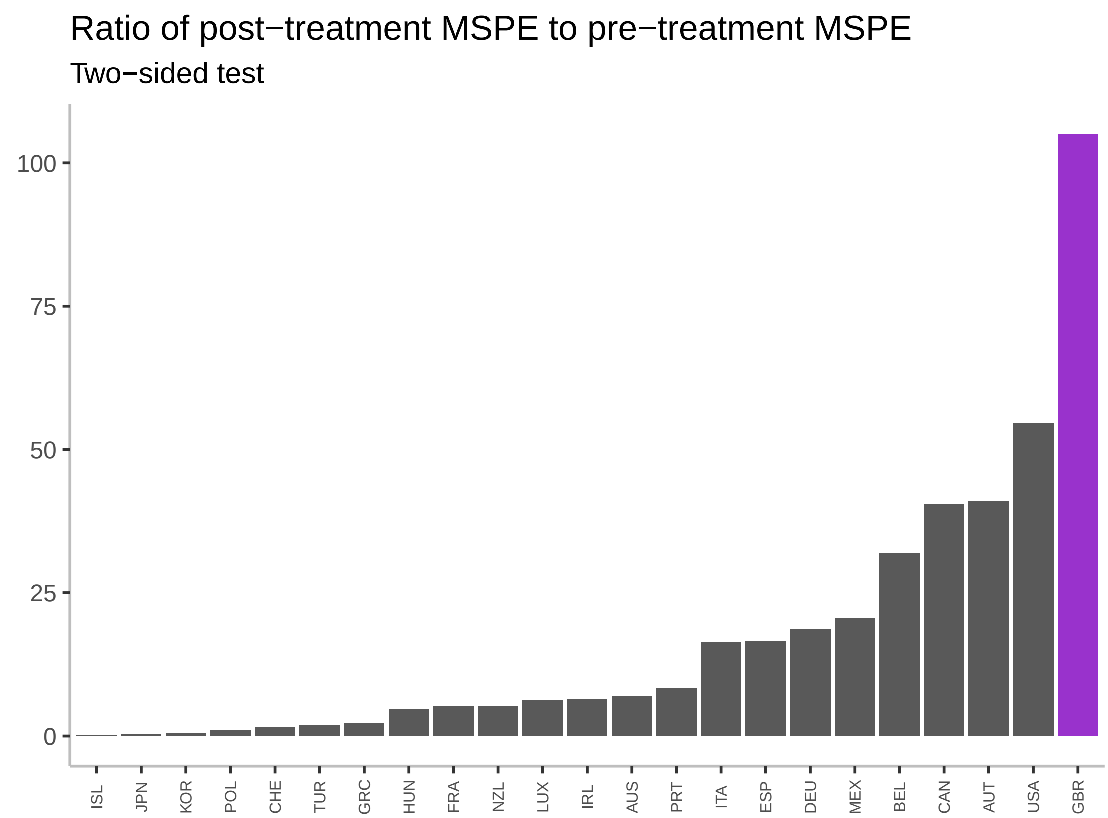


Figure A.120: Ratio of post-treatment Mean Square Prediction Error (MSPE) to pre-treatment MSPE for all countries in the sample.

A.8 The landscape of the UK's climate change policies, 1988-2015

Climate change emerged onto the British policy-making agenda in the late 1980s years during Margaret Thatcher's Conservative government. Thatcher herself surprised observers by personally advocating for climate action and supporting such global institutions as the Intergovernmental Panel on Climate Change (IPCC)[1]. Thatcher's climate

[1] For instance, see Thatcher's General Assembly address on 8 November 1989: https://www.theguardian.com/environment/2005/jun/30/climatechange.climatechangeenvironment1.

motives have been the subject of substantial debate; some suggest her science degree predisposed her to trust climate scientists while others suggest she embraced the climate file to boost her international profile (Oshitani, 2006). Whatever the reason, her efforts legitimized climate change on the British political agenda, and led to funding for new climate science research. Under Thatcher, the UK also set its first carbon pollution reduction target: carbon pollution stabilization at 1990 levels by 2005.

In 1995, under Major, Environment Minister John Gummer eventually got cabinet approval for the more ambitious target of 5-10% reduction below 1990 levels by 2010; cabinet target support was apparently a function of promises that the targets would meet themselves without policy interventions (Oshitani, 2006). Despite these targets – and contemporaneous declines in British carbon pollution – Conservative governments across the 1990s enacted few deliberate climate policies. While open to climate science, Thatcher remained wary of climate mitigation measures. Post-government, she would take a skeptical tone and criticize climate policy instruments as costly agents of socialism (Thatcher, 2003, 449-451). Nonetheless, as in other advanced economies, carbon taxation emerged onto the British policy-making agenda during the early 1990s, pushed forward by then Environment Minister Christopher Patten and his senior advisor David Pearce[2]. Reform efforts were backed from outside government by the environmental lobby, including Friends of the Earth. However, carbon-dependent economic actors inside and outside government stymied serious consideration of the idea. Within government, officials from the Treasury, the Department of Trade and Industry and the Department of Energy all expressed reservations. Similarly, British officials opposed EU-level consideration of a carbon tax at the time, a function both of sectoral interests and skepticism of EU-level

[2]Environmental taxes more broadly had emerged onto the agenda into the late 1980s after media attention to a Department of Energy report on the topic; the topic was then described in the Conservative government's 1992 environment White Paper (Jordan et al., 2003).

policy-making on the issue (Oshitani, 2006; Skjærseth and Wettestad, 2008). During this time, both business communities, through the Confederation of British Industry (CBI) *and* the labor community, through the Trade Unions Congress, opposed any carbon tax as a function of its potential economic and job impacts (Oshitani, 2006).

Labour took power in 1997 under Prime Minister Tony Blair. Blair immediately offered strong rhetorical commitment to climate policy-making, particularly within British foreign policy. Blair was quick to criticize American counterparts for being climate laggards, both during his first US visit for the Denver G-7 meeting and in his July 2003 address to Congress. Under Blair, the UK also upped its carbon pollution target to 20% below 1990 levels by 2010. However, Labour domestic reforms were generally more modest and centered on the Blair government's 2001 British Climate Change Programme (BCP) report of climate policy-making priorities. (The BCP was later updated in 2006). The BCP itself was partly inspired by the government-commissioned Marshall report, led by a former head of the CBI, that explored prospects for market-oriented instruments in environmental policy-making (Darkin, 2006).

Blair's Labour government embraced market-based policy instruments in a way previous Conservative administrations had not (Jordan et al., 2003). This BCP thus set the stage for such climate policy-making instruments as the UK Renewables Obligation (RO) in 2002 (a hybrid policy that combined a renewable mandate with tradeable compliance certificates) and an Energy Efficiency Commitment (a regulation directed at energy suppliers' home energy provisions). However, the most contentious policy measure was Labour's Climate Change Levy. The policy's structure and components is described in the main text.

Even with its flexible provisions, industry mobilized against the CCL. For instance, business lobbies and environmental groups clashed over the policy's projected impacts. A report by the Engineering Employers' Federation, the UK Steel Association and the Chemical Industries Association suggested the policy would kill 95,000 manufacturing jobs. By contrast, environmental groups forecasted net employment gains of 12,000 by 2002. The intensity of this business opposition apparently tempered Gordon Brown's (then British Chancellor) enthusiasm for climate policy-making over the following decade, including during his subsequent term as Prime Minister (Carter, 2014)[3]. Moreover, partly responding to dramatic fuel price protests in September 2000, Labour leaders were also acutely concerned about the consumer costs of their policies; the real yield of consumer fuel taxes including the CCL actually decreased by 4% from 2000 to 2007 (McLean, 2008).

In fact, the UK ETS itself emerged out of an effort by British industry, beginning early in Blair's first term to unsuccessfully pre-empt a then-proposed CCL. In 1998, business actors in consultation with government agreed to explore emissions trading through the Advisory Committee on Business and the Environment (Skjærseth and Wettestad, 2008). British business associations then formed an emissions trading group (ETG) in 1999 to design the architecture for a CCL alternative, backed by such companies as BP with existing in-house carbon prices and financial interests in London who wanted to exploit new market opportunities associated with Kyoto Protocol carbon markets (Jordan et al., 2003). The emergence of explicit business community splits on climate policy continued to surface throughout the early 2000s. For instance, a number of major British corporations founded the Corporate Leaders Group on Climate Change to advance business interests related to climate risk mitigation, from Lloyds to Shell to Tesco (Carter, 2014). The group's communications appear to have increased Blair's interest in undertaking cli-

[3]Labour actors also felt they had received insufficient backing from green groups during the CCL debate (Carter, 2008).

mate reforms, partly because the emergence of pro-climate business interests genuinely surprised him (Carter, 2008). Splits in British business interests were also reflected within older lobby groups. For instance, CBI set up a task-force in 2005 to work through cross-cutting cleavages within the British business community on the basis of divergent climate reform interests (Lockwood, 2013).

Domestic climate policy-making under Blair proceeded in parallel to EU-level efforts to negotiate a common climate policy. Unlike Germany, the British government was one of the strongest backers of emissions trading within the EU beginning in the late 1990s when other major actors were skeptical of EU-level action. British positioning on this issue was particularly salient given the country's previous hard opposition to EU-level carbon taxation (Skjærseth and Wettestad, 2008). British preferences may have been driven by a desire to pre-empt tax instruments and by growing domestic interest in emissions trading by business actors hoping to pre-empt carbon taxation domestically (Skjærseth and Wettestad, 2008). However, despite this idea leadership, British political officials imagined a far less ambitious EU proposal than would eventually emerge; they viewed EU-level carbon pricing as best organized through coordination of domestic systems rather than centralized policy. Accordingly, the UK lobbied the EU to design the EU ETS using a weak and voluntary architecture similar to its domestic scheme, not the substantially more ambitious scheme that would eventually emerge (Jordan et al., 2003; Skjærseth and Wettestad, 2008). In this way, the simple fact of British leadership in pushing EU emissions trading confounds the unambitious content of potential policies. Like Germany, the UK was forced into the EU ETS, and shut down its domestic ETS in 2006 after the EU-level policy superseded it. By 2009, the EU ETS covered almost half of British carbon emissions (Bowen and Rydge, 2011). At the same time, partly because of British policy learning during its domestic emissions trading experiment, the UK was a

European leader in managing EU ETS implementation (Skjærseth and Wettestad, 2008).

Bibliography

Bowen, A. and Rydge, J. (2011). Climate-Change Policy in the United Kingdom. *OECD Economics Department Working Papers*, 886:2–42.

Carter, N. (2008). Combating Climate Change in the UK: Challenges and Obstacles. *The Political Quarterly*, 79(2):194–205.

Carter, N. (2014). The politics of climate change in the UK. *WIREs Climate Change*, 5(3):423–433.

Darkin, B. (2006). Pledges, politics and performance: an assessment of UK climate policy. *Climate Policy*, 6(3):257–274.

Jordan, A., Wurzel, R. K., Zito, A. R., and Brückner, L. (2003). Policy Innovation or 'Muddling Through'? 'New' Environmental Policy Instruments in the United Kingdom. *Environmental Politics*, 12(1):179–200.

Lockwood, M. (2013). The political sustainability of climate policy: The case of the UK Climate Change Act. *Global Environmental Change*, 23(5):1339–1348.

McLean, I. (2008). Climate Change and UK Politics: From Brynle Williams to Sir Nicholas Stern. *The Political Quarterly*, 79(2):184–193.

Oshitani, S. (2006). *Global Warming Policy in Japan and Britain: Interactions Between Institutions and Issue Characteristics*. Manchester University Press, Manchester, UK.

Principality of Liechtenstein (2017). *Liechtenstein's Greenhouse Gas Inventory: 1990-2015*. Office of Environment, Principality of Liechtenstein, Vaduz, Liechtenstein.

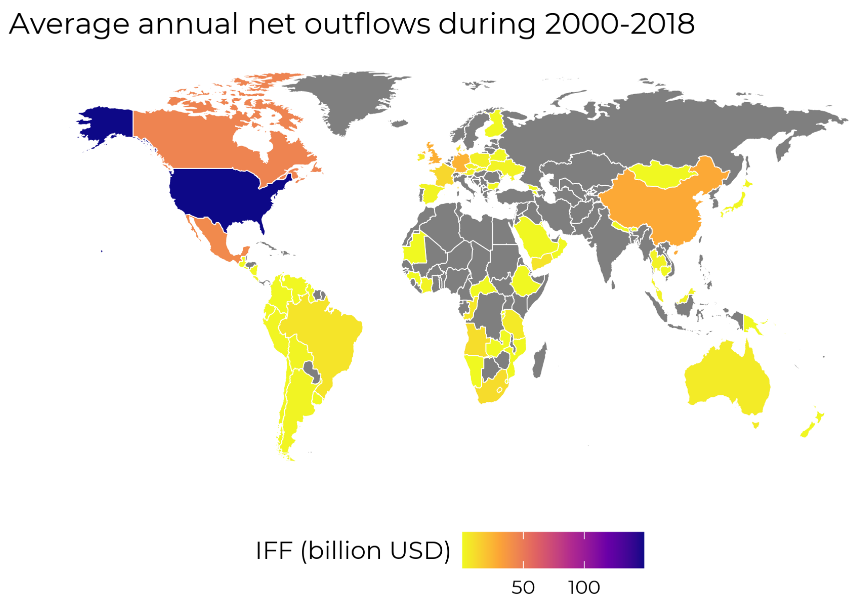
Skjærseth, J. B. and Wettestad, J. (2008). *EU Emissions Trading - Initiation, Decision-making and Implementation*. Routledge, London, UK.

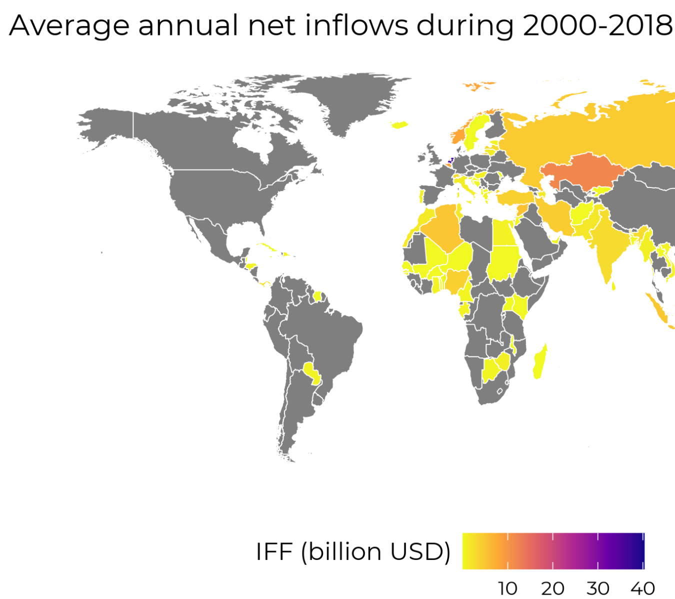
Thatcher, M. (2003). *Statecraft: Strategies for a Changing World*. Harper Collins, London, UK.

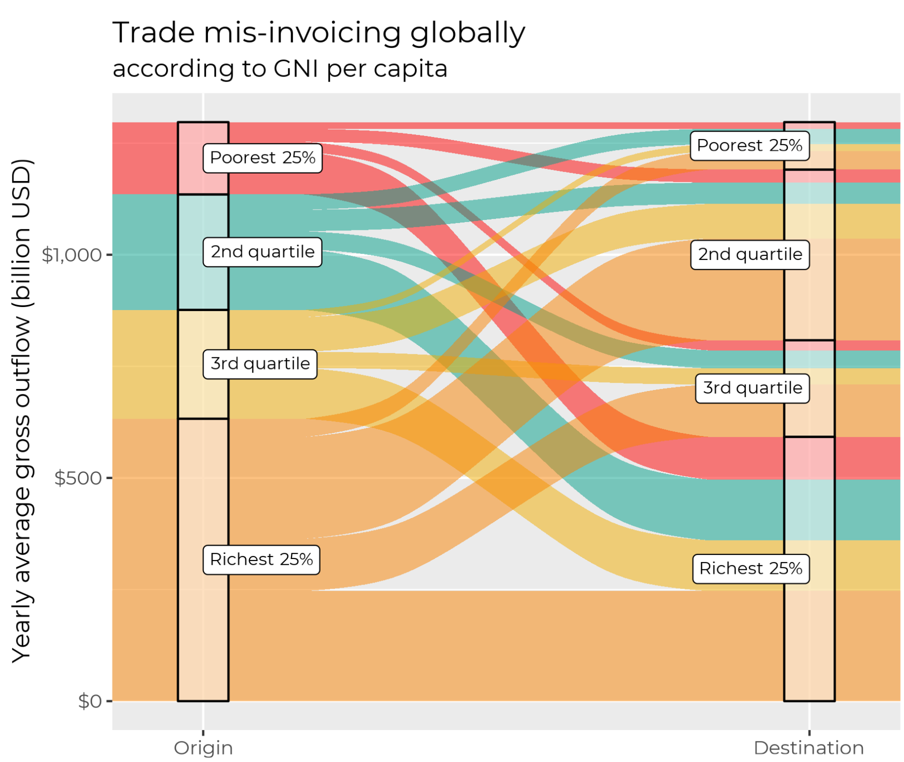
Appendix B

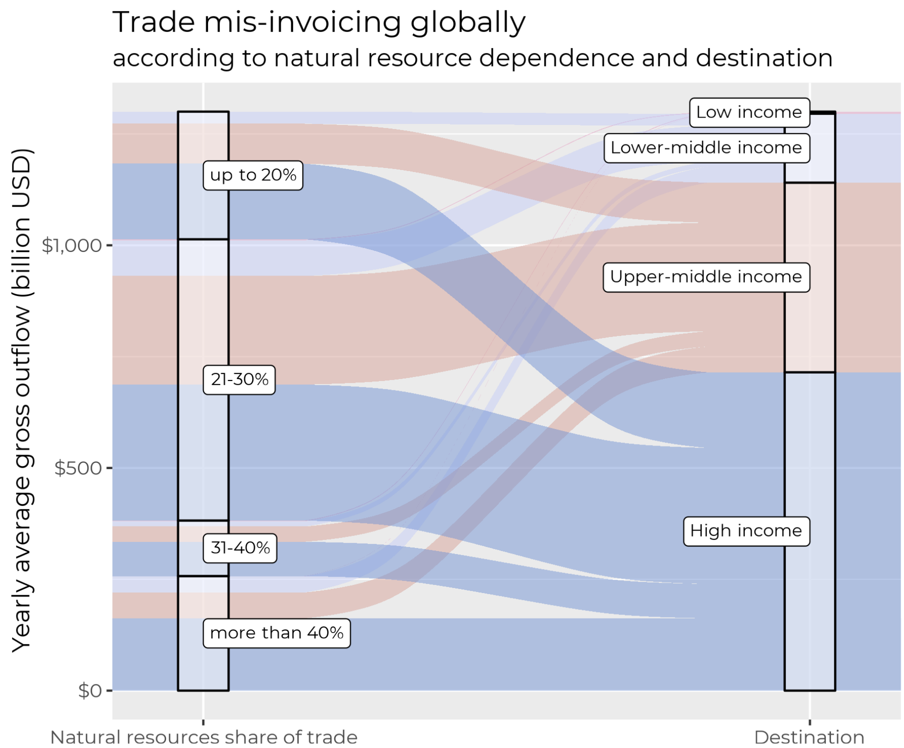
Appendix for Chapter 3

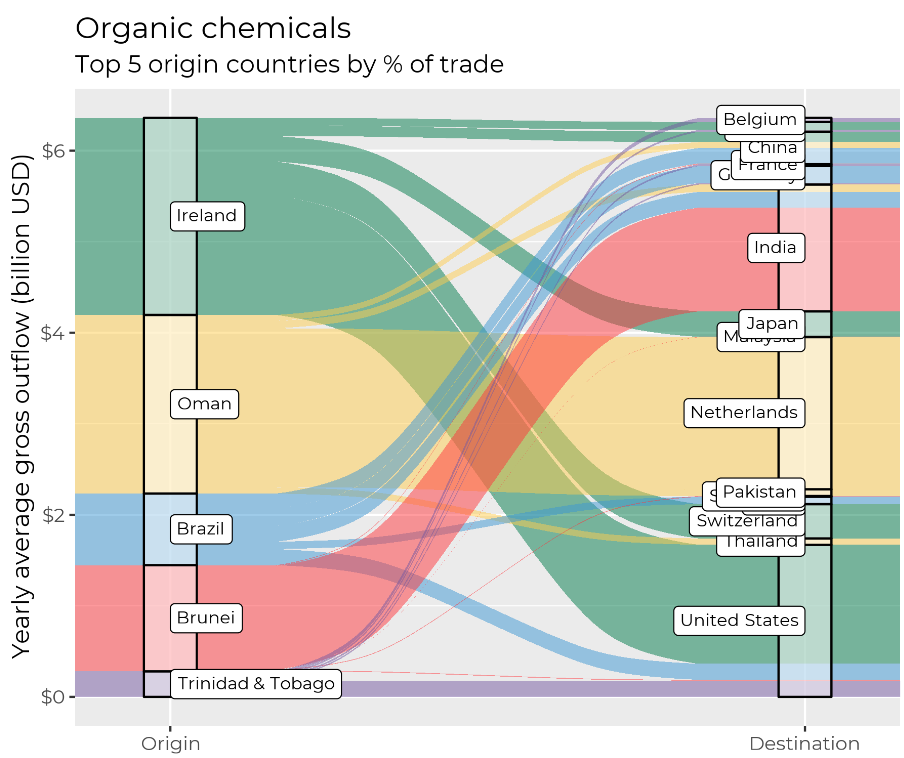
B.1 Selected figures from the "atlas"




Average annual net inflows during 2000-2018
IFF (billion USD)
10
20
30
40

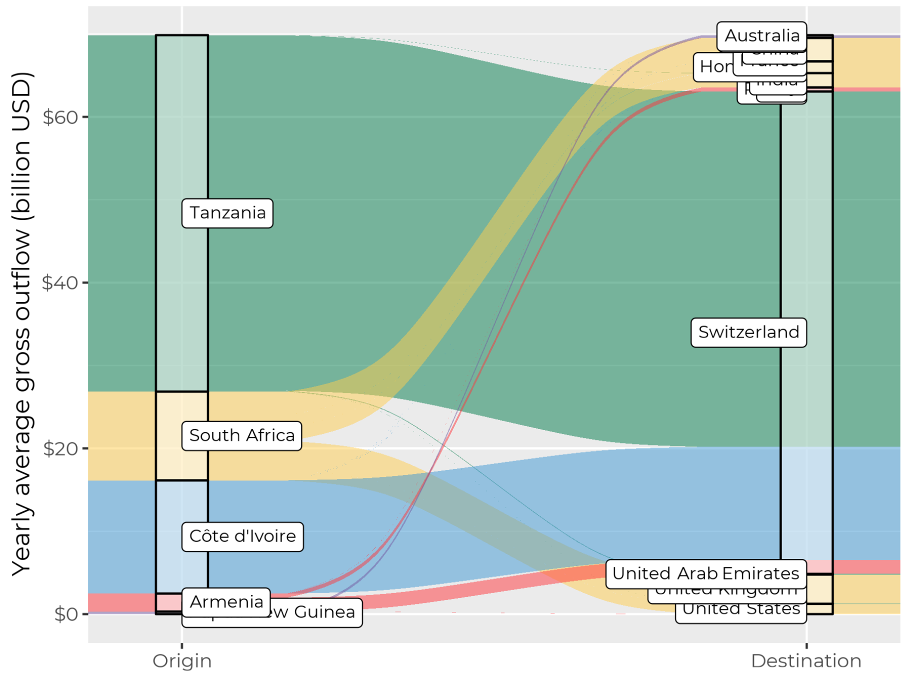

Trade mis-invoicing globally
according to GNI per capita
Yearly average gross outflow (billion USD)
$1,000
$500
$0
Poorest 25%
2nd quartile
3rd quartile
Richest 25%
Poorest 25%
2nd quartile
3rd quartile
Richest 25%
Origin
Destination


Trade mis-invoicing globally
according to natural resource dependence and destination
Yearly average gross outflow (billion USD)
$1,000
$500
$0
up to 20%
21-30%
31-40%
more than 40%
Low income
Lower-middle income
Upper-middle income
High income
Natural resources share of trade
Destination


Organic chemicals
Top 5 origin countries by % of trade
Yearly average gross outflow (billion USD)
$6
$4
$2
$0
Ireland
Oman
Brazil
Brunei
Trinidad & Tobago
Belgium
China
India
Japan
Netherlands
Pakistan
Switzerland
United States
Origin
Destination

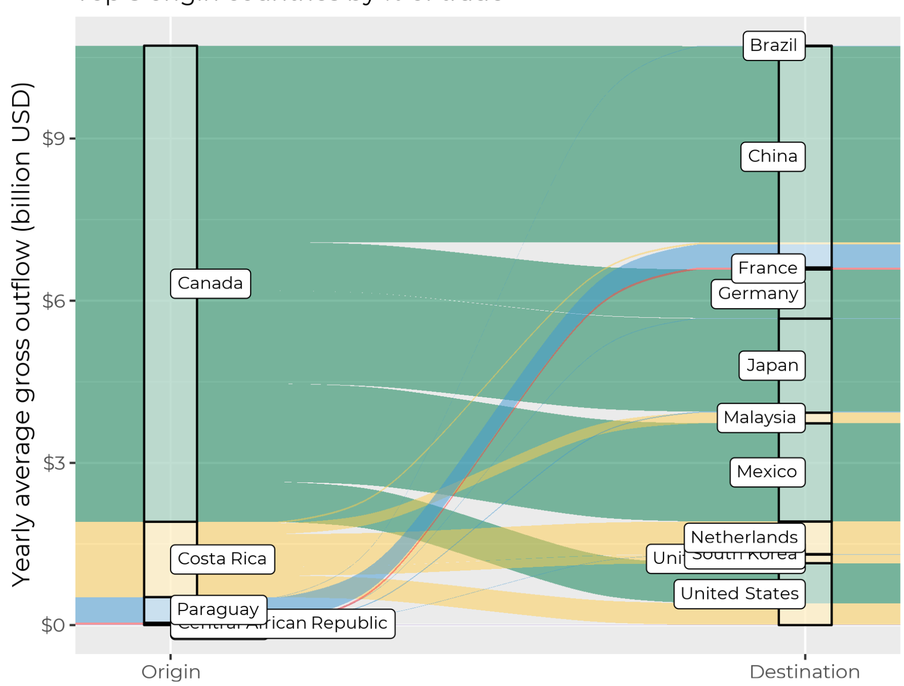
Pearls, precious stones and metals, jewelry, and coins

Top 5 origin countries by % of trade



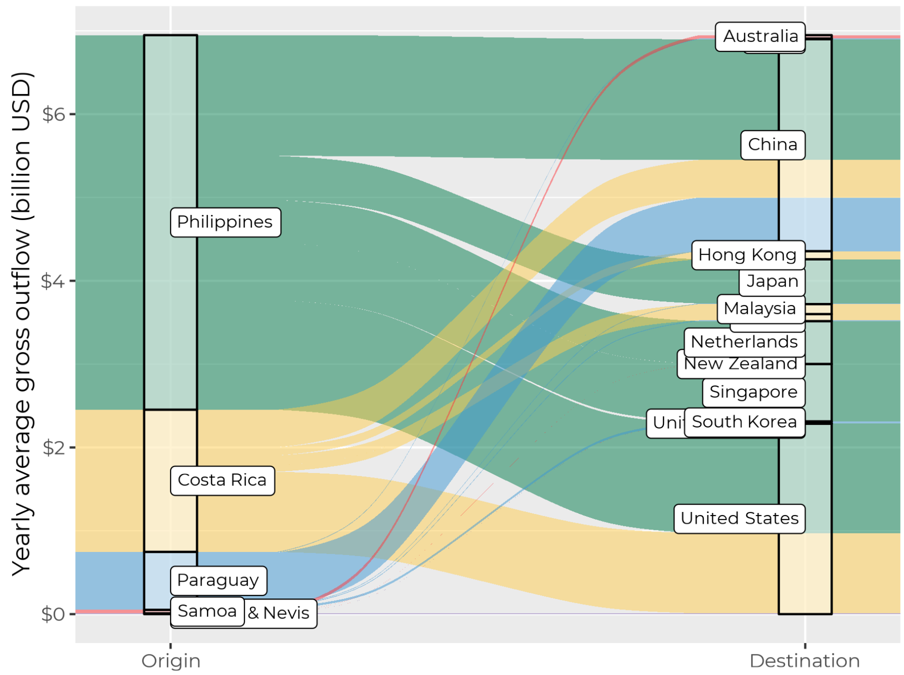
Nuclear reactors, boilers, and machinery

Top 5 origin countries by % of trade



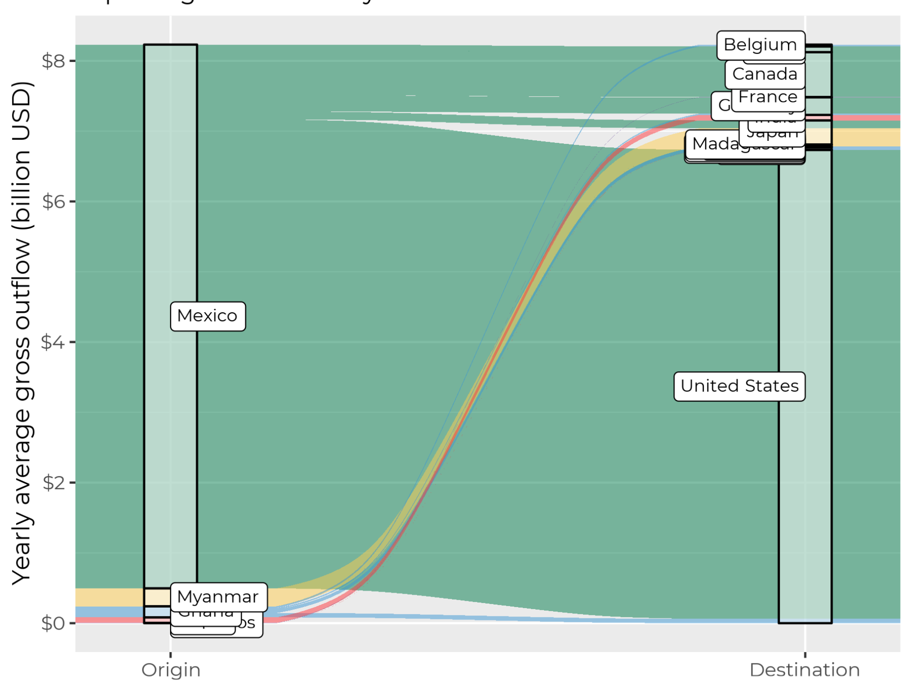
Electrical equipment, sound recorders, and televisions

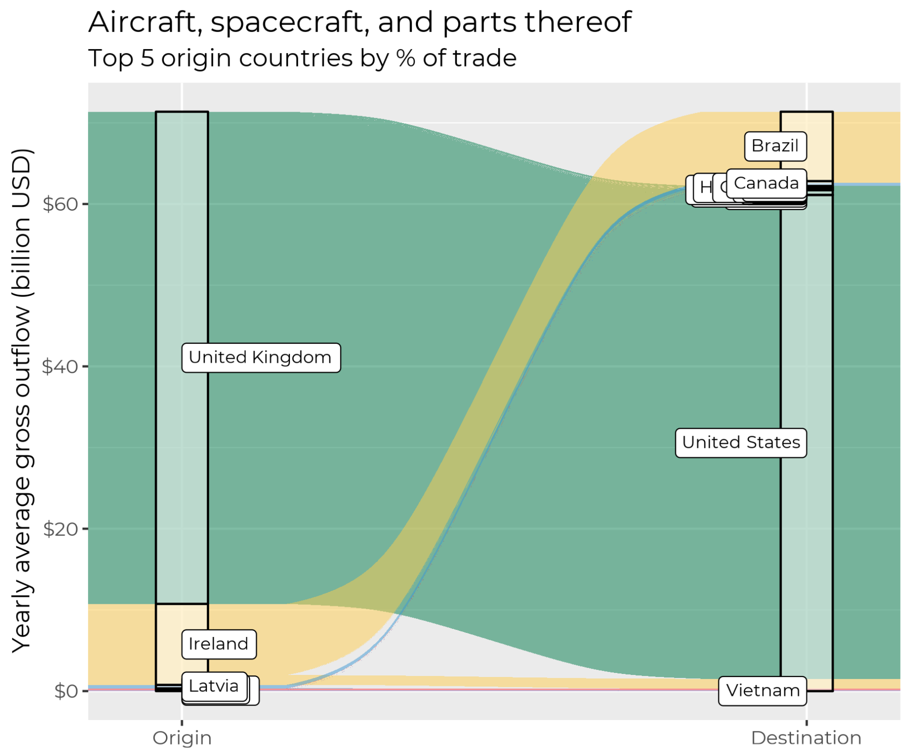
Top 5 origin countries by % of trade

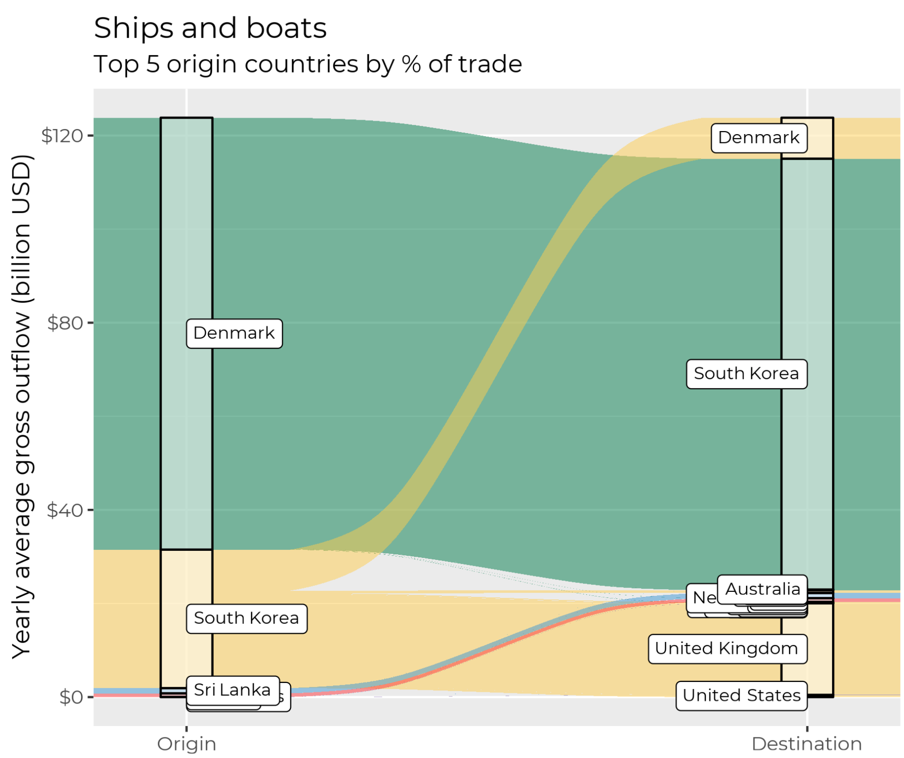


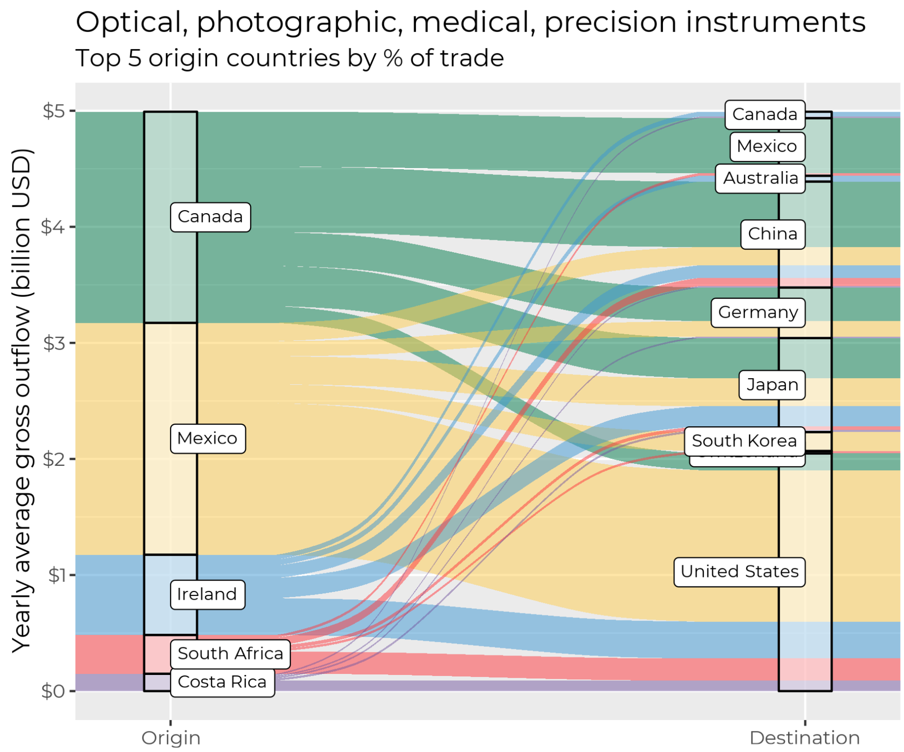
Vehicles other than railway or tramways

Top 5 origin countries by % of trade




Aircraft, spacecraft, and parts thereof
Top 5 origin countries by % of trade
Yearly average gross outflow (billion USD)
$0
$20
$40
$60
United Kingdom
Ireland
Latvia
Brazil
Canada
United States
Vietnam
Origin
Destination


Ships and boats
Top 5 origin countries by % of trade
Yearly average gross outflow (billion USD)
$0
$40
$80
$120
Denmark
South Korea
Sri Lanka
Denmark
South Korea
Australia
United Kingdom
United States
Origin
Destination



B.2 Full country results

Country	ISO	Region	Income group	Million USD	% of GDP	% of trade
Afghanistan	AFG	Asia	LIC	$6	0.03%	0.07%
Albania	ALB	Europe	UMC	$174	1.71%	3.39%
Algeria	DZA	Africa	LMC	$4,225	3.25%	5.46%
Angola	AGO	Africa	LMC	$11,009	14.51%	16.27%
Antigua & Barbuda	ATG	Americas	HIC	$0	0.02%	0.04%
Argentina	ARG	Americas	UMC	$5,689	1.50%	5.24%
Armenia	ARM	Asia	UMC	$652	6.20%	11.71%
Aruba	ABW	Americas	HIC	$32	1.24%	2.37%
Australia	AUS	Oceania	HIC	$16,819	1.68%	5.01%
Austria	AUT	Europe	HIC	$5,104	1.39%	1.83%
Azerbaijan	AZE	Asia	UMC	$3,149	7.09%	15.54%
Bahamas	BHS	Americas	HIC	$147	1.44%	3.93%
Bahrain	BHR	Asia	HIC	$611	2.59%	2.30%

Country	ISO	Region	Income group	Million USD	% of GDP	% of trade
Bangladesh	BGD	Asia	LMC	$1,949	1.97%	5.15%
Barbados	BRB	Americas	HIC	$82	2.10%	4.67%
Belarus	BLR	Europe	UMC	$2,465	4.31%	3.86%
Belgium	BEL	Europe	HIC	$9,341	2.06%	1.22%
Belize	BLZ	Americas	UMC	$34	2.45%	3.13%
Benin	BEN	Africa	LMC	$53	0.70%	2.97%
Bermuda	BMU	Americas	HIC	$30	0.45%	2.99%
Bhutan	BTN	Asia	LMC	$14	1.22%	1.25%
Bolivia	BOL	Americas	LMC	$2,379	9.90%	15.24%
Bosnia & Herzegovina	BIH	Europe	UMC	$535	3.08%	3.51%
Botswana	BWA	Africa	UMC	$483	3.39%	4.00%
Brazil	BRA	Americas	UMC	$23,646	1.59%	7.98%
Brunei	BRN	Asia	HIC	$1,764	11.09%	13.54%
Bulgaria	BGR	Europe	UMC	$1,422	3.13%	3.25%
Burkina Faso	BFA	Africa	LIC	$81	0.70%	1.83%
Burundi	BDI	Africa	LIC	$8	0.39%	1.04%
Cambodia	KHM	Asia	LMC	$2,256	20.54%	22.54%
Cameroon	CMR	Africa	LMC	$380	1.67%	5.42%
Canada	CAN	Americas	HIC	$65,240	4.73%	8.38%
Cape Verde	CPV	Africa	LMC	$0	0.02%	0.03%
Central African Republic	CAF	Africa	LIC	$14	0.83%	4.52%
Chile	CHL	Americas	HIC	$5,161	2.90%	4.90%
China	CHN	Asia	UMC	$59,179	1.18%	2.49%
Colombia	COL	Americas	UMC	$3,829	1.55%	5.26%
Comoros	COM	Africa	LMC	$3	0.35%	2.05%
Congo	COG	Africa	LMC	$8,040	54.23%	54.65%
Costa Rica	CRI	Americas	UMC	$3,380	11.12%	15.96%
Côte d'Ivoire	CIV	Africa	LMC	$3,846	14.27%	20.13%
Croatia	HRV	Europe	HIC	$573	1.09%	1.79%
Cuba	CUB	Americas	UMC	$133	0.37%	1.83%
Cyprus	CYP	Asia	HIC	$150	0.69%	1.25%
Czech Republic	CZE	Europe	HIC	$5,172	2.71%	2.12%
Denmark	DNK	Europe	HIC	$12,314	3.70%	6.17%

Country	ISO	Region	Income group	Million USD	% of GDP	% of trade
Dominica	DMA	Americas	UMC	$4	0.87%	1.64%
Dominican Republic	DOM	Americas	UMC	$1,730	3.56%	9.02%
Ecuador	ECU	Americas	UMC	$2,556	3.85%	8.53%
Egypt	EGY	Africa	LMC	$6,756	3.73%	11.21%
El Salvador	SLV	Americas	LMC	$974	5.78%	8.26%
Estonia	EST	Europe	HIC	$626	3.22%	2.38%
Eswatini	SWZ	Africa	LMC	$1	0.03%	0.03%
Ethiopia	ETH	Africa	LIC	$683	1.94%	5.62%
Fiji	FJI	Oceania	UMC	$67	1.82%	2.18%
Finland	FIN	Europe	HIC	$3,884	1.64%	2.90%
France	FRA	Europe	HIC	$18,752	0.76%	1.77%
Gabon	GAB	Africa	UMC	$258	3.22%	5.65%
Gambia	GMB	Africa	LIC	$8	0.58%	1.64%
Georgia	GEO	Asia	UMC	$383	3.72%	6.98%
Germany	DEU	Europe	HIC	$33,306	1.00%	1.52%
Ghana	GHA	Africa	LMC	$2,342	8.29%	15.10%
Greece	GRC	Europe	HIC	$1,653	0.67%	2.00%
Grenada	GRD	Americas	UMC	$0	0.07%	0.13%
Grenadines	VCT	Americas	UMC	$4	0.64%	1.18%
Guatemala	GTM	Americas	UMC	$1,574	4.46%	9.10%
Guinea	GIN	Africa	LIC	$66	1.41%	2.62%
Guinea-Bissau	GNB	Africa	LIC	$0	0.01%	0.07%
Guyana	GUY	Americas	UMC	$86	3.22%	3.54%
Honduras	HND	Americas	LMC	$614	4.65%	7.78%
Hong Kong	HKG	Asia	HIC	$35,582	15.44%	3.52%
Hungary	HUN	Europe	HIC	$2,680	2.10%	1.53%
Iceland	ISL	Europe	HIC	$690	4.00%	7.28%
India	IND	Asia	LMC	$13,544	0.96%	3.07%
Indonesia	IDN	Asia	UMC	$9,339	1.69%	3.91%
Iran	IRN	Asia	UMC	$4,166	1.36%	3.64%
Ireland	IRL	Europe	HIC	$10,842	4.73%	5.64%
Israel	ISR	Asia	HIC	$4,081	1.89%	3.77%
Italy	ITA	Europe	HIC	$3,259	0.17%	0.40%

Country	ISO	Region	Income group	Million USD	% of GDP	% of trade
Jamaica	JAM	Americas	UMC	$333	2.71%	4.96%
Japan	JPN	Asia	HIC	$20,064	0.40%	1.57%
Jordan	JOR	Asia	UMC	$554	2.39%	2.51%
Kazakhstan	KAZ	Asia	UMC	$8,778	5.44%	9.84%
Kenya	KEN	Africa	LMC	$507	1.89%	4.62%
Kuwait	KWT	Asia	HIC	$1,163	0.95%	1.26%
Kyrgyz Republic	KGZ	Asia	LMC	$418	7.22%	7.25%
Lao PDR	LAO	Asia	LMC	$1,118	7.55%	14.65%
Latvia	LVA	Europe	HIC	$557	1.98%	2.22%
Lebanon	LBN	Asia	UMC	$419	1.57%	3.23%
Lesotho	LSO	Africa	LMC	$0	0.01%	0.01%
Lithuania	LTU	Europe	HIC	$1,620	3.96%	3.26%
Macao	MAC	Asia	HIC	$320	2.08%	3.60%
Macedonia	MKD	Europe	UMC	$399	4.39%	4.47%
Madagascar	MDG	Africa	LIC	$308	3.72%	9.15%
Malawi	MWI	Africa	LIC	$50	0.90%	1.71%
Malaysia	MYS	Asia	UMC	$14,984	7.57%	4.86%
Maldives	MDV	Asia	UMC	$9	0.49%	0.83%
Mali	MLI	Africa	LIC	$210	2.33%	4.55%
Malta	MLT	Europe	HIC	$434	4.50%	4.28%
Mauritania	MRT	Africa	LMC	$47	0.80%	1.18%
Mauritius	MUS	Africa	HIC	$261	2.86%	4.07%
Mexico	MEX	Americas	UMC	$57,367	5.48%	9.72%
Moldova	MDA	Europe	LMC	$193	3.52%	3.44%
Mongolia	MNG	Asia	LMC	$454	7.66%	8.47%
Morocco	MAR	Africa	LMC	$2,411	2.84%	4.98%
Mozambique	MOZ	Africa	LIC	$635	5.42%	9.47%
Myanmar	MMR	Asia	LMC	$1,574	2.53%	7.13%
Namibia	NAM	Africa	UMC	$662	5.71%	5.00%
Nepal	NPL	Asia	LMC	$538	3.26%	8.64%
Netherlands	NLD	Europe	HIC	$10,133	1.30%	1.20%
New Zealand	NZL	Oceania	HIC	$2,619	1.94%	4.44%
Nicaragua	NIC	Americas	LMC	$1,276	10.42%	12.40%

Country	ISO	Region	Income group	Million USD	% of GDP	% of trade
Niger	NER	Africa	LIC	$58	0.78%	2.59%
Nigeria	NGA	Africa	LMC	$8,712	2.80%	8.98%
Norway	NOR	Europe	HIC	$10,729	2.80%	5.65%
Oman	OMN	Asia	HIC	$4,062	7.82%	7.85%
Pakistan	PAK	Asia	LMC	$2,440	1.25%	4.08%
Palau	PLW	Oceania	HIC	$1	0.42%	0.76%
Panama	PAN	Americas	HIC	$636	2.61%	3.52%
Papua New Guinea	PNG	Oceania	LMC	$276	7.58%	7.30%
Paraguay	PRY	Americas	UMC	$1,240	5.20%	9.10%
Peru	PER	Americas	UMC	$4,588	4.11%	9.92%
Philippines	PHL	Asia	LMC	$8,939	5.48%	8.47%
Poland	POL	Europe	HIC	$6,291	1.44%	2.07%
Portugal	PRT	Europe	HIC	$1,256	0.59%	1.03%
Qatar	QAT	Asia	HIC	$3,608	4.03%	5.45%
Russia	RUS	Europe	UMC	$41,694	3.00%	7.44%
Rwanda	RWA	Africa	LIC	$53	0.63%	1.89%
Samoa	WSM	Oceania	UMC	$16	2.15%	3.65%
São Tomé and Príncipe	STP	Africa	LMC	$0	0.13%	0.29%
Saudi Arabia	SAU	Asia	HIC	$3,678	0.79%	1.16%
Senegal	SEN	Africa	LMC	$263	1.68%	3.69%
Seychelles	SYC	Africa	HIC	$29	2.11%	1.51%
Singapore	SGP	Asia	HIC	$27,794	13.26%	4.85%
Slovak Republic	SVK	Europe	HIC	$3,127	3.47%	2.35%
Slovenia	SVN	Europe	HIC	$708	1.65%	1.51%
Solomon Islands	SLB	Oceania	LMC	$19	1.82%	2.59%
South Africa	ZAF	Africa	UMC	$23,565	8.30%	17.90%
South Korea	KOR	Asia	HIC	$13,473	1.20%	1.84%
Spain	ESP	Europe	HIC	$6,438	0.52%	1.19%
Sri Lanka	LKA	Asia	LMC	$1,108	2.76%	5.45%
St. Kitts & Nevis	KNA	Americas	HIC	$17	2.55%	5.85%
St. Lucia	LCA	Americas	UMC	$22	1.63%	3.24%
Sudan	SDN	Africa	LIC	$444	1.06%	2.99%
Suriname	SUR	Americas	UMC	$59	1.60%	1.82%

Country	ISO	Region	Income group	Million USD	% of GDP	% of trade
Sweden	SWE	Europe	HIC	$3,932	0.84%	1.41%
Switzerland	CHE	Europe	HIC	$28,904	4.87%	6.36%
Syria	SYR	Asia	LIC	$990	1.00%	3.75%
Tanzania	TZA	Africa	LMC	$6,700	17.24%	40.51%
Thailand	THA	Asia	UMC	$11,964	4.13%	3.70%
Togo	TGO	Africa	LIC	$78	2.36%	4.44%
Tonga	TON	Oceania	UMC	$1	0.15%	0.31%
Trinidad & Tobago	TTO	Americas	HIC	$4,277	18.59%	19.51%
Tunisia	TUN	Africa	LMC	$1,354	3.43%	4.20%
Turkey	TUR	Asia	UMC	$6,898	1.12%	2.83%
Uganda	UGA	Africa	LIC	$126	0.77%	2.45%
Ukraine	UKR	Europe	LMC	$11,316	8.74%	10.64%
United Arab Emirates	ARE	Asia	HIC	$20,708	5.90%	3.50%
United Kingdom	GBR	Europe	HIC	$45,373	1.90%	4.73%
United States	USA	Americas	HIC	$220,848	1.49%	6.92%
Uruguay	URY	Americas	HIC	$955	2.56%	6.86%
Vanuatu	VUT	Oceania	LMC	$1	0.10%	0.22%
Venezuela	VEN	Americas	UMC	$4,610	2.84%	6.80%
Vietnam	VNM	Asia	LMC	$9,053	7.69%	5.61%
Yemen	YEM	Asia	LIC	$8,623	36.55%	57.51%
Zambia	ZMB	Africa	LMC	$497	2.66%	4.28%
Zimbabwe	ZWE	Africa	LMC	$694	5.44%	8.06%

Table B.1: Average gross annual outflows during 2000-2018.

Appendix C

Appendix for Chapter 4

C.1 Codebook

C.1.1 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

C.1.2 Predictors

Gravity variables

Code name	Description	Data source	Type	Unit of observation
`GDP`	Gross domestic product (thousands, current US$)	*Gravity Database*, CEPII	Continuous	it, jt
`pop`	Population (thousands)	*Gravity Database*, CEPII	Continuous	it, jt
`dist`	Distance between most populated city of each country (km)	*Gravity Database*, CEPII	Continuous	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	Continuous	*it*, *jt*
`RegQual`	Regulatory quality, percentile rank	*Worldwide Governance Indicators*, Kaufmann and Kraay	Continuous	*it*, *jt*
`RuleLaw`	Rule of law, percentile rank	*Worldwide Governance Indicators*, Kaufmann and Kraay	Continuous	*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

C.2 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*.[1] 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: 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 C.1 below.

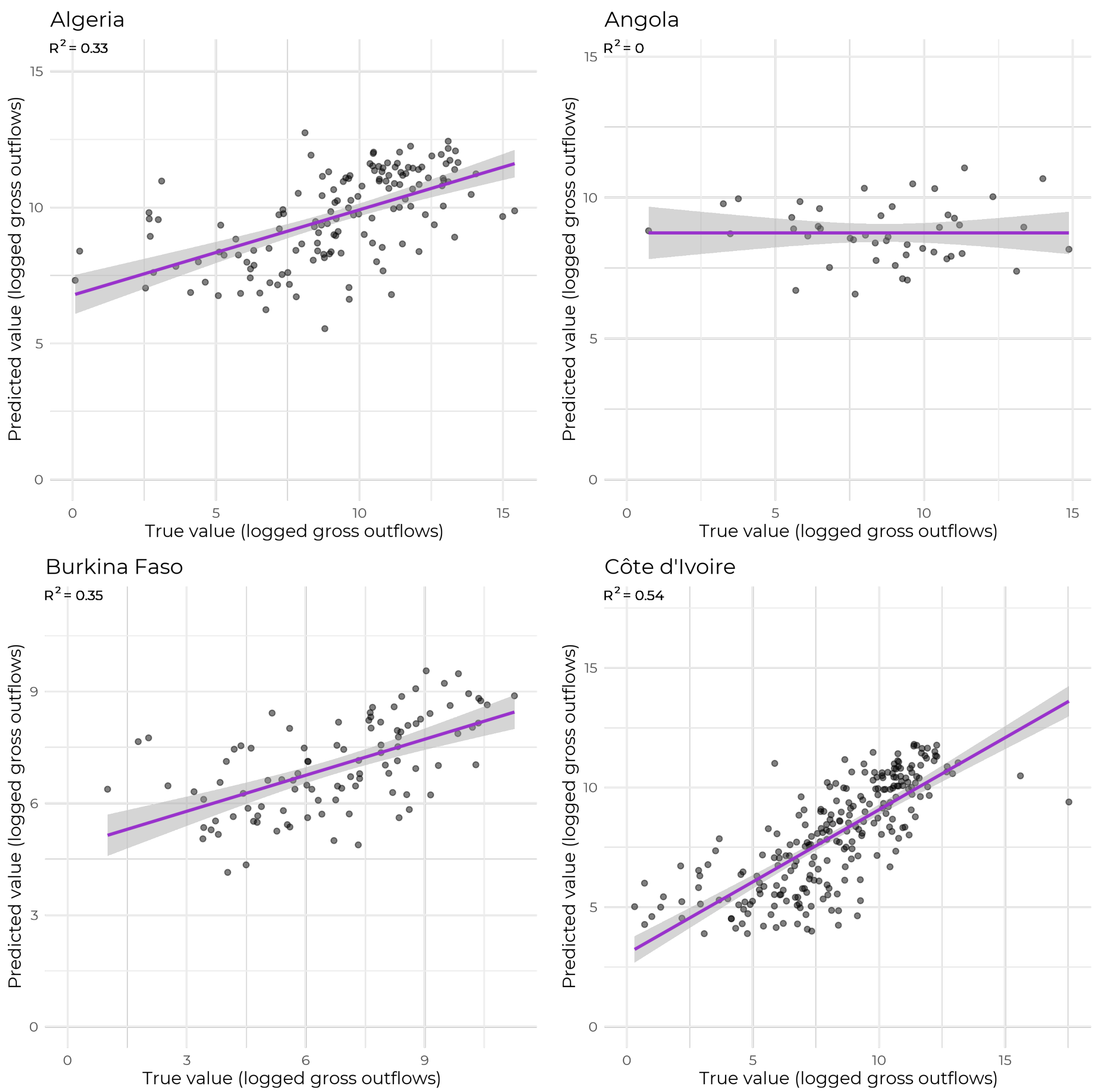
Hyperparameter	**Tuning**
Number of trees	1278
Maximum depth of individual trees	195
Minimum number of observations to split on at an internal node	12
Minimum number of observations in a leaf (terminal node)	1
Maximum number of random features to consider at each split	All features
Use bootstrapped samples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

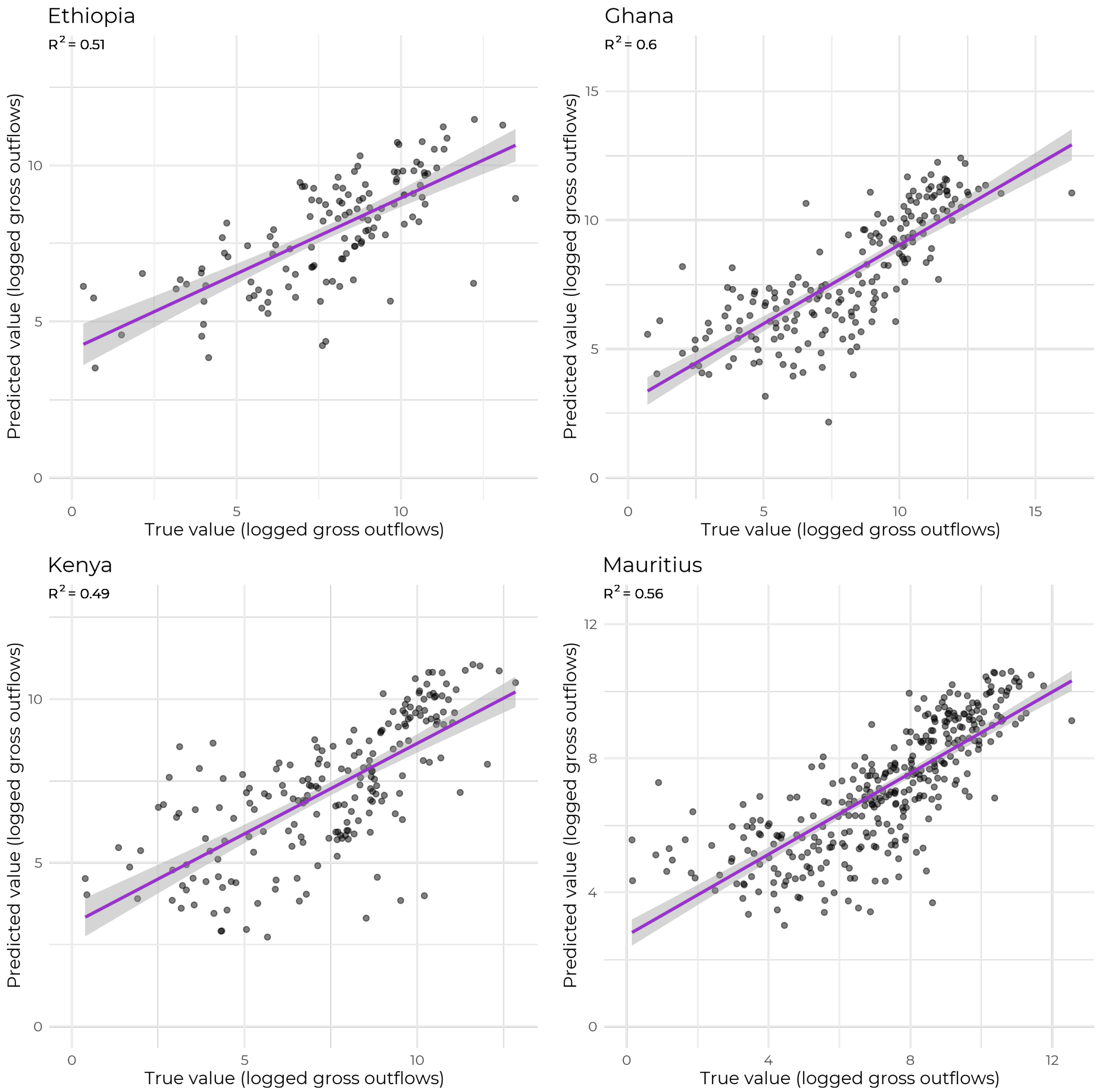
Table C.1: Tuned hyperparameters for the Random Forest estimator following randomized search strategy with 5-fold cross-validation. The search resulted in the same configuration of hyperparameters for both outflows and inflows.

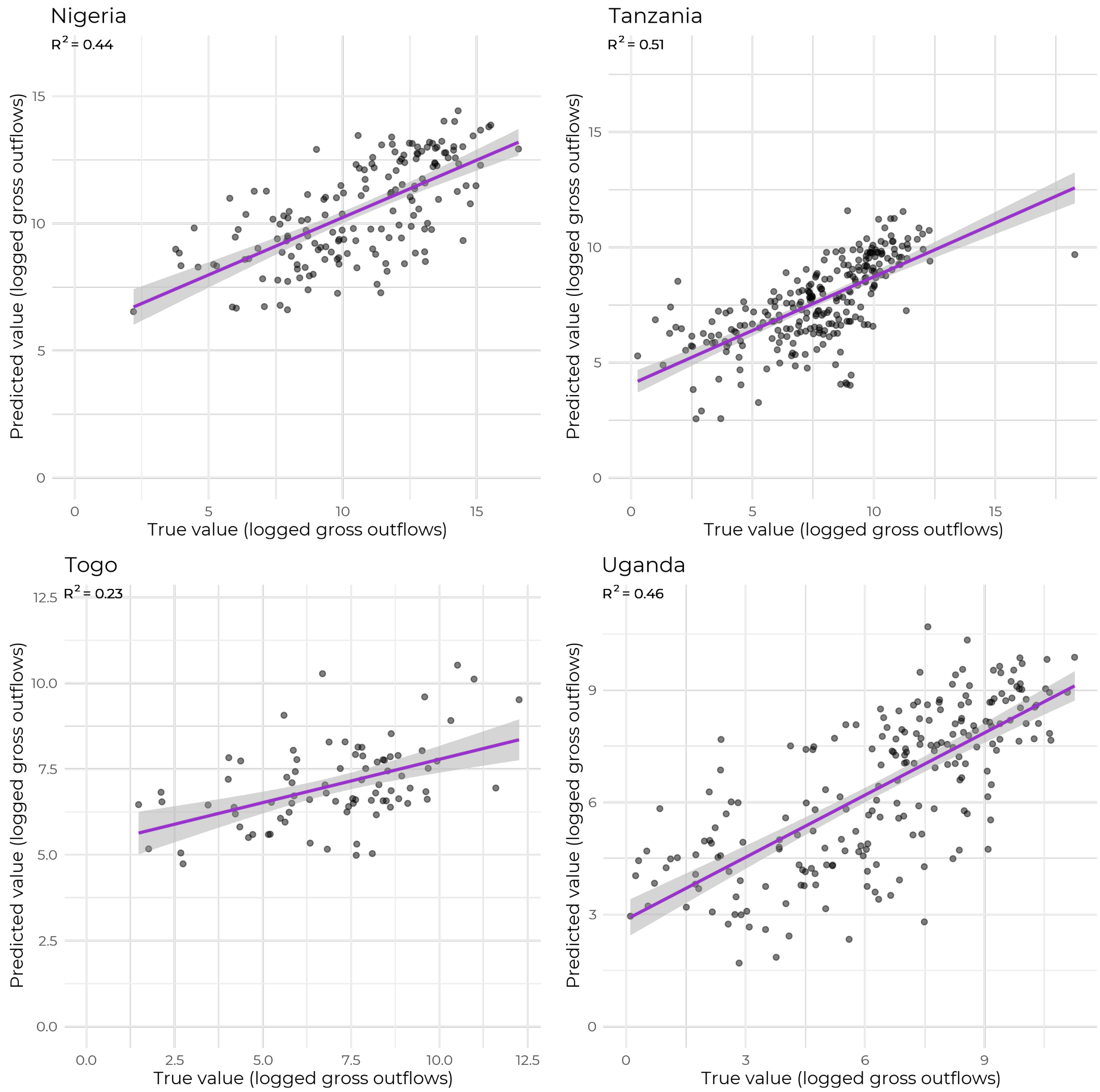
[1] Implemented using the *RandomizedSearchCV* procedure of *scikit-learn* (random seed 1509).

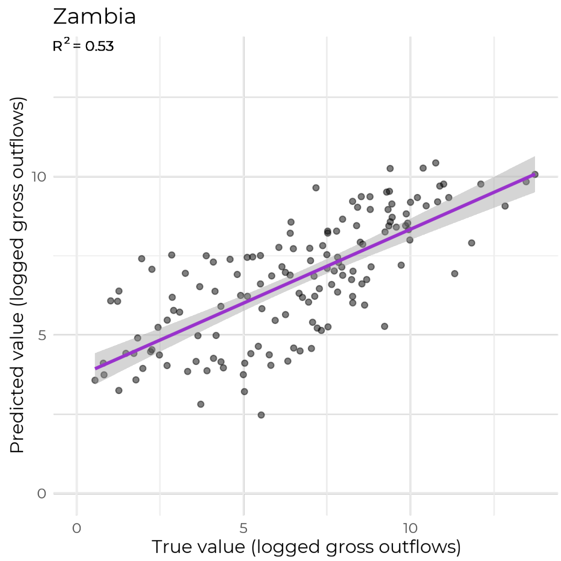
C.3 Cross-validated predictions for all African countries

C.3.1 Gross outflows



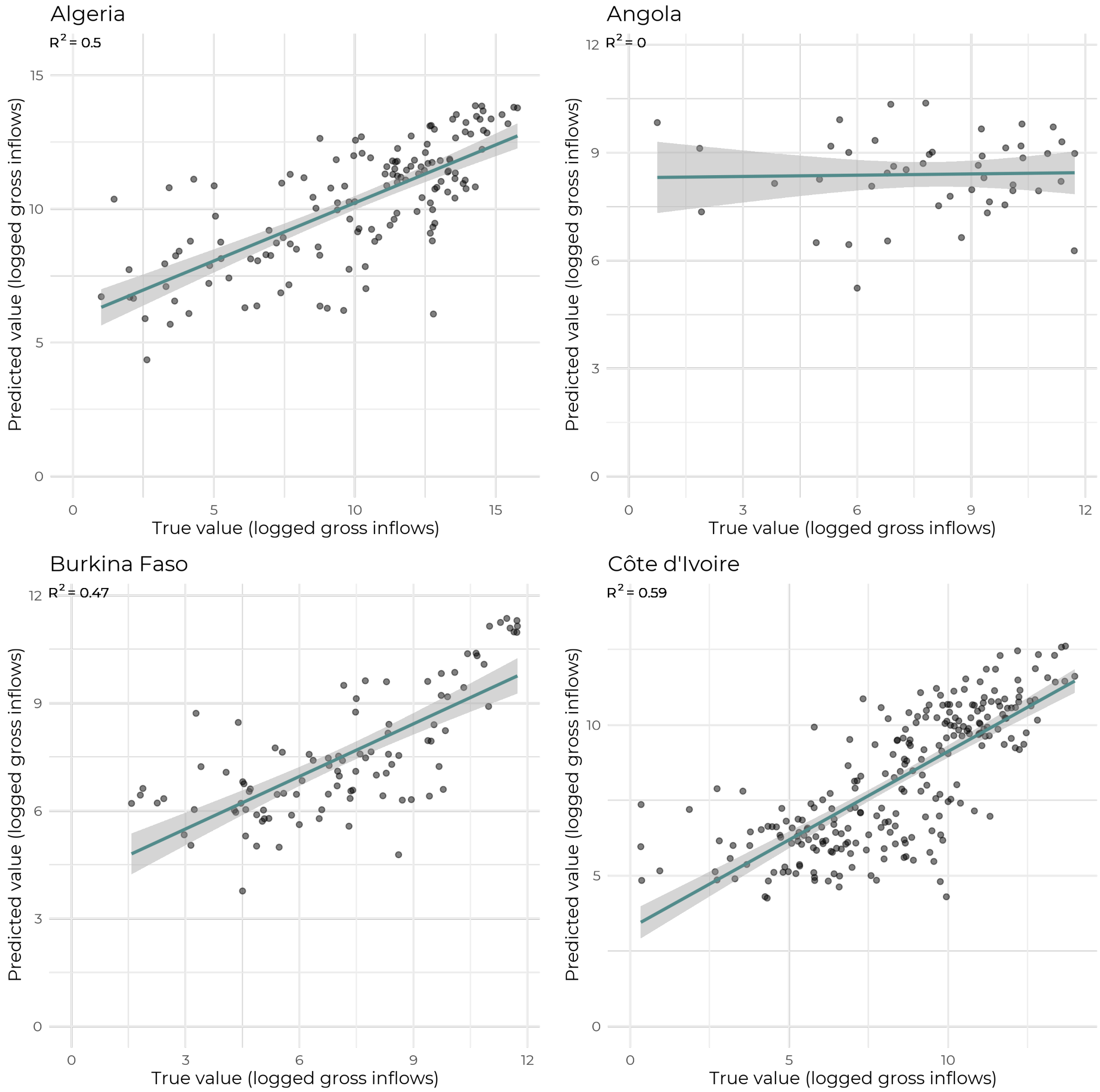

Ethiopia
R^2 = 0.51
Predicted value (logged gross outflows)
True value (logged gross outflows)
0
5
10
Ghana
R^2 = 0.6
15
Kenya
R^2 = 0.49
Mauritius
R^2 = 0.56
4
8
12

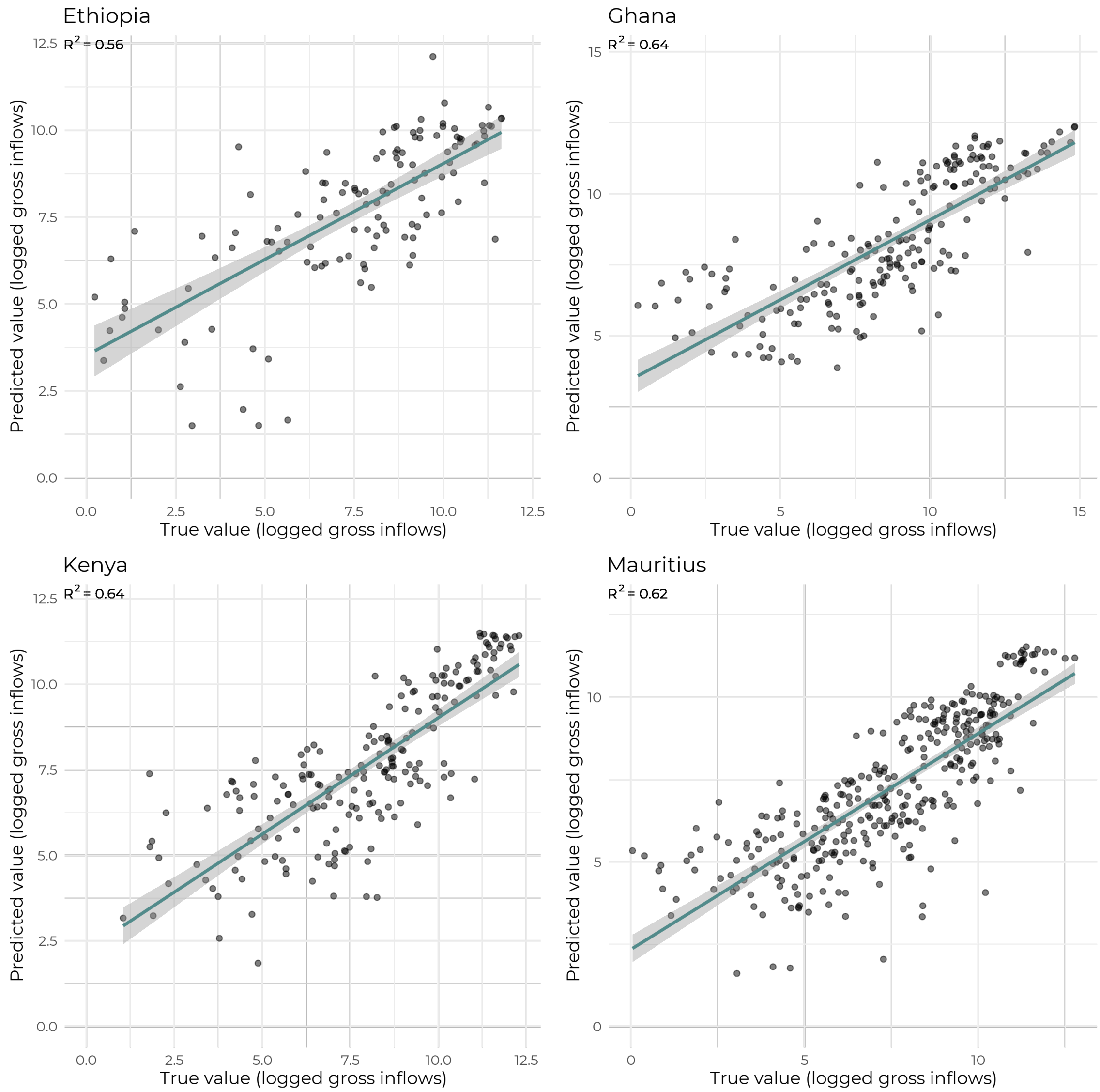

Nigeria
R² = 0.44
Predicted value (logged gross outflows)
True value (logged gross outflows)
Tanzania
R² = 0.51
Predicted value (logged gross outflows)
True value (logged gross outflows)
Togo
R² = 0.23
Predicted value (logged gross outflows)
True value (logged gross outflows)
Uganda
R² = 0.46
Predicted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

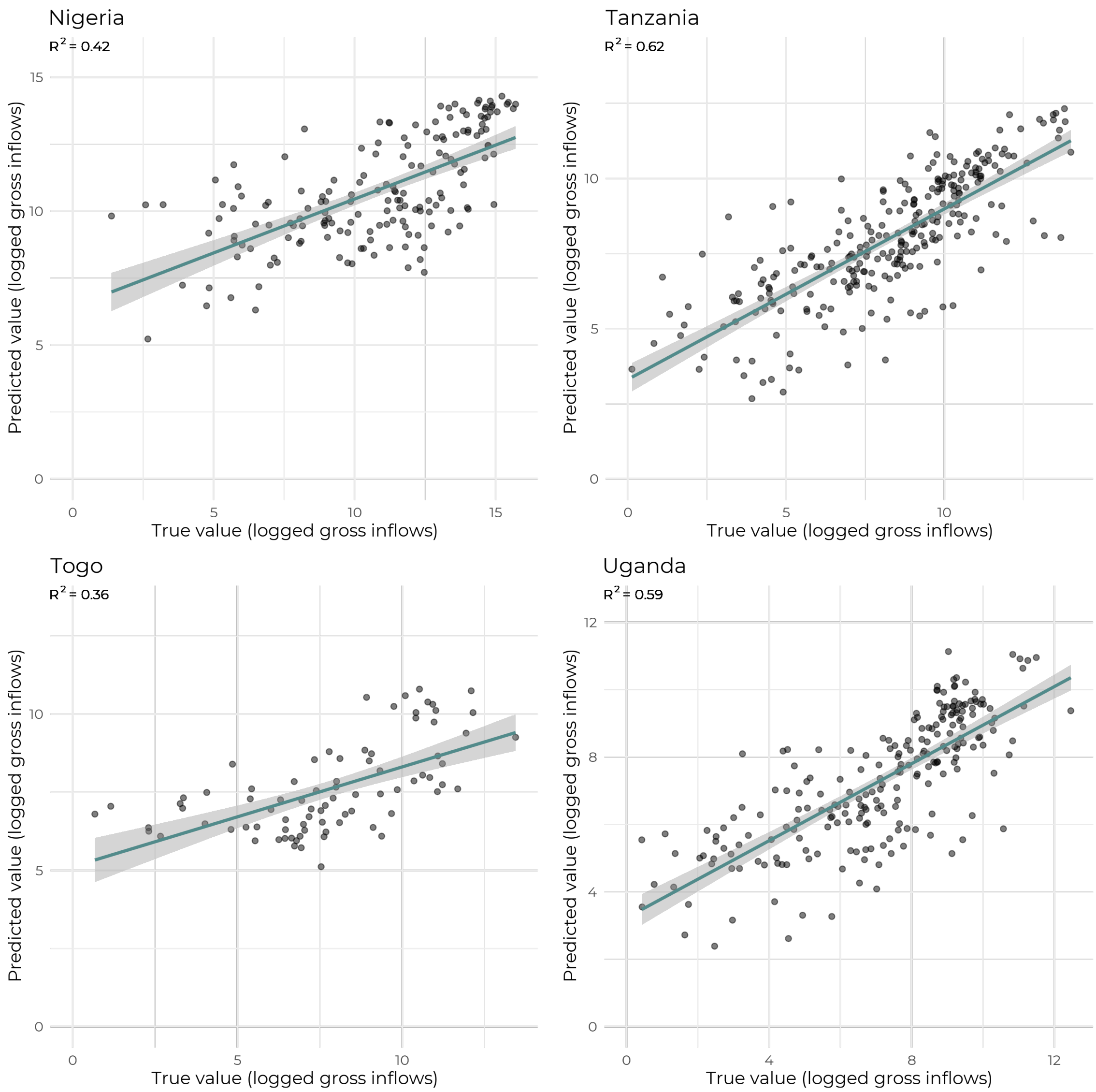


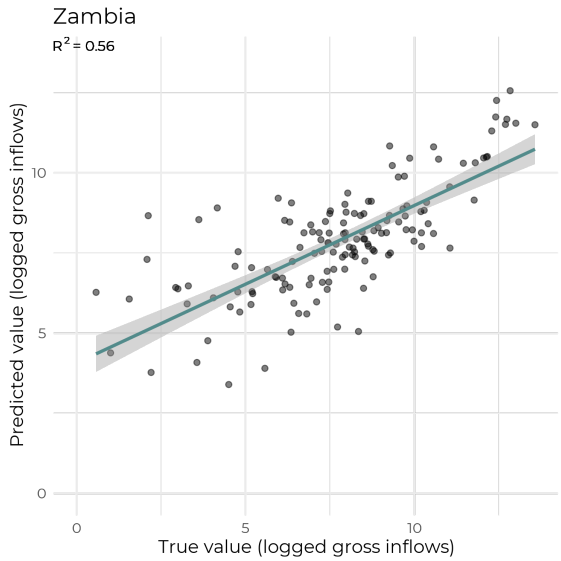
Zambia
R^2 = 0.53
Predicted value (logged gross outflows)
True value (logged gross outflows)
0
5
10
0
5
10

C.3.2 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)
0
5
10
15
4
8
12


Zambia
R^2 = 0.56
Predicted value (logged gross inflows)
True value (logged gross inflows)
0
5
10
0
5
10

C.4 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 C.2: 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.