Analytics in Economics and Business

Dario Comanducci (elaborato per MD2SL 2024)

1 Gli effetti in Africa a lungo termine della tratta degli schiavi

Il lavoro pubblicato da Nunn in [8] affronta l'impatto della tratta degli schiavi sull'attuale povertà degli stati africani, impiegando un approccio standard basato su variabili strumentali. Utilizzando lo stesso dataset di [8], si è voluto qui applicare un'analisi simile attraverso l'approccio *Double ML*, impiegando come modelli di Machine Learning dei modelli ad albero (Bagging e Random Forest). I risultati ottenuti sono in linea con quelli in [8], mostrando un coefficiente negativo statisticamente significativo tra la variabili di risposta (PIL pro capite odierno degli Stati africani) ed il trattamento (la tratta degli schiavi).

Il presente lavoro è quindi organizzato come segue: § 2 e § 3 riportano rispettivamente un sunto dell'analisi storica condotta in [8] e la descrizione delle variabili nel dataset associato all'articolo stesso (visualizzando, in aggiunta a [8], i dati anche su mappe geografiche); per quanto riguarda invece lo svolgimento vero e proprio dell'elaborato, § 4 contiene un'analisi preliminare dei dati a disposizione, § 5 descrive come è stato organizzato l'approccio Double ML ai fini dell'analisi causale, ed infine § 6 riporta i risultati sperimentali sia replicando alcuni esiti ottenuti in [8], sia applicando la strategia delineata in § 5. Per concludere, § 7 riepiloga per sommi capi quanto svolto, mentre in appendice A sono riportati gli script Python principali.

2 Analisi storica per la tratta degli schiavi

Una spiegazione del sottosviluppo dell'Africa è la sua storia, caratterizzata da due eventi principali: la tratta degli schiavi ed il colonialismo. Mentre numerosi studi si sono concentrati sul legame tra l'esperienza coloniale dei Paesi africani ed il loro attuale sviluppo economico, [8] è stato un lavoro seminale per l'impatto che invece ha avuto la tratta degli schiavi sull'economia odierna di tali Stati.

Il dominio coloniale è ufficialmente durato dal 1885 fin verso il 1960, per un totale di circa 75 anni; a paragone, l'Africa ha sperimentato simultaneamente per un periodo di quasi 500 anni, dal 1400 al 1900, quattro traffici di schiavi:

- la *tratta transatlantica* con gli schiavi che vennero trasportati dall'Africa occidentale, centro-occidentale e orientale verso l'America;
- la *tratta transahariana*, dal sud del Sahara all'Africa settentrionale;
- la *tratta del Mar Rosso*, dall'entroterra del Mar Rosso verso il Medio Oriente e l'India;

- la *tratta dell'Oceano Indiano*, dall'Africa orientale con destinazioni in Medio Oriente e in India o nelle isole delle piantagioni nell'Oceano Indiano.

A differenza dei precedenti traffici di schiavi, individui della stessa o di etnie simili si schiavizzavano a vicenda. Ciò ebbe conseguenze particolarmente dannose, tra cui la frammentazione sociale ed etnica, l'instabilità politica e un conseguente indebolimento degli Stati, nonché la corruzione delle istituzioni giudiziarie. Analizziamo quanto detto più in dettaglio.

Frammentazione sociale ed etnica Gli schiavi erano in maggioranza catturati da villaggi o Stati che si razziavano a vicenda: ciò impedì la formazione di comunità più grandi e identità etniche più ampie anche laddove gruppi di villaggi si erano precedentemente sviluppati in federazioni su larga scala.

Instabilità politica Il clima di incertezza e insicurezza portava gli individui a procurarsi armi che erano fornite dagli Europei in cambio di schiavi, spesso ottenuti tramite rapimenti locali. Questo fatto aumentò ulteriormente la necessità di schiavizzare gli altri per proteggersi: come conseguenza di questo circolo vizioso non solo le comunità razziavano altre comunità, ma anche che i membri di una comunità rapivano altri appartenenti alla stessa comunità. L'instabilità dovuta a tali conflitti interni ha comportato in molti casi il crollo delle preesistenti forme di governo, spesso sostituite da razziatori di schiavi guidati da qualche signore della guerra; tali bande non furono in genere capaci di svilupparsi poi in Stati grandi e stabili.

Corruzione La tratta degli schivi causò anche la corruzione di strutture legislative precedentemente stabilite, contribuendo all'instabilità politica: divenne infatti prassi ottenere schiavi tramite accuse false di stregoneria o altri crimini, convertendo in condanne di schiavitù le precedenti sanzioni giudiziarie (percosse, pagamento di indennizzi o esilio).

La tratta degli schiavi ha avuto quindi un ruolo chiave sia nell'elevato frazionamento etnico dell'Africa di oggi (a sua volta legato allo sviluppo economico), sia nel sottosviluppo politico (fattore centrale per gli odierni Stati deboli dell'Africa).

2.1 La selezione dei Paesi africani nella tratta degli schiavi

Una spiegazione alternativa è che i Paesi inizialmente più sottosviluppati dal punto di vista economico e sociale abbiano "scelto" di partecipare alla tratta degli schiavi e che questi Paesi continuino a essere i più sottosviluppati oggi; tuttavia contro questa ipotesi possono essere portate varie evidenze storiche.

Nei primi scambi commerciali, che riguardavano in gran parte merci diverse dagli schiavi, solo le società africane con istituzioni sufficientemente sviluppate erano in grado di consentire il commercio con gli europei: ad esempio i portoghesi non trovarono alcuna società adeguata per il commercio europeo fino al fiume Zaire in quanto le comunità costiere locali erano troppo piccole per popolazione e territorio, con istituzioni economiche e sociali inadatte; viceversa dovettero raggiungere il regno del Congo (dotato di un governo centralizzato, una moneta nazionale e con reti commerciali ben sviluppate) per trovare una società in grado di sostenere il commercio con l'Europa.

Quando la domanda europea si rivolse quasi esclusivamente agli schiavi, la preferenza per il commercio con le parti più sviluppate dell'Africa continuò:

poiché le aree più prospere erano anche le più densamente popolate, un gran numero di schiavi poteva essere ottenuto facilmente istigando guerre civili o conflitti; inoltre le società meno sviluppate, anche più violente e ostili, furono spesso in grado di resistere agli europei per secoli (i portoghesi furono costretti a concentrare i loro sforzi lungo la costa più a Sud, in particolare).

Una seconda potenziale fonte di selezione potrebbe essere stata rappresentata dal fatto che le società che inizialmente avevano la schiavitù domestica potrebbero essersi selezionate per la tratta degli schiavi: evidenze storiche indicano che la schiavitù domestica fosse presente nelle aree dell'Africa coinvolte nelle più antiche tratte islamiche di schiavi, ma non è chiaro se la schiavitù domestica sia stata una causa o una conseguenza delle tratte esterne di schiavi. Sono comunque emerse prove indicanti che la schiavitù domestica potrebbe non essere esistita prima della tratta transatlantica [6, 9]. Inoltre vari studi linguistici hanno mostrato che nell'Africa centro-occidentale ed in altre zone dell'Africa non esisteva una parola per "schiavo": semmai il significato delle parole inizialmente usate per indicare "servo" o "prigioniero" cambiò in "schiavo commercializzato" successivamente al XVI secolo, propagandosi dai porti costieri impegnati nel commercio degli schiavi nelle comunità dell'entroterra, anch'esse coinvolte nel commercio.

3 Le variabili d'interesse

Per avvalorare su dati osservazionali l'ipotesi avanzata, in [8] è proposto un modello di regressione lineare che, oltre a risposta e trattamento, impiega diverse variabili di controllo ed alcune variabili strumentali. Tutti i dati nel dataset per il modello impiegato in [8] sono riassunti in Tab. 1 e di seguito descritti.[1]

3.1 Risposta e trattamento

Come indice della ricchezza o povertà di ciascun Paese africano è stato impiegato il PIL (valori aggiornati all'anno 2000); in particolare, in [8] la variabile di risposta y viene definita come il logaritmo naturale del PIL pro capite:

$$y_n = \ln(\text{PIL pro capite per la nazione } n) \tag{1}$$

Il "trattamento" invece è rappresentato dalle deportazioni di schiavi subite da ciascuna nazione africana: in [8] il numero di schiavi deportati da ciascun paese africano è stato poi normalizzato rispetto all'area di ogni nazione, prendendone infine il logaritmo naturale; pertanto il *trattamento* è modellato per ciascuna nazione n come[2]

$$d_n = \ln \frac{\text{numero di schiavi deportati dalla nazione } n}{\text{superficie della nazione } n} \tag{2}$$

Per una descrizione di come siano state ricostruite le stime sulle deportazioni degli schiavi per i vari Paesi, si faccia riferimento a [8].[3]

[1] Dataset disponibile in https://nathannunn.sites.olt.ubc.ca/files/2022/02/DATA-22The-Long-Term-Effects-of-Africas-Slave-Trades22-slave_trade_qje_replication_dataset.zip

[2] Per i Paesi che non hanno subito deportazioni, il numero di schiavi è stato posto pari a 0.1 per evitare che il logaritmo naturale producesse $-\infty$ come valore.

[3] Brevemente, attraverso informazioni storiche ricostruite da più fonti sulla base dei registri di spedizione e vendita, le stime sono state ottenute combinando il numero di schiavi deportati dai porti africani con le percentuali per etnia in modo da ripartire gli schiavi nei vari Paesi.

Tab. 1: Variabili presenti nel dataset usato in [8]

Tipologia	Variabile	Descrizione originale tratta dal dataset
risposta y	`ln_maddison_pcgdp2000`	Log per capita GDP
trattamento d	`ln_export_area`	Log slave exports normalized by land area
passato colon.	`colony0`	Colonizer indicator: not colonized
	`colony1`	Colonizer indicator: Britan
	`colony2`	Colonizer indicator: France
	`colony3`	Colonizer indicator: Portugal
	`colony4`	Colonizer indicator: Belgium
	`colony5`	Colonizer indicator: Spain
	`colony6`	Colonizer indicator: UN
	`colony7`	Colonizer indicator: Italy
geografia/clima	`abs_latitude`	Absolute latitude
	`longitude`	Longitude
	`rain_min`	Min of monthly average rainfall (mm)
	`humid_max`	Max monthly afternoon avg humidity (%)
	`low_temp`	Min of avg monthly low temp (C)
	`ln_coastline_area`	Ocean proximity Log(coastline/land_area)
religione	`islam`	Percent Islamic
legislazione	`legor_fr`	Legal origin indicator: French
	`legor_uk`	Legal origin indicator: British
regioni africane	`island_dum`	Indicator variable for small islands
	`region_n`	Region indicator: North
	`region_s`	Region indicator: South
	`region_w`	Region indicator: West
	`region_e`	Region indicator: East
	`region_c`	Region indicator: Central
prod. mineraria	`ln_avg_gold_pop`	Log gold production per capita
	`ln_avg_oil_pop`	Log oil production per capita
	`ln_avg_all_diamonds_pop`	Log diamond production per capita
IV	`atlantic_distance_minimum`	Min Atlantic distance (000s of kms)
	`indian_distance_minimum`	Min Indian distance (000s of kms)
	`saharan_distance_minimum`	Min Saharan distance (000s of kms)
	`red_sea_distance_minimum`	Min Red Sea distance (000s of kms)
extra	`ln_export_pop`	Log slave exports norm. by historic popul.
	`ln_pop_dens_1400`	Log population density in 1400
	`ethnic_fractionalization`	Ethnic fractionalization
	`state_dev`	State development
	`land_area`	Land area in millions of square kms

Nel dataset associato a [8] la risposta y è data da `ln_maddison_pcgdp2000`, mentre il trattamento d corrisponde a `ln_export_area`. Fig. 1 illustra y e d per ogni Paese africano (codici dei Paesi in Tab. 2).[4]

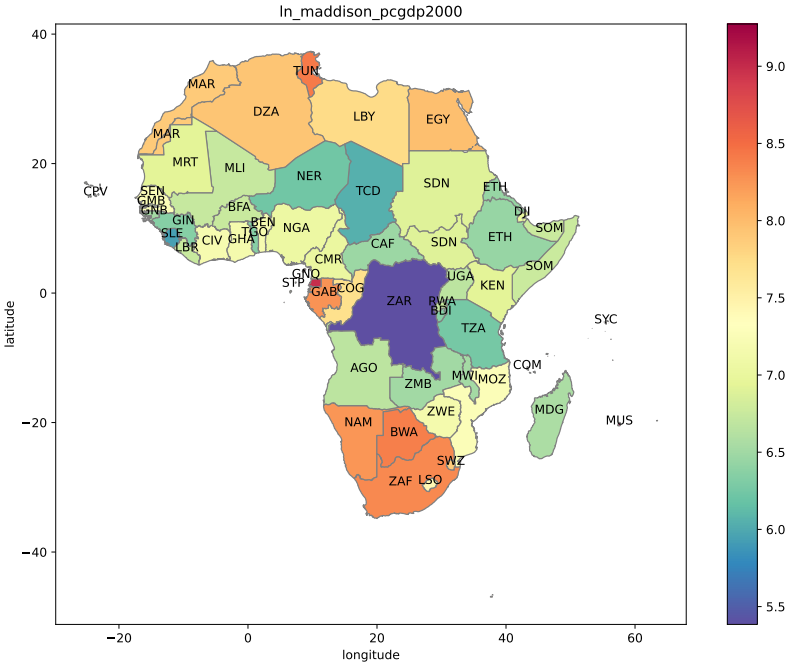
3.2 Variabili di controllo

3.2.1 Controlli di base

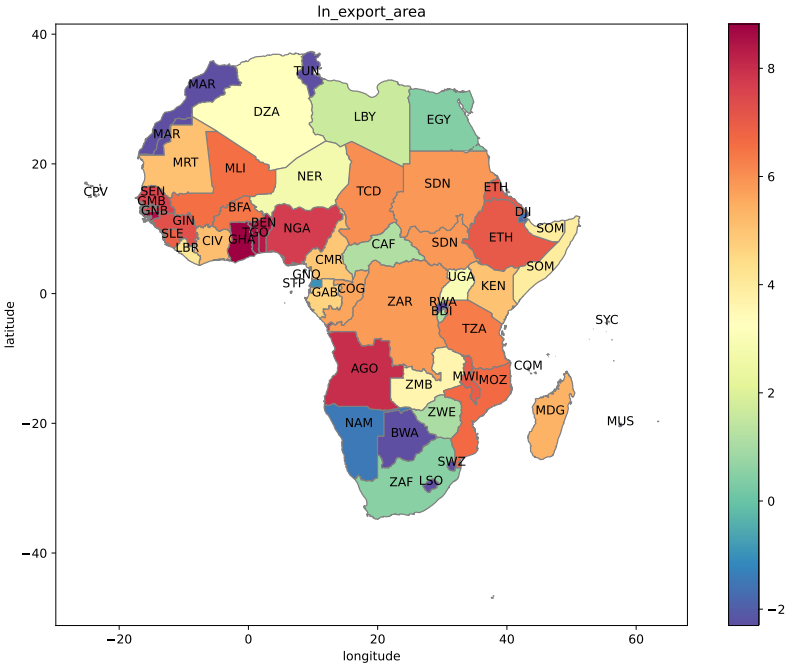
L'equazione base del modello impiegato in [8] è

$$y_n = \beta_0 + \beta_1 d_n + \boldsymbol{\gamma}^\top \mathbf{x}_n + \boldsymbol{\delta}^\top \mathbf{c}_n + \varepsilon_n \tag{3}$$

[4] Rispetto ai dati in [8] gli Stati africani sono cambiati, ad esempio è sorto lo Stato del Sudan del Sud: in tutte le mappe (shapefile da https://www.naturalearthdata.com/) di questo report si è cercato di fondere, laddove necessario, gli attuali Paesi in modo da replicare la conformazione geopolitica corrispondente a [8].

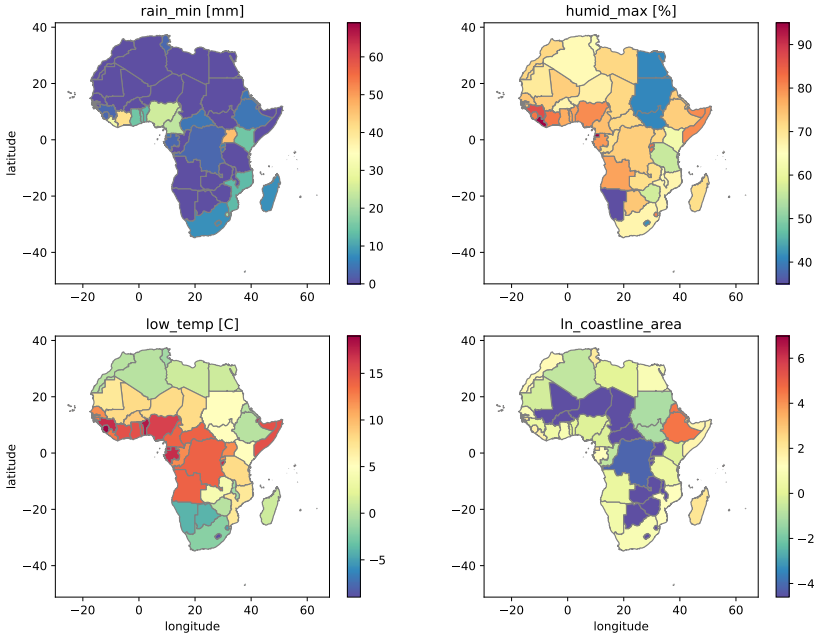


(a) $y =$ `ln_maddison_pcgdp2000`

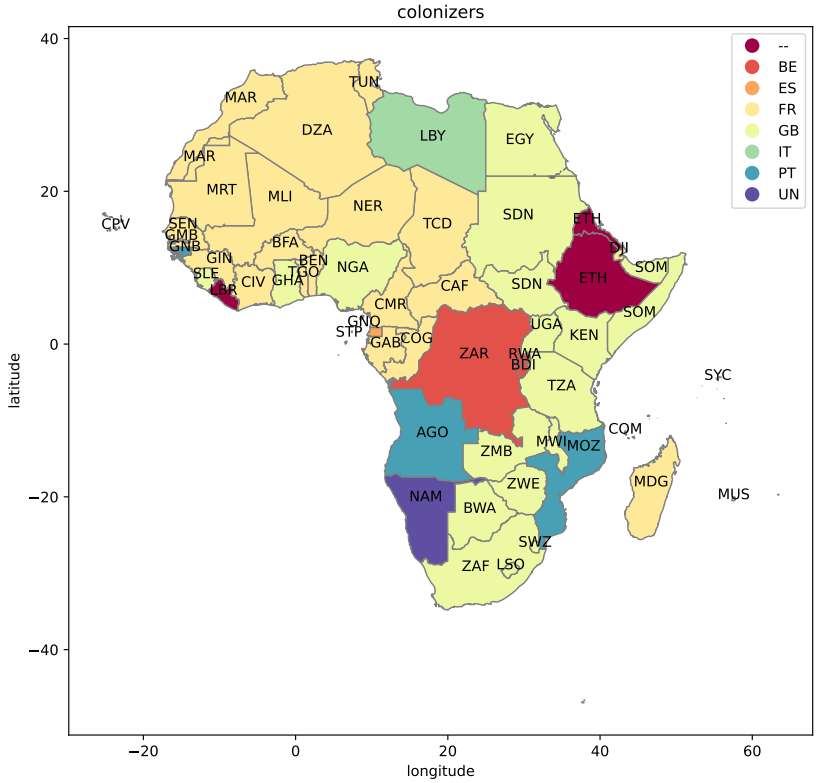


(b) $d =$ `ln_export_area`

Fig. 1: Risposta y e trattamento d per ogni Paese africano [8]. *Spicca in (a) il PIL pro capite molto basso della Repubblica Democratica del Congo (ex Zaire), probabilmente per la cleptocrazia del dittatore Mobutu fino alla fine degli anni '90 (i dati del PIL risalgono al 2000).*



(a) Variabili geoclimatiche (tranne latitudine e longitudine dei centroidi)



(b) Ultimi colonizzatori di ciascun Paese

Fig. 2: Controlli basilari usati in [8].

Tab. 2: Codici impiegati nelle mappe per i Paesi africani.

isocode	Paese	isocode	Paese
AGO	Angola	MDG	Madagascar
BDI	Burundi	MLI	Mali
BEN	Benin	MOZ	Mozambique
BFA	Burkina Faso	MRT	Mauritania
BWA	Botswana	MUS	Mauritius
CAF	Central African Republic	MWI	Malawi
CIV	Ivory Coast	NAM	Namibia
CMR	Cameroon	NER	Niger
COG	Congo	NGA	Nigeria
COM	Comoros	RWA	Rwanda
CPV	Cape Verde Islands	SDN	Sudan
DJI	Djibouti	SEN	Senegal
DZA	Algeria	SLE	Sierra Leone
EGY	Egypt	SOM	Somalia
ETH	Ethiopia	STP	Sao Tome & Principe
GAB	Gabon	SWZ	Swaziland
GHA	Ghana	SYC	Seychelles
GIN	Guinea	TCD	Chad
GMB	Gambia	TGO	Togo
GNB	Guinea-Bissau	TUN	Tunisia
GNQ	Equatorial Guinea	TZA	Tanzania
KEN	Kenya	UGA	Uganda
LBR	Liberia	ZAF	South Africa
LBY	Libya	ZAR	Dem. Rep. of Congo
LSO	Lesotho	ZMB	Zambia
MAR	Morocco	ZWE	Zimbabwe

dove oltre alla risposta y, al trattamento d ed al rumore ε, compaiono anche delle variabili di controllo $\mathbf{x}$ e $\mathbf{c}$ (Fig. 2):

- $\mathbf{x}$ contiene variabili che mirano a catturare le differenze nella geografia e nel clima dei Paesi (Fig. 2a) ossia[5]
  - distanza dall'equatore `abs_latitude` (il valore assoluto della latitudine del centroide di ogni paese, misurata in gradi),
  - longitudine `longitude` (longitudine del centroide di ogni paese, misurata anch'essa in gradi),
  - precipitazioni minime mensili `rain_min` (le precipitazioni totali medie, misurate in millimetri, nel mese più secco dell'anno),
  - umidità massima media `humid_max` (la media dell'umidità massima pomeridiana, misurata in percentuale, durante il mese più caldo dell'anno),
  - temperatura minima media `low_temp` (la temperatura media mensile più bassa misurata in gradi Celsius),

[5] Per cinque Stati in cui il baricentro si trova al di fuori dei confini terrestri del paese (Gambia, Somalia, Capo Verde, Mauritius e Seychelles) viene utilizzato il punto all'interno del Paese più vicino al baricentro. I dati meteorologici sono stati raccolti in un periodo di 30 anni.

  - vicinanza all'oceano `ln_coastline_area` misurata dal logaritmo naturale della linea costiera (misurata in migliaia di chilometri) diviso per la superficie di ogni paese;

- **c** è un vettore di variabili dummy per descrivere il dominio coloniale (l'altro evento significativo del passato dell'Africa), ed indicano l'ultima dominazione coloniale prima dell'indipendenza (Fig. 2b):
  - `colony0`, non colonizzato;
  - `colony1`, Regno Unito (GB);
  - `colony2`, Francia (FR);
  - `colony3`, Portogallo (PT);
  - `colony4`, Belgio (BE);
  - `colony5`, Spagna (ES);
  - `colony6`, ONU (UN);
  - `colony7`, Italia (IT).

Tutti i fattori in **x**, eccetto la longitudine, influenzano se un paese ha un clima tropicale, il che influisce sulla prevalenza di malattie infettive e sulla produttività agricola; la longitudine è inclusa anche per catturare le differenze tra le parti orientali e occidentali del continente africano.

3.2.2 Nord Africa, isole e restanti aree geografiche

Altre variabili di controllo aggiuntive servono per tener conto delle potenziali differenze tra isole o Paesi nordafricani e il resto dell'Africa.

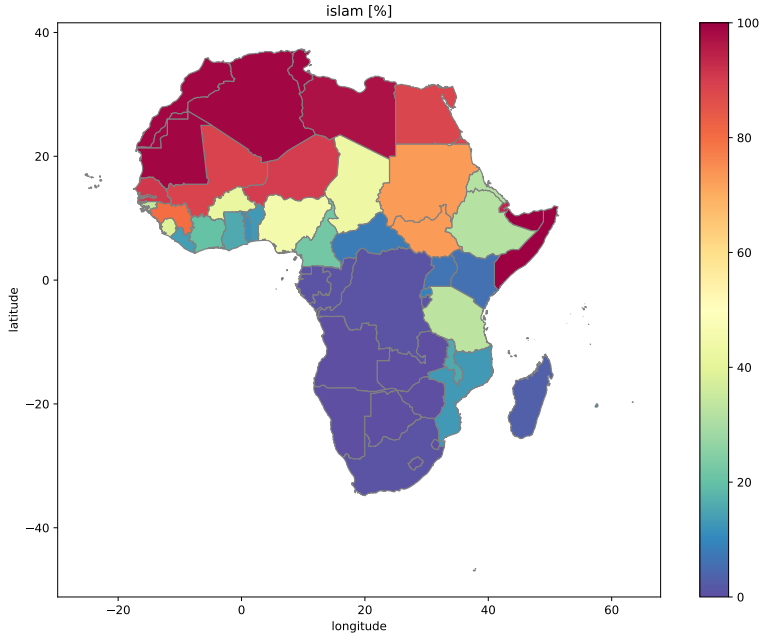
Le due differenze fondamentali tra il Nord Africa e il resto del continente sono che i paesi nordafricani sono prevalentemente islamici e con sistemi legali basati sul diritto civile francese (Fig. 3). Per catturare queste differenze, in [8] il modello di Eq. (3) è esteso includendo anche la percentuale della popolazione con fede islamica (`islam` nel dataset) e una variabile `legor_fr` indicatrice dell'origine francese del sistema legislativo; per completezza il dataset comprende anche una dummy `legor_uk` per l'eventuale origine britannica del sistema legislativo.

Vengono inoltri aggiunti degli effetti fissi (1 o 0) per indicare se uno stato è composto da piccole isole (`island_dum`) o meno, e le rispettive dummy per denotare gli stati che si trovano nel Nord, Sud, Ovest, Est e nella regione centrale dell'Africa (rispettivamente `region_n`, `region_s`, `region_w`, `region_e` e `region_c` nel dataset).

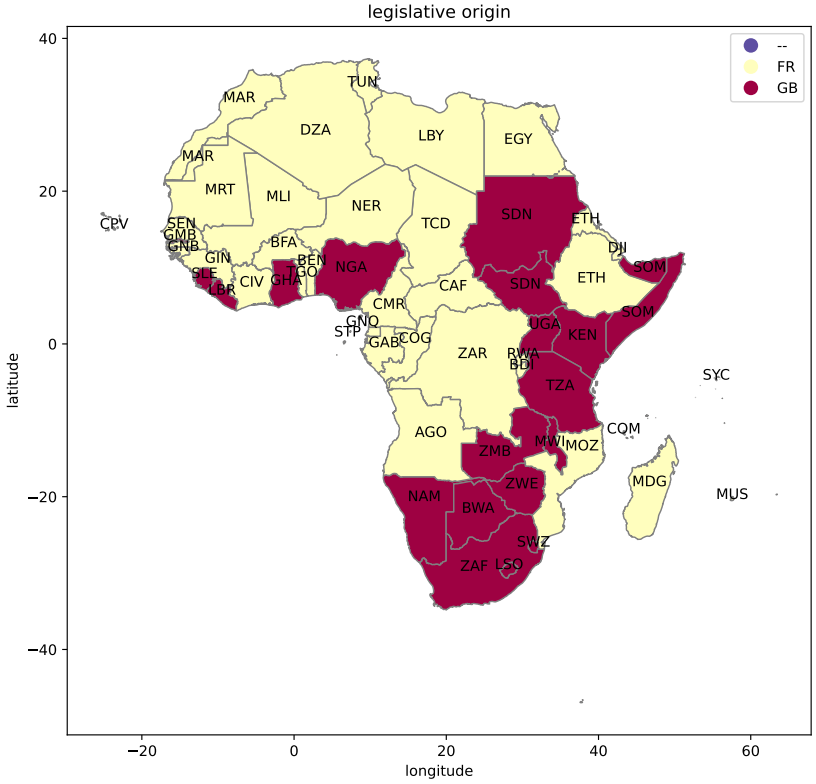
3.2.3 Produzione mineraria

L'ultimo fattore da controllare sono le differenze nelle dotazioni di risorse naturali dei vari paesi; in [8] vengono quindi incluse le produzioni di oro, petrolio e diamanti, misurate come il logaritmo naturale della produzione media annuale per mille abitanti dal 1970 al 2000 (Fig. 4):

- l'oro estratto è misurato in kilogrammi (`ln_avg_gold_pop`);
- il petrolio greggio è misurato in migliaia di tonnellate (`ln_avg_oil_pop`);
- i diamanti includono sia le pietre preziose che i diamanti industriali e sono misurati in migliaia di carati (`ln_avg_all_diamonds_pop`).

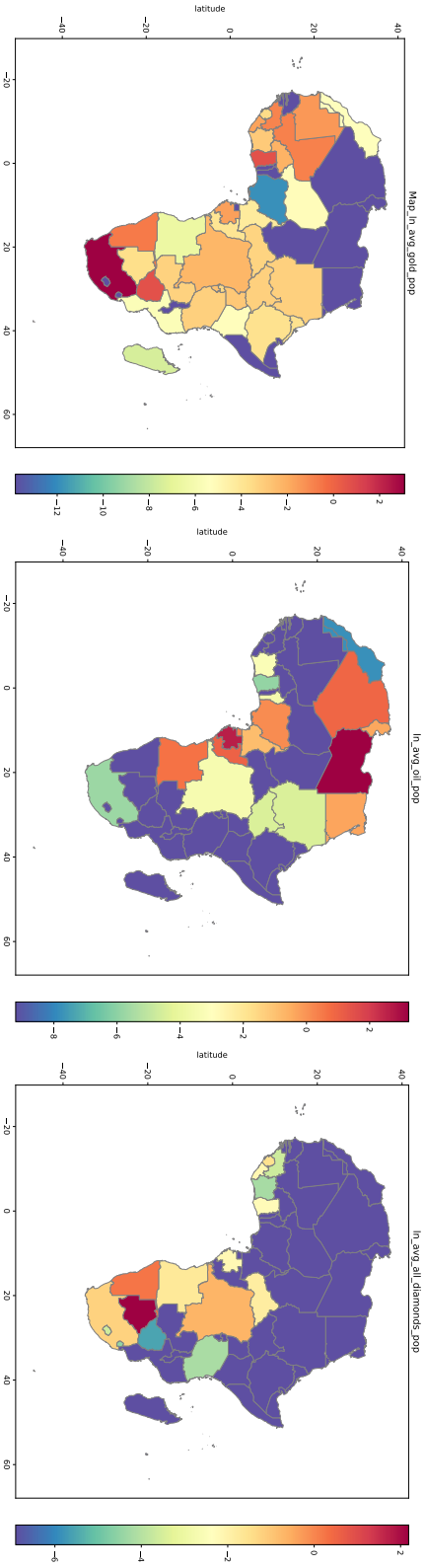


(a) Percentuale di popolazione con fede islamica



(b) Base del sistema legislativo

Fig. 3: Caratterizzazione del Nord-Africa rispetto al resto del continente [8].



(a) Oro (b) Petrolio (c) Pietre preziose

Fig. 4: Risorse minerarie per ciascun Paese misurate come il logaritmo naturale della produzione media annuale (kg per l'oro, 10^6 kg per il petrolio, 10^3 carati per le pietre preziose) ogni mille abitanti [8].

3.3 Variabili strumentali

Per capire se la tratta degli schiavi abbia un impatto causale sul PIL attuale delle nazioni africane in [8] viene fatto ricorso ad alcune variabili strumentali (IV) correlate con le esportazioni di schiavi, ossia le distanze tra ciascun Paese africano e le località in cui gli schiavi venivano richiesti. Poiché la posizione dell'offerta non è stata influenzata dalla localizzazione della domanda, tali IV costituiscono degli strumenti validi.[6]

In [8] le IV sono costruite sulle distanze di ciascun Paese rispetto alle destinazioni più importanti di ciascuna tratta degli schiavi (Fig. 5).

Tratta transatlantica La *distanza di navigazione* `atlantic_distance_minimum` dal punto della costa africana più prossimo al centroide del Paese al più vicino mercato principale in America.

Tratta sull'Oceano Indiano La *distanza di navigazione* `indian_distance_minimum` dal punto della costa a minima distanza dal centroide del Paese alla più vicina delle due principali destinazioni, cioè Mauritius e Mascate (in Oman).

Tratta trans-sahariana La *distanza via terra* `saharan_distance_minimum` dal centroide di un Paese al porto di esportazione più vicino, cioè Algeri, Tunisi, Tripoli, Bengasi e Il Cairo.

Tratta del Mar Rosso La *distanza via terra* `red_sea_distance_minimum` dal centroide dei Paesi al porto di esportazione più vicino tra Massaua, Suakin e Gibuti.

Per i Paesi insulari, non è possibile raggiungere i porti della tratta degli schiavi del Sahara o del Mar Rosso viaggiando via terra: per questi Paesi in [8] è stata utilizzata la somma della distanza di navigazione e della distanza via terra.

3.4 Variabili extra nel dataset

Il dataset disponibile per [8] contiene infine quattro variabili extra, usate nell'articolo per validare alcune asserzioni ma non coinvolte nell'analisi causale:

- `ln_export_pop`, il logaritmo naturale delle esportazioni totali di schiavi normalizzate per popolazione storica;
- `ln_pop_dens_1400`, il logaritmo della densità demografica nel 1400
- `ethnic_fractionalization`, frazionamento etnico;
- `state_dev`, sviluppo dello Stato;
- `land_area`, superficie terrestre in milioni di km quadrati.

[6] La localizzazione della domanda di schiavi africani era determinata da una serie di fattori, tutti non collegati all'offerta di schiavi, come ad esempio: nelle Indie Occidentali e negli Stati Uniti meridionali, gli schiavi venivano richiesti nella coltivazione di zucchero e tabacco per il clima adatto a tali prodotti; l'esistenza di miniere d'oro e d'argento è stata determinante per la domanda di schiavi in Brasile; nel Sahara settentrionale, in Arabia e in Persia, gli schiavi erano necessari per lavorare nelle miniere di sale e nell'area del Mar Rosso venivano utilizzati come sommozzatori di perle.

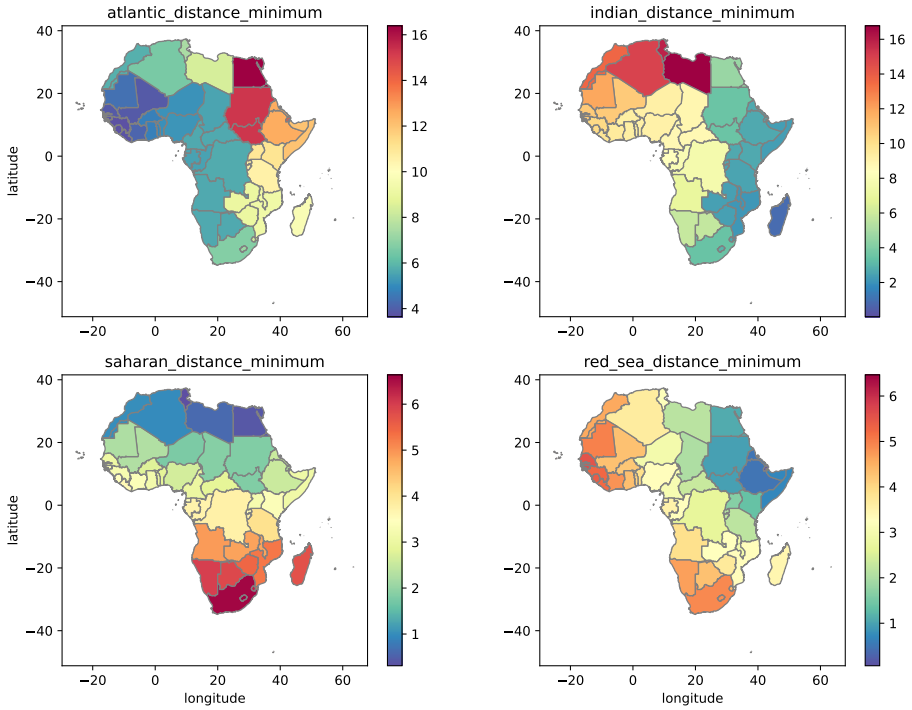


Fig. 5: Varabili strumentali in [8].

4 Analisi dei dati

Le variabili elencate in Tab. 1 si dividono in

- continue, ossia
  - `ln_maddison_pcgdp2000`
  - `ln_export_area`
  - `abs_latitude`, `longitude`
  - `rain_min`, `humid_max`, `low_temp`, `ln_coastline_area`
  - `islam`
  - `ln_avg_gold_pop`, `ln_avg_oil_pop`, `ln_avg_all_diamonds_pop`
  - `atlantic_distance_minimum`, `indian_distance_minimum`, `saharan_distance_minimum`, `red_sea_distance_minimum`
- discrete (binarie/categoriche)
  - `colony0` ... `colony7` (passato coloniale)
  - `legor_fr` e `legor_uk` (origine legislativa)
  - `island_dum`, `region_n`, `region_s`, `region_w`, `region_e`, `region_c` (regioni geografiche)

Analizziamo di seguito come i vari dati si relazionano tra di loro (per semplicità omettiamo analisi incrociate tra variabili di tipo continuo e discreto).

4.1 Variabili continue

Iniziamo analizzando le distribuzioni delle variabili continue e delle loro correlazioni (Fig. 6): dati i pochi esempi (gli Stati africani in esame sono solo 52), gli istogrammi non forniscono molte indicazioni sulle distribuzioni delle variabili considerate se non che alcune variabili sembrano mostrare una distribuzione bimodale come

- il trattamento stesso `ln_export_area`, con alcuni Paesi che hanno subito poco tratta degli schiavi (riscontrabile anche da Fig. 1b);
- ciascuna risorsa mineraria, con molti paesi privi della risorsa in questione (tuttavia, da Fig. 4 si nota anche che laddove manca una risorsa mineraria, lo stesso Paese dispone di almeno una delle altre due);
- la religione islamica, confermata da Fig. 3a;
- la vicinanza all'oceano, per come è stata definita `ln_coastline_area` in quanto i paesi dell'Africa centrale sono privi di linea costiera (Fig. 2a);
- la variabile strumentale `indian_distance_minimum`, forse per la dislocazione dei due soli mercati di destinazione nella tratta indiana (Fig. 5).

Viceversa l'analisi di correlazione mostra varie relazioni ossia

- esiste una forte correlazione tra `longitude` e tre variabili strumentali[7] (correlazione positiva per `atlantic_distance_minimum`, negativa per `indian_distance_minimum` e `red_sea_distance_minimum`);
- la tratta transahariana (`saharan_distance_minimum`) è invece correlata con la percentuale di fede islamica nei vari Stati (`islam`);
- le distanze della tratta transatlantica sono negativamente correlate con le distanze della tratta indiana e quella del mar Rosso;
- le distanze della tratta indiana sono negativamente correlate anche a quelle della tratta transahariana;
- la temperatura minima (`low_temp`) è negativamente correlata con la distanza dall'equatore (`abs_latitude`) e positivamente con l'umidità (`humid_max`).

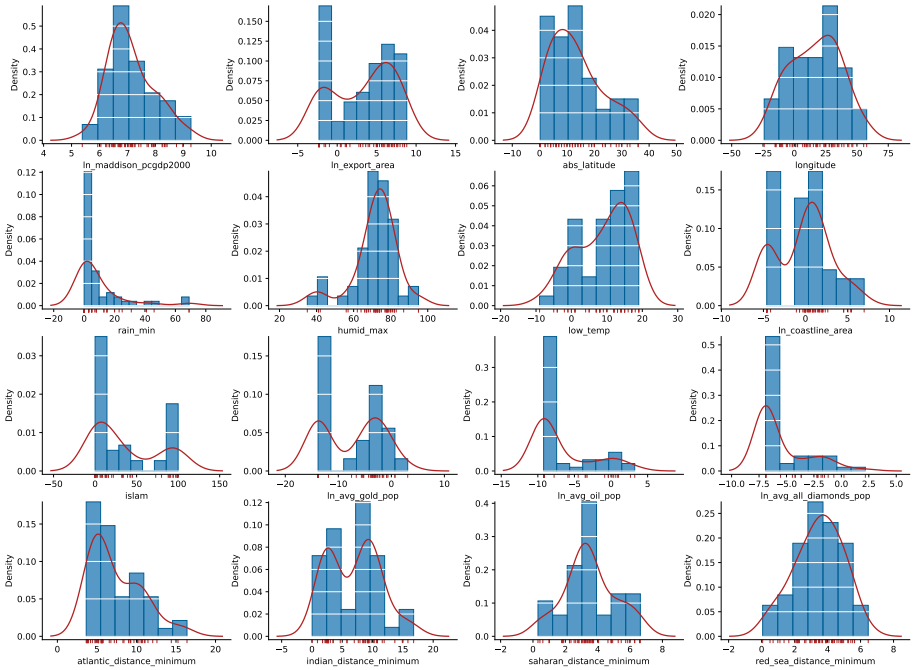
Tralasciando `low_temp`, `abs_latitude` e `humid_max`, Fig. 7 riporta i grafici di dispersione tra le variabili appena menzionate.

4.2 Variabili discrete

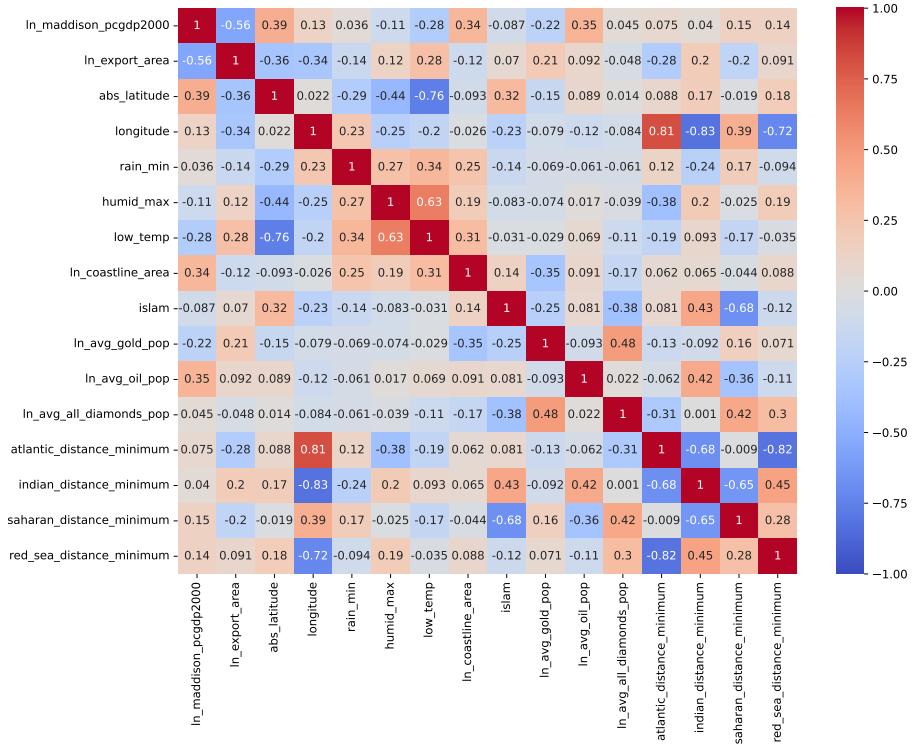
In Tab 3 sono riportate le matrici di contingenza per le variabili discrete, "riconvertite" in categoriche avendole raggruppate per tipologia:

- passato coloniale (si faccia riferimento anche a Fig. 2b);
- origine legislativa (si veda anche Fig. 3b);
- regioni geografiche.

[7] Ricordiamo che le IV sono date dalle distanze di ciascun Paese rispetto alle destinazioni più importanti per ciascuna tratta degli schiavi.



(a) distribuzione campionaria



(b) matrice di correlazione

Fig. 6: Analisi d'insieme delle variabili continue.

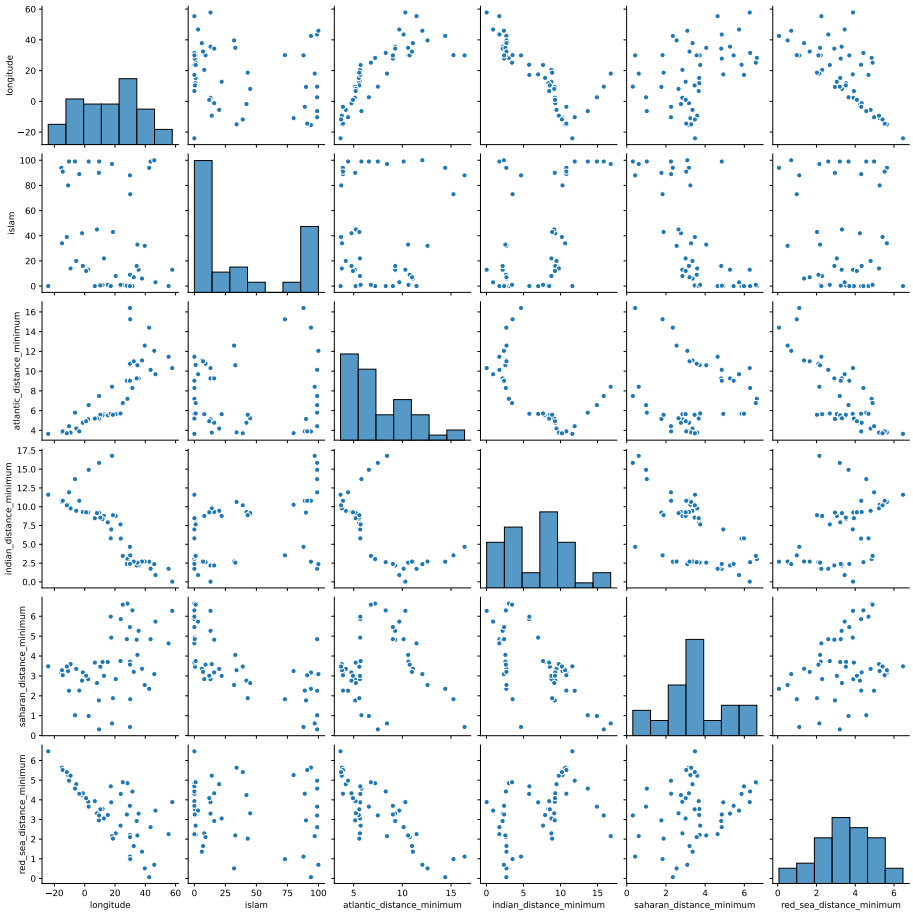


Fig. 7: grafici di dispersione tra le variabili continue con maggior correlazione.

Come in effetti si poteva dedurre osservando Fig. 2b e Fig. 3b, per i dati in [8], da tali tabelle si evince che l'Africa ha subito prevalentemente l'influenza francese e quella britannica: i Francesi maggiormente nei Paesi nord-occidentali e centrali dell'Africa; gli Inglesi nel Sud e nell'Est del continente africano.

5 Analisi causale tramite Double ML

Mentre in [8] l'analisi causale per l'impatto della tratta degli schivi sul PIL attuale degli Stati africani è stato condotta in maniera classica tramite delle variabili strumentali applicando la tecnica Two Stage LS (Least Squares), si vuole qui analizzare il problema con modelli di Machine Learning (ML) secondo l'approccio *Double ML* applicato ad un *modello parzialmente lineare* (PLM) tra la risposta Y ed il trattamento D, data una serie di controlli $\mathbf{X}$ [4]:

$$y_n = \alpha d_n + g(\mathbf{x}_n) + \varepsilon_{yn} \qquad \mathbb{E}[\varepsilon_y | X, D] = 0 \tag{4a}$$

$$d_n = m(\mathbf{x}_n) + \varepsilon_{dn} \qquad \mathbb{E}[\varepsilon_d | X] = 0 \tag{4b}$$

Tab. 3: Matrici di contingenza per le variabili discrete (binarie) raggruppate per tipologia

(a)

		base legislativa		Aree geografiche					
passato coloniale	Tot.	legor_fr	legor_uk	island_dum	region_n	region_s	region_w	region_e	region_c
colony_0 −−	2	1	1	0	0	0	1	1	0
colony_1 GB	18	2	16	1	1	7	4	6	0
colony_2 FR	21	21	0	2	3	0	9	4	5
colony_3 PT	5	5	0	2	0	1	2	0	2
colony_4 BE	3	3	0	0	0	0	0	2	1
colony_5 ES	1	1	0	0	0	0	0	0	1
colony_6 UN	1	0	1	0	0	1	0	0	0
colony_7 IT	1	1	0	0	1	0	0	0	0
Tot.	52	34	18	5	5	9	16	13	9

(b)

		Aree geografiche					
base legisl.	Tot.	island_dum	region_n	region_s	region_w	region_e	region_c
legor_fr	34	5	5	1	11	8	9
legor_uk	18	0	0	8	5	5	0
Tot.	52	5	5	9	16	13	9

essendo $g(\mathbf{x})$ e $m(\mathbf{x})$ due modelli predittivi di Machine Learning (ML), mentre ε_{yn} e ε_{dn} costituiscono i residui.

Parzializzando Eq. (4) rispetto a $\mathbf{X}$ nello stile di Frisch-Waugh-Lovell, per eliminare il bias dei modelli ML, il problema diventa quello di determinare due modelli $\hat{l}(\mathbf{x})$ e $\hat{m}(\mathbf{x})$ secondo

$$y_n = l(\mathbf{x}_n) + v_n \tag{5a}$$
$$d_n = m(\mathbf{x}_n) + u_n \tag{5b}$$
$$v_n = \alpha u_n \tag{5c}$$

con α dato da OLS (Ordinary Least Squares) su $\hat{v} = y - \hat{l}(\mathbf{x})$ e $\hat{u} = d - \hat{m}(\mathbf{x})$.

5.1 Cross-fitting

Al fine di ridurre il problema del bias nei modelli ML dovuto ad overfitting, la stima di α avviene attraverso una procedura di cross-fitting[8] [1, 4, 5] consentendo al contempo l'uso di tutti i dati:

- il dataset è partizionato in K sottoinsiemi disgiunti $\{I_1 \dots I_k\}$;
- per $k = 1 \dots K$ i modelli $l(\mathbf{x})$ e $m(\mathbf{x})$ in Eq. (5) vengono addestrati sui dati di $K-1$ insiemi $\{I_j : j \neq k\}$, per poi calcolare i valori predetti $\hat{l}_k(\mathbf{x}_n)$ e $\hat{m}_k(\mathbf{x}_n)$ sul restante insieme I_k (ossia solo per $\mathbf{x}_n \in I_k$);
- infine, calcolando i residui per ciascun insieme I_k

  $$\hat{v}_n = y_n - \hat{l}_k(\mathbf{x}_n) \qquad \text{(solo per } \mathbf{x}_n, y_n \in I_k\text{)}$$
  $$\hat{u}_n = d_n - \hat{m}_k(\mathbf{x}_n) \qquad \text{(solo per } \mathbf{x}_n, d_n \in I_k\text{)}$$

  otteniamo la stima di α attraverso la regressione su tutti i $\hat{v}_n$ rispetto ai valori di $\hat{u}_n$ risolvendo $\hat{\mathbf{u}}\alpha = \hat{\mathbf{v}}$, cioè

  $$\hat{\mathbf{u}} = [\hat{u}_1 \ \dots \ \hat{u}_N]^\top \tag{6a}$$
  $$\hat{\mathbf{v}} = [\hat{v}_1 \ \dots \ \hat{v}_N]^\top \tag{6b}$$
  $$\hat{\alpha} = \frac{\hat{\mathbf{u}}^\top \hat{\mathbf{v}}}{\hat{\mathbf{u}}^\top \hat{\mathbf{u}}} \tag{6c}$$

5.2 Modelli ML impiegati

Tra i modelli di predizione sono stati considerate due tipologie di modelli ad albero: *Bagging* e *Random Forest*. Entrambi i metodi sono basati sull'impiego di B alberi ciascuno addestrato su una variante con reimbussolamento del dataset originale (*boostrap*), avendo in uscita la media dei valori predetti da ogni albero:

$$f(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^{B} f_b(\mathbf{x})$$

Gli alberi delle Random Forest ad ogni split impiegano solo sottoinsiemi casuali delle variabili a disposizione, mentre nel Bagging si sfruttano sempre tutti i

[8] In [4] la versione qui impiegata è indicata come algoritmo DML2.

dati; pertanto, in aggiunta al numero B di alberi, le Random Forest hanno un ulteriore iper-parametro: il numero s di variabili impiegabili ad ogni split.

L'impiego di un ensemble di alberi e del bootstrap sono in questo contesto molto significativi, essendo il dataset in questione composto da sole 52 osservazioni (tante quanti gli Stati africani), a fronte di una trentina di variabili.

5.2.1 Modalità di scelta per gli iperparametri

Esistono diversi modi per combinare la procedura di cross-fitting con la cross-validazione per impostare gli iper-parametri nei modelli ML [3]:

- *Full sample*, in cui gli iper-parametri vengono ottimizzati tramite cross-validazione sull'intero dataset, su partizioni indipendenti da quelle impiegate per il cross-fitting successivo;

- *Split sample*, dove il dataset è diviso in due parti, una impiegata per la cross-validazione degli iper-parametri e l'altra per il cross-fitting;

- *On folds*, con la stessa partizione del cross-fittimg, ottimizzando ogni volta gli iper-parametri sulla porzione di dati per l'addestramento dei modelli.

Secondo [3] in generale la modalità Full sample oppure On folds hanno prestazioni simili, ed entrambe le modalità sono preferibili alla Split sample specialmente su dataset piccoli. Pertanto in questo lavoro gli iperparametri per Bagging e Random Forest verranno selezionati secondo la modalità Full sample.

I modelli di ML verranno valutati anche per modellare le variabili strumentali.

5.2.2 Selezione dei modelli ML

Avendo deciso di valutare l'impiego di più possibili tipologie di modelli ML (Bagging e Random Forest), è necessario stabilire quale adottare nel cross-fitting per la predizione di y e di d; a tale scopo il criterio impiegato è quello di scegliere il modello che predice meglio: dagli errori di predizione riassunti dal valor medio dei residui al quadrato sui valori forniti da ciascun modello per $\hat{v}$ e $\hat{u}$ (Eq. (6))

$$\mathbb{E}[v^2] = \frac{1}{N}\sum_n \hat{v}_n^2 \tag{7a}$$

$$\mathbb{E}[u^2] = \frac{1}{N}\sum_n \hat{u}_n^2 \tag{7b}$$

vengono selezionati i modelli con gli errori più bassi [5, pp. 252-253].

6 Risultati sperimentali

Prima di applicare l'approccio Double ML in § 6.2, si è voluto verificare che i valori nel dataset confermino i risultati dell'artcolo, replicandone alcuni in § 6.1.

6.1 Replica delle stime riportate nell'articolo

6.1.1 Stime OLS senza IV

In [8] la dipendenza tra la variabile d'uscita y = `ln_maddison_pcgdp2000` ed il trattamento d = `ln_export_area` è stata valutata aggiungendo via via i vari controlli nella stima di un modello lineare tramite Ordinary Least Squares (OLS); proviamo quindi a replicare alcuni dei risultati ottenunti in [8]:[9]

1. all'inizio con solo l'effetto fisso del passato coloniale (`colony0`...`colony7`);[10]
2. poi inserendo anche i controlli geo-climatici (`abs_latitude`, `longitude`, `rain_min`, `humid_max`, `low_temp`, `ln_coastline_area`);
3. successivamente tenendo conto dei fattori caratterizzanti il Nord-Africa e le piccole isole (`islam`, `legor_fr`, `island_dum`, `region_n`);[11]
4. infine sono state prese in considerazione anche le produzioni minerarie (ossia `ln_avg_gold_pop`, `ln_avg_oil_pop`, `ln_avg_all_diamonds_pop`).

A differenza di [8], dove è stato analizzato anche il caso senza i Paesi del Nord-Africa e quelli formati da piccole isole in quanto possibili outlier, Tab. 4 riporta solo le stime del modello lineare su tutti i Paesi poiché l'esclusione di tali Stati non cambia eccessivamente il valore del coefficiente associato al trattamento; com'era lecito aspettarsi, i risultati in Tab. 4 corrispondono ai valori in [8].

6.1.2 Two Stage LS

L'uso delle quattro variabili strumentali (z_1 = `atlantic_distance_minimum`, z_2 = `indian_distance_minimum`, z_3 = `saharan_distance_minimum` e z_4 = `red_sea_distance_minimum`) richiede un doppio passaggio di stime OLS:

- nel primo passaggio viene modellata la dipendenza tra il trattamento d e le variabili strumentali $\mathbf{z} = [z_1\ z_2\ z_3\ z_4]^\top$

$$\begin{aligned} d_n &= \beta_0' + \mu_1' z_{1n} + \mu_2' z_{2n} + \mu_3' z_{3n} + \mu_4' z_{4n} + \nu_1' x_{1n} + \ldots + \nu_K' x_{Kn} + \epsilon_n' \\ &= \beta_0' + \boldsymbol{\mu}'^\top \mathbf{z}_n + \boldsymbol{\nu}'^\top \mathbf{x}_n + \epsilon_n' \end{aligned} \tag{8}$$

  dove $\boldsymbol{\nu}' = [\nu_1' \ldots \nu_K']^\top$ sono i coefficienti di regressione associati ad eventuali controlli $\mathbf{x} = [x_1 \ldots x_K]^\top$ inseriti nel modello, e $\boldsymbol{\mu}' = [\mu_1' \ldots \mu_K']^\top$;

- una volta ricavate le stime per β_0', $\mu_1' \ldots \mu_4'$ e $\nu_1' \ldots \nu_K'$, si passa a calcolare le stime per β_0'', α e $\boldsymbol{\nu}'' = [\nu_1'' \ldots \nu_K'']^\top$

$$y_n = \beta_0'' + \alpha \underbrace{(\beta_0' + \boldsymbol{\mu}'^\top \mathbf{z}_n + \boldsymbol{\nu}'^\top \mathbf{x}_n)}_{d_n''} + \boldsymbol{\nu}''^\top \mathbf{x}_n + \epsilon_n'' \tag{9}$$

  con particolare interesse al coefficiente α che determina l'influenza del trattamento sulla risposta y.

[9] Nonostante il dataset contenga variabili preposte ad identificare l'area geografica dei vari Paesi, in [8] esse non vengono esplicitamente usate a parte `island_dum`, `region_n`.

[10] Essendo variabili disgiunte ed esaustive, una di loro è ridondante: pertanto, senza perdere di generalità, verrà omessa `colony0` (in [8] non viene indicata la variabile coloniale esclusa).

[11] Anche in questo caso `legor_uk` è ridondante, essendo complementare rispetto a `legor_fr` (da Fig. 3b o Tab. 3): quindi è sufficiente impiegare solo `legor_fr`.

Tab. 4: Stime OLS, casi 1–4; i p-value hanno livello di significativà al 5% (lo stesso dicasi nelle altre tabelle del report).

	caso 1			caso 2			caso 3			caso 4		
variabili	coeff.	std. err.	p-value	coeff.	std. err.	p-value	coeff.	std. err.	p-value	coeff.	std. err.	p-value
intercetta	7.2153	0.467	0.000	5.9441	1.108	0.000	5.7431	1.094	0.000	6.0241	1.055	0.000
ln_export_area	−0.1119	0.024	0.000	−0.0762	0.029	0.013	−0.0854	0.035	0.020	−0.1032	0.034	0.005
colony1	0.3621	0.475	0.450	0.8392	0.552	0.137	1.2496	0.579	0.038	1.0750	0.551	0.060
colony2	0.3135	0.470	0.508	0.7651	0.560	0.180	0.5444	0.541	0.322	0.3683	0.513	0.478
colony3	0.2108	0.531	0.693	0.4489	0.592	0.453	−0.0247	0.585	0.967	−0.1035	0.560	0.855
colony4	−0.8893	0.585	0.136	0.0823	0.770	0.915	−0.4588	0.727	0.532	−0.5897	0.678	0.391
colony5	1.6561	0.790	0.042	2.1373	0.839	0.015	1.3864	0.819	0.100	0.7023	0.843	0.411
colony6	0.8621	0.793	0.283	1.3410	0.928	0.157	1.5813	0.956	0.108	1.6933	0.901	0.070
colony7	0.7156	0.780	0.364	0.7518	0.833	0.372	0.6273	0.806	0.442	0.0429	0.783	0.957
abs_latitude				0.0158	0.017	0.349	0.0194	0.018	0.277	0.0228	0.017	0.177
longitude				0.0005	0.005	0.928	−0.0038	0.006	0.499	−0.0039	0.005	0.468
rain_min				−0.0006	0.007	0.926	0.0001	0.007	0.985	−0.0010	0.006	0.880
humid_max				0.0094	0.012	0.428	0.0090	0.012	0.440	0.0148	0.011	0.189
low_temp				−0.0189	0.028	0.510	−0.0049	0.027	0.859	−0.0151	0.026	0.567
ln_coastline_area				0.0847	0.039	0.035	0.0954	0.042	0.031	0.0825	0.040	0.046
islam							−0.0084	0.003	0.006	−0.0059	0.003	0.061
legor_fr							0.7551	0.503	0.143	0.6427	0.470	0.182
island_dum							−0.3983	0.529	0.457	−0.1504	0.516	0.773
region_n							0.3818	0.484	0.436	−0.3039	0.517	0.561
ln_avg_gold_pop										0.0110	0.017	0.531
ln_avg_oil_pop										0.0780	0.027	0.008
ln_avg_all_diamonds_pop										−0.0392	0.043	0.373

In [8] sono stati analizzati i seguenti casi:[12]

1. impiego solo delle IV, senza alcun controllo addizionale;

2. inserimento delle variabili coloniali `colony1`...`colony7` tra i controlli

3. aggiunta dei controlli geo-climatici `abs_latitude`, `longitude`, `rain_min`, `humid_max`, `low_temp`, `ln_coastline_area`.

Tab. 5 riporta le stima ottenute per il primo passaggio del Two Stage LS nei casi appena indicati; infine da Tab. 6 recuperiamo la stima di α, il coefficiente di regressione corrispondente alla variabile d'': anche in questo caso i valori ottenuti concordano con quelli in [8]. Si noti che le variabili di controllo nei casi 2 e 3 di Tab. 5-6 corrispondono a quelle dei casi 1 e 2 di Tab. 4 (stime OLS, § 6.1.1).

6.2 Stime tramite Double ML

Dalle varie combinazioni dei controlli $\mathbf{x}$ analizzati in [8] (e riprodotti in § 6.1), è stato preso come caso di studio sull'approccio Double ML quello che impiega le variabili coloniali e geo-climatiche (caso 2 in § 6.1.1, caso 3 in § 6.1.2), ossia il caso con maggior numero di controlli tra quelli comuni in Tab. 4–6.

6.2.1 Ottimizzazione degli iperparametri

Per il Bagging gli iper-parametri da impostare sono (facendo riferimento alla classe `RandomForestRegressor` fornita dalla libreria Python *scikit-learn*):

- il numero B di alberi (`n_estimators`);

- il numero massimo di elementi di un nodo foglia come criterio di arresto nella crescita dell'albero (`min_samples_leaf`)

Per le Random Forest si aggiunge ai parametri del Bagging anche il valore s di variabili considerate casualmente ad ogni split (`max_features`).[13]

Tab. 7 riporta i valori degli iper-parametri sia per i modelli $y = l(\mathbf{x})$, $d = m(\mathbf{x})$ e per i modelli relativi alle quattro variabili di controllo $z_i = h_i(\mathbf{x})$, $(i = 1 \dots 4)$: in tutti i casi è stata impiegata la strategia *Leave-One-Out* [7, p. 204], dato il numero ridotto di osservazioni nel dataset (listato in Cod. 1, in appendice).

6.2.2 Esperimenti

Seguono alcuni esperimenti ottenuti sia avendo implementato la funzione di stima per l'approccio Double ML, sia applicando le funzioni della libreria DoubleML [3]; sono stati inoltre condotti dei test Monte Carlo per analizzare meglio la variabilità su α in funzione della partizione dei dati nel cross-fitting.

[12] Tralasciamo di nuovo l'analisi senza le piccole isole ed i Paesi del Nord Africa.

[13] Il Bagging è un caso particolare di Random Forest, dove s è pari al numero di variabili a disposizione (`max_features=None` nel costruttore di `RandomForestRegressor`).

Tab. 5: Stime IV (Stage I), casi 1–3

	caso 1			caso 2			caso 3		
variabili	coeff.	std. err.	p-value	coeff.	std. err.	p-value	coeff.	std. err.	p-value
intercetta	29.1097	6.959	0.000	38.3928	8.853	0.000	39.1188	15.857	0.019
atlantic_distance_minimum	−1.3140	0.357	0.001	−1.7414	0.425	0.000	−1.3163	0.761	0.093
indian_distance_minimum	−1.0954	0.380	0.006	−1.4262	0.531	0.010	−1.0810	0.697	0.130
saharan_distance_minimum	−2.4349	0.823	0.005	−3.0011	1.049	0.007	−1.1393	1.595	0.480
red_sea_distance_minimum	−0.0019	0.710	0.998	−0.1521	0.813	0.853	−1.2214	1.820	0.507
colony1				−0.3592	2.605	0.891	−2.4563	3.436	0.480
colony2				−2.0516	2.564	0.428	−3.3958	3.461	0.334
colony3				−0.6215	2.959	0.835	−3.8935	3.669	0.296
colony4				−3.6295	3.141	0.255	−5.8491	4.490	0.201
colony5				−6.5378	4.251	0.132	−8.0328	4.667	0.094
colony6				−3.0404	4.454	0.499	−8.7829	6.012	0.153
colony7				3.9700	5.582	0.481	3.1688	6.361	0.622
abs_latitude							0.0113	0.124	0.928
longitude							−0.1088	0.131	0.414
rain_min							−0.0437	0.039	0.266
humid_max							−0.0930	0.080	0.256
low_temp							0.1687	0.167	0.320
ln_coastline_area							0.0142	0.287	0.961

Tab. 6: Stime IV (Stage II), casi 1–3

	caso 1			caso 2			caso 3		
variabili	coeff.	std. err.	p-value	coeff.	std. err.	p-value	coeff.	std. err.	p-value
intercetta	7.8114	0.187	0.0	7.7160	0.490	0.000	8.6018	1.595	0.000
d''	−0.2079	0.049	0.0	−0.2014	0.040	0.000	−0.2857	0.096	0.005
`colony1`				0.1513	0.469	0.749	0.0595	0.640	0.926
`colony2`				0.1437	0.463	0.758	−0.1358	0.676	0.842
`colony3`				0.0419	0.522	0.936	−0.5212	0.720	0.474
`colony4`				−1.2520	0.586	0.038	−1.6229	1.063	0.135
`colony5`				1.0675	0.800	0.189	0.0892	1.219	0.942
`colony6`				0.2304	0.807	0.777	−0.2794	1.155	0.810
`colony7`				0.3594	0.772	0.644	0.1694	0.854	0.844
`abs_latitude`							−0.0075	0.019	0.699
`longitude`							−0.0101	0.007	0.164
`rain_min`							−0.0119	0.008	0.155
`humid_max`							−0.0022	0.013	0.864
`low_temp`							0.0205	0.033	0.534
`ln_coastline_area`							0.0231	0.047	0.624

Tab. 7: Iper-parametri ottimali per i modelli $y = l(\mathbf{x})$, $d = m(\mathbf{x})$ e $z_i = h_i(\mathbf{x})$ ($i = 1 \dots 4$, per le variabili di controllo) con cross-validazione Leave-One-Out testando i valori `n_estimators` $\in \{10, 50, 100, 500, 1000, 5000, 10000\}$; `min_samples_leaf` $\in \{3, 5, 7\}$; `max_features` $\in \{5, 10, 12\}$.

Alg. ML	iper-parametro	l	m	h_1	h_2	h_3	h_4
Bagging	`n_estimators`	100	5000	50	500	10	50
	`min_samples_leaf`	7	3	5	3	3	3
Random Forest	`n_estimators`	100	1000	50	10	10	50
	`min_samples_leaf`	5	3	5	3	3	3
	`max_features`	10	12	12	10	12	12

Tab. 8: Stime fornite da Cod. 2 per il modello parzialmente lineare.

$l(\mathbf{x})$ $m(\mathbf{x})$	Bagging Bagging	Bagging Rand. Forest	Rand. Forest Bagging	Rand. Forest Rand. Forest
α	-0.0834	-0.0840	-0.0888	-0.0892
std. err.	0.0302	0.0302	0.0294	0.0293
p-value	0.0081	0.0076	0.0039	0.0037
RMSE y	0.7606	0.7606	0.7487	0.7487
RMSE d	3.3412	3.3390	3.3412	3.3390

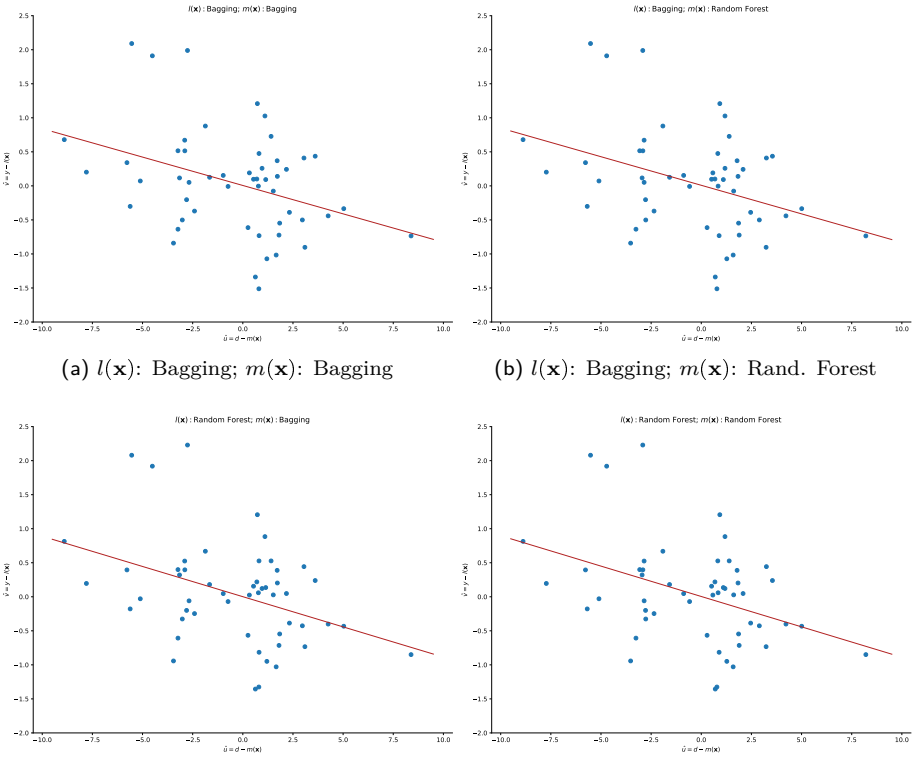
Implementazione dell'approccio Double ML Attraverso la funzione `dmlFit`[14] riportata nel listato Cod 2 (in appendice) sono stati valutati i risultati per le quattro possibili combinazioni dei modelli di Bagging e Random Forest su $y = l(\mathbf{x})$ e $d = m(\mathbf{x})$, senza l'impiego delle variabili strumentali: avendo imposto gli iper-parametri di Tab. 7, ricaviamo da Tab. 8 che la combinazione migliore è fornita dall'impiego di Random Forest sia per $y = l(\mathbf{x})$ che $d = m(\mathbf{x})$ (RMSE y e RMSE d minimi), a cui corrisponde $\alpha = -0.0892$ con p-value pari a 0.0037 ed errore standard 0.0293; in ogni caso, anche i valori forniti dalle altre combinazioni di modelli sono paragonabili. Fig. 8 mostra la retta di regressione sui residui $(\hat{u}, \hat{v})$ per le quattro combinazioni in Tab. 8.[15]

Impiego della libreria DoubleML In alternativa all'implementazione naïve di Cod. 2, è possibile utilizzare la libreria DoubleML [2]: il listato in Cod. 3 mostra l'uso della funzione `DoubleMLPLR` in relazione all'impiego di Random Forest sia per $l(\mathbf{x})$ che $m(\mathbf{x})$. Tab. 9 riporta i valori ottenuti che, pur in linea con quelli

[14] la funzione `dmlFit` è ispirata al notebook "Double/Debiased Machine Learning for the Partially Linear Regression Model" [5, cap. 10]. Rispetto ad Eq. (5), nel notebook la regressione per α sulle N coppie (u_n, v_n) prevede anche il termine noto (cioè $v_n = \tau_0 + \alpha u_n$): l'implementazione in Cod 2 ha mantenuto l'approccio del notebook, in quanto i valori medi $\bar{u}$ e $\bar{v}$ riscontrati sono comunque circa zero e pertanto trascurabili nella formula di regressione per α in presenza di termine noto:

$$\alpha = \frac{\sum_n (u_n - \bar{u})(v_n - \bar{v})}{\sum_n (u_n - \bar{u})^2} \approx \frac{\sum_n u_n v_n}{\sum_n u_n^2} \quad \text{per } \bar{u}, \bar{v} \approx 0$$

[15] In tutti i casi i dati sono stati partizionati in $K = 4$ insiemi disgiunti per il cross-fitting, così da dividere equamente le osservazioni dei $52 = 4 \cdot 13$ Stati africani (e per abbassare i tempi di calcolo rispetto a $K = 13$).



(a) $l(\mathbf{x})$: Bagging; $m(\mathbf{x})$: Bagging

(b) $l(\mathbf{x})$: Bagging; $m(\mathbf{x})$: Rand. Forest

(c) $l(\mathbf{x})$: Rand. Forest; $m(\mathbf{x})$: Bagging

(d) $l(\mathbf{x})$: Rand. Forest; $m(\mathbf{x})$: Rand. Forest

Fig. 8: Rette di regressione $\hat{v} = \alpha\hat{u}$ per i vari casi di modelli ML in Tab. 8.

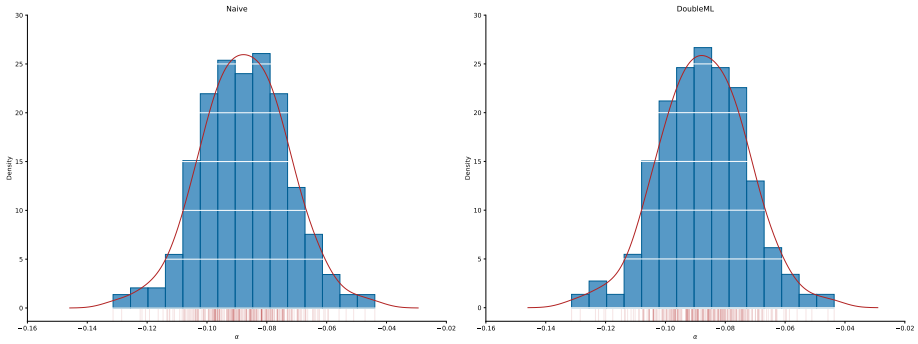
Tab. 9: Valori ottenuti impiegando la libreria DoubleML (Cod. 3).

α	std. err.	p-value	RMSE y	RMSE d
-0.1031	0.0349	0.0032	0.8012	3.0574

del caso analogo in Tab. 8, differiscono da essi. L'ipotesi è che la causa di tali differenze sia imputabile a qualche variante implementativa, in particolare nella modalità con cui i dati vengono partizionati per il cross-fitting.

Confronto Monte Carlo In ogni caso i risultati in Tab. 8 sono stati ottenuti su una particolare partizione del dataset: rimescolando la partizione in maniera casuale, le stime fornite cambiano; per cui con approccio Monte Carlo variando casualmente la partizione per $T = 250$ volte,[16] otteniamo le distribuzioni campionare di α visibili in Fig. 9 sia per l'implementazione naïve di Cod. 2 e sia ricorrendo a DoubleML (Cod. 3). Notiamo che le due distribuzioni sono molto simili, mostrando in particolare valor medio $\bar{\alpha}$ e deviazione standard σ_α quasi uguali: $\bar{\alpha} \approx -0.087$ e $\sigma_\alpha \approx 0.015$ in entrambi i casi.

[16] Il valore di T è stato limitato da motivi di potenza computazionale.



(a) $\bar{\alpha} = -0.0874$, $\sigma_\alpha = 0.0145$ (b) $\bar{\alpha} = -0.0873$, $\sigma_\alpha = 0.0146$

Fig. 9: Analisi Monte Carlo per la stima di α su $T = 1000$ istanze: in (a), la distribuzione campionaria fornita dall'implementazione naïve di Cod 2; in (b) quella impiegando la libreria DoubleML (Cod. 3).

Tab. 10: Valori ottenuti tramite la libreria DoubleML applicando anche le variabili di controllo (Cod. 4).

α	std. err.	p-value	RMSE y	RMSE d
-0.1643	0.0862	0.0567	0.8012	3.0574

Impiego delle variabili strumentali Come ultimo esperimento, la libreria DoubleML è stata sfruttata per aggiungere al modello causale anche le variabili strumentali proposte in [8]: a tale scopo si è ricorsi alla funzione `DoubleMLPLIV` (listato in Cod. 4), che riceve anche un modello ML per le variabili strumentali: con riferimento a Tab. 7, il modello scelto è una Random Forest con parametri `n_estimators` = 500, `max_features` = 12, `min_sample_leaf` = 3.[17] Anche in questo caso si ha un valore $\alpha < 0$ statisticamente significativo (Tab. 10), ma con un lieve maggior impatto ($\alpha = -0.1643$) rispetto a Tab. 9 ($\alpha = -0.1031$).

Con $T = 250$ esperimenti Monte Carlo, analogamente a quanto fatto sui modelli senza variabili strumentali, otteniamo che la distribuzione campionaria trovata per α ha valor medio $\bar{\alpha} = -0.2590$ e deviazione standard $\sigma_\alpha = 0.0879$ (Fig. 10): la distribuzione ottenuta è piuttosto differente rispetto a quelle di Fig. 9, mostrando una certa asimmetria e maggiore dispersione; in ogni caso il valor medio per α è sensibilmente maggiore rispetto a quello ottenuto senza variabili strumentali (ossia $\bar{\alpha} \approx -0.087$).

7 Conclusioni

L'elaborato si è basato sul lavoro svolto in [8], dove viene analizzato attraverso delle variabili strumentali l'impatto a lungo termine che la tratta degli schiavi

[17] Il razionale dietro alla scelta fatta è che, per le variabili strumentali z_i, `min_sample_leaf` = 3 è il valore più ricorrente sia nel caso di Bagging che Random Forest; inoltre nelle Random Forest `max_features` = 12 è, oltre ad essere anch'esso il valore più gettonato, prossimo a 13 (caso in cui si ha Bagging) ed infine è stato scelto il valore massimo di alberi `n_estimators` = 500 perché per Bagging e Random Forest un incremento degli stimatori non crea overfitting.

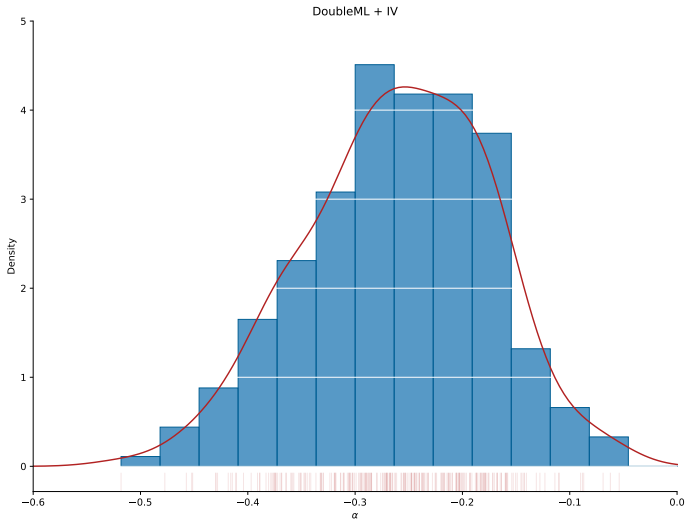


Fig. 10: Distribuzione campionaria su $T = 250$ esperimento Monte Carlo per la stima di α impiegando anche le variabili strumentali (funzione `DoubleMLPLIV` in DoubleML, Cod. 4): valor medio $\bar{\alpha} = -0.2590$, deviazione standard $\sigma_\alpha = 0.0879$.

(il trattamento) ha avuto sull'economia odierna degli Stati africani (risposta caratterizzata dal PIL pro capite degli Stati).

Riprendendo il dataset annesso all'articolo, in § 6.1 sono stati prima replicati alcuni risultati sull'impatto del trattamento mostrati nel paper sia per quanto riguarda le stime OLS (senza IV, in Tab. 4) che quelle ottenute ricavate tramite Two-Stage LS con le IV (Tab. 5–6).

Successivamente in § 6.2 è stata condotta l'analisi causale con approccio Double ML, applicata ad una delle tante combinazioni di controlli valutate in [8] (passato coloniale e caratteristiche geo-climatiche dei vari Paesi): dal confronto in Tab. 11 tra i valori in [8] con quelli ottenuti tramite Double ML concludiamo che le stime ricavate attraverso l'impiego di tecniche Machine Learning sono in linea con quelle ottenute da [8] in maniera classica.

Il presente lavoro potrebbe essere esteso impiegando un maggior numero di controlli e aumentando il numero T degli esperimenti Monte Carlo, allo scopo di ottenere risultati più accurati: le restrizioni nell'elaborato sono dipese dai tempi di calcolo per la limitata potenza computazionale disponibile.

Tab. 11: Stime di α a confronto con i valori riportati in [8], avendo impiegato come controlli il passato coloniale e le caratteristiche geo-climatiche dei vari Paesi: i valori per [8] corrispondono al caso 2 in Tab. 4 (No IV) ed al caso 3 in Tab. 6 (IV); i valori per `dmlFit` e DoubleML sono presi dai valor medi dei test Monte Carlo (Fig. 9–10)

	[8] (Nunn)	`dmlFit`	DoubleML
No IV	−0.0762	−0.0874	−0.0873
IV	−0.2857		−0.2590

A Script principali

Il codice è stato sviluppato su Python 3.9.19.

Listato 1: Codice per gli iper-parametri ottimali.

```
import pandas as pd
import numpy as np
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import LeaveOneOut
from sklearn.model_selection import GridSearchCV
#-----------------------------------------------------------------------
def loocv_Bagging(outcome, x, gridParams, rstate):
    grid_search =
        GridSearchCV(estimator=RandomForestRegressor(random_state=rstate,
                                                     max_features=None),
                                 param_grid=gridParams,
                                 cv=LeaveOneOut(),
                                 scoring = 'neg_root_mean_squared_error',
                                 n_jobs=-1)
    grid_search.fit(x, outcome)
    return grid_search.best_params_

def loocv_RandomForest(outcome, x, gridParams, rstate):
    grid_search =
        GridSearchCV(estimator=RandomForestRegressor(random_state=rstate),
                             param_grid=param_grid_RF,
                             cv=LeaveOneOut(),
                             scoring = 'neg_root_mean_squared_error',
                             n_jobs=-1)
    grid_search.fit(x, outcome)
    return grid_search.best_params_
#-----------------------------------------------------------------------

rstate = 313
np.random.seed(rstate)

param_grid_BG = {
    'n_estimators': [10, 50, 100, 500, 1000, 5000, 10000],
    'min_samples_leaf': [3, 5, 7]
}

param_grid_RF = {
    'n_estimators': [10, 50, 100, 500, 1000, 5000, 10000],
    'min_samples_leaf': [3, 5, 7],
    'max_features': [5, 10, 12]
}

ctrlDML = ['colony1', 'colony2', 'colony3', 'colony4', 'colony5',
     'colony6', 'colony7',
        'abs_latitude', 'longitude', 'rain_min', 'humid_max',
            'low_temp', 'ln_coastline_area']
x = df[ctrlDML]
d = df.ln_export_area
```

```
y = df.ln_maddison_pcgdp2000
iv1 = df.atlantic_distance_minimum
iv2 = df.indian_distance_minimum
iv3 = df.saharan_distance_minimum
iv4 = df.red_sea_distance_minimum

BGy = loocv_Bagging(y, x, param_grid_BG, rstate)
print("BG y: " + str(BGy))
BGd = loocv_Bagging(d, x, param_grid_BG, rstate)
print("BG d: " + str(BGd))

BGiv1 = loocv_Bagging(iv1, x, param_grid_BG, rstate)
print("BG iv1: " + str(BGiv1))
BGiv2 = loocv_Bagging(iv2, x, param_grid_BG, rstate)
print("BG iv2: " + str(BGiv2))
BGiv3 = loocv_Bagging(iv3, x, param_grid_BG, rstate)
print("BG iv3: " + str(BGiv3))
BGiv4 = loocv_Bagging(iv4, x, param_grid_BG, rstate)
print("BG iv4: " + str(BGiv4))

RFy = loocv_RandomForest(y, x, param_grid_RF, rstate)
print("RF y: " + str(RFy))
RFd = loocv_RandomForest(d, x, param_grid_RF, rstate)
print("RF d: " + str(RFd))

RFiv1 = loocv_RandomForest(iv1, x, param_grid_RF, rstate)
print("RF iv1: " + str(RFiv1))
RFiv2 = loocv_RandomForest(iv2, x, param_grid_RF, rstate)
print("RF iv2: " + str(RFiv2))
RFiv3 = loocv_RandomForest(iv3, x, param_grid_RF, rstate)
print("RF iv3: " + str(RFiv3))
RFiv4 = loocv_RandomForest(iv4, x, param_grid_RF, rstate)
print("RF iv4: " + str(RFiv4))
```

Listato 2: Codice per Double ML applicato a Eq. (5).

```
# original code: Double/Debiased Machine Learning for the Partially
    Linear Regression Model
# url: https://colab.research.google.com/github/CausalAIBook/
#             MetricsMLNotebooks/blob/main/PM4/
#             python_debiased_ml_for_partially_linear_model_growth.ipynb

from sklearn.model_selection import cross_val_predict, KFold
import statsmodels.formula.api as smf

def dmlFit(X, D, y, modely, modeld, nfolds, rState=3131):
    '''DML for the Partially Linear Model setting with cross-fitting'''

    # shuffled k-folds
    cv = KFold(n_splits=nfolds, shuffle=True, random_state=rState)
    # out-of-fold predictions for y
    yhat = cross_val_predict(modely, X, y, cv=cv, n_jobs=-1)
    # out-of-fold predictions for D
```

```
    Dhat = cross_val_predict(modeld, X, D, cv=cv, n_jobs=-1)

    # calculate outcome and treatment residuals
    resy = y - yhat
    resD = D - Dhat

    # final stage ols
    dml_data = pd.concat([pd.Series(resy, name='resy'),
                          pd.Series(resD, name='resD')], axis=1)
    # regular ols
    ols_mod = smf.ols(formula='resy ~ 1 + resD', data=dml_data).fit()

    intercept = ols_mod.params[0]
    point = ols_mod.params[1]
    stderr = ols_mod.bse[1]
    pvalue = ols_mod.pvalues[1]
    return (intercept, point, stderr, pvalue, resy, resD)
```

Listato 3: Codice per l'impiego della libreria DoubleML.

```
import doubleml as dml

rstate = 3131
np.random.seed(rstate)
kfolds = 4

data = x.copy()
data['y'] = y
data['d'] = d
dml_data = dml.DoubleMLData(data, y_col='y', d_cols='d')

modely = RandomForestRegressor(n_estimators= 100, min_samples_leaf=5,
    max_features=10, n_jobs=-1, random_state=313)
modeld = RandomForestRegressor(n_estimators=1000, min_samples_leaf=3,
    max_features=12, n_jobs=-1, random_state=313)
dml_plr_obj = dml.DoubleMLPLR(dml_data, ml_l=modely, ml_m=modeld,
    n_folds=kfolds)
resDoubleML = dml_plr_obj.fit()
print(resDoubleML)
```

Listato 4: Codice per le variabili strumentali con la libreria DoubleML.

```
np.random.seed(rstate)

dataIV = x.copy()
dataIV['y'] = y
dataIV['d'] = d
dataIV['z1'] = df.atlantic_distance_minimum
dataIV['z2'] = df.indian_distance_minimum
dataIV['z3'] = df.saharan_distance_minimum
dataIV['z4'] = df.red_sea_distance_minimum
dml_dataIV = dml.DoubleMLData(dataIV, y_col='y', d_cols='d',
    z_cols=['z1','z2','z3','z4'])
```

```
modely = RandomForestRegressor(n_estimators= 100, min_samples_leaf=5,
    max_features=10, n_jobs=-1, random_state=313)
modeld = RandomForestRegressor(n_estimators=1000, min_samples_leaf=3,
    max_features=12, n_jobs=-1, random_state=313)
modelz = RandomForestRegressor(n_estimators= 500, min_samples_leaf=3,
    max_features=12, n_jobs=-1, random_state=313)
dml_pliv_obj = dml.DoubleMLPLIV(dml_dataIV, ml_l=modely, ml_m=modelz,
    ml_r=modeld, n_folds=kfolds)
resDoubleML_IV = dml_pliv_obj.fit()
print(resDoubleML_IV)
```

Riferimenti bibliografici

[1] Achim Ahrens, Christian B. Hansen, Mark E. Schaffer, and Thomas Wiemann. ddml: Double/debiased machine learning in stata. *The Stata Journal*, 24(1):3–45, 2024.

[2] Philipp Bach, Victor Chernozhukov, Malte S. Kurz, and Martin Spindler. DoubleML – An object-oriented implementation of double machine learning in Python. *Journal of Machine Learning Research*, 23(53):1–6, 2022.

[3] Philipp Bach, Oliver Schacht, Victor Chernozhukov, Sven Klaassen, and Martin Spindler. Hyperparameter tuning for causal inference with double machine learning: A simulation study. In Francesco Locatello and Vanessa Didelez, editors, *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pages 1065–1117. PMLR, 01–03 April 2024.

[4] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018.

[5] Victor Chernozhukov, Christian Hansen, Nathan Kallus, Martin Spindler, and Vasilis Syrgkanis. *Applied Causal Inference Powered by ML and AI*. 2024. CausalML-book.org; arXiv:2403.02467.

[6] John D. Fage. *An Introduction to the History of West Africa*. Cambridge University Press,, Cambridge, 3rd ed edition, 1962.

[7] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, and Jonathan Taylor. *An Introduction to Statistical Learning with Applications in Python*. Springer Texts in Statistics. Springer, 2023.

[8] Nathan Nunn. The Long-term Effects of Africa's Slave Trades. *The Quarterly Journal of Economics*, 123(1):139–176, 2008.

[9] W. Rodney. *A History of the Upper Guinea Coast, 1545-1800*. Oxford studies in African affairs. Clarendon Press, 1970.