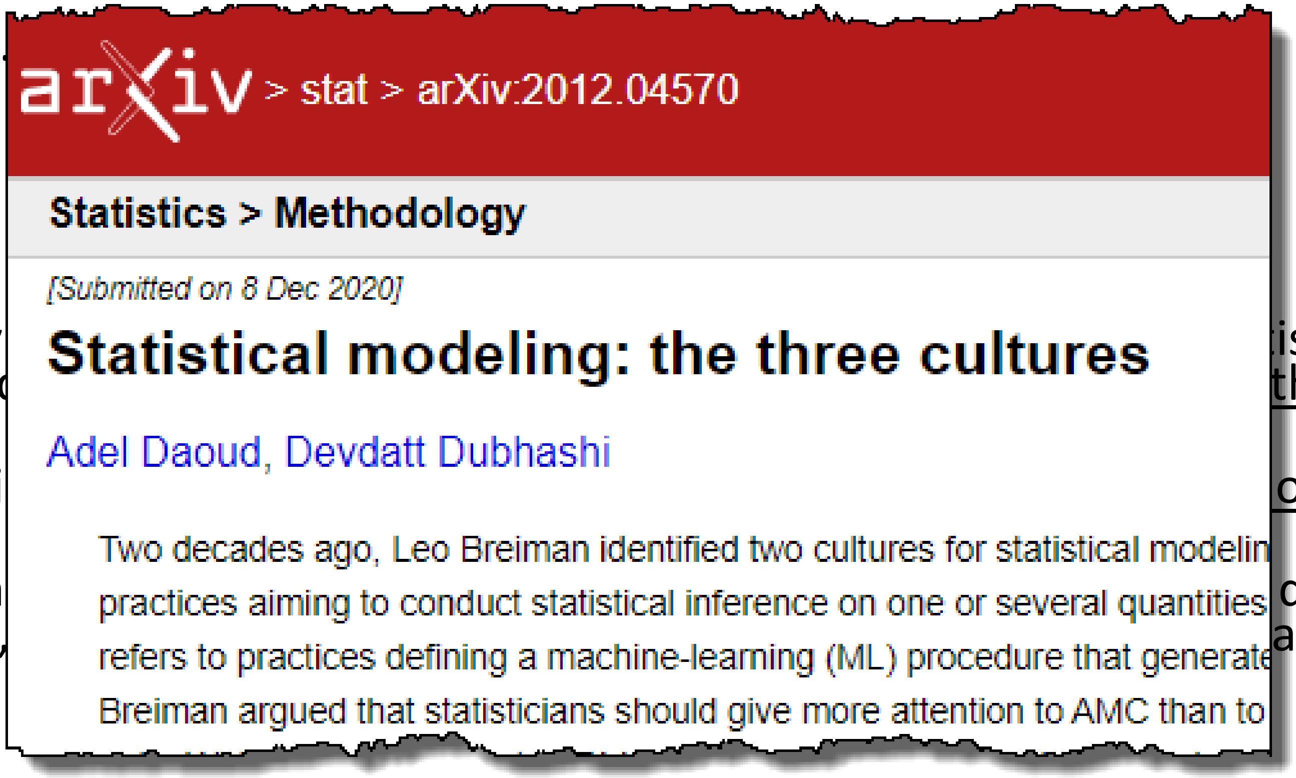
On the hybrid modeling culture: using AI methods to boost global health and poverty research

Adel Daoud, PhD,

1. Associate Professor, Institute for Analytical Sociology, Linköping University,
2. Affiliated Associate Professor in Data Science and Artificial Intelligence, Dep. of Computer Science, Chalmers Technical University, Sweden,
3. Academy Research Fellow, Royal Swedish Academy of Letters, History, and Antiquities, Sweden,
4. Member of the Swedish Young Academy,
5. Visiting scientist, Harvard School of Public Health, Harvard University, U.S.
6. Principal investigator, the *AI and Global Development Lab* (global-lab.ai).

Structure o

- Aim
  - Scholars tend to v … istical modeling, belonging to two … the data modeling culture.
  - Via extensive mixi … odeling culture has evolved.
  - In this talk, I will a … default mode of scientific thinking, … alth and social sciences.


arXiv > stat > arXiv:2012.04570

Statistics > Methodology

[Submitted on 8 Dec 2020]

Statistical modeling: the three cultures

Adel Daoud, Devdatt Dubhashi

Two decades ago, Leo Breiman identified two cultures for statistical modelin
practices aiming to conduct statistical inference on one or several quantities
refers to practices defining a machine-learning (ML) procedure that generate
Breiman argued that statisticians should give more attention to AMC than to

- Outline
  1. Thinking predictively, inferentially, and causally.
  2. The rise of a hybrid modeling culture: using prediction for (causal) inference.
  3. A framework for using ML on three archetypical scientific problems.
     1. Focus on effect modification, also known as conditional average treatment effect (CATE).
  4. Summary.

Thinking predictively, inferentially, and causally

Inference, the aim of the data modeling culture

- Leo Breiman, 2001, “Statistical modeling: the two cultures”, *Statistical Science*, identified the two first cultures:

Statistics starts with data. Think of the data as being generated by a black box in which a vector of input variables **x** (independent variables) go in one side, and on the other side the response variables **y** come out. Inside the black box, nature functions to associate the predictor variables with the response variables, so the picture is like this:

There are two goals in analyzing the data:

Prediction. To be able to predict what the responses are going to be to future input variables;

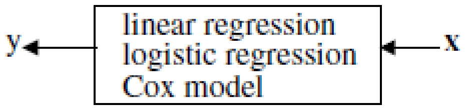
Information. To extract some information about how nature is associating the response variables to the input variables.

The Data Modeling Culture

The analysis in this culture starts with assuming a stochastic data model for the inside of the black box. For example, a common data model is that data are generated by independent draws from

response variables = f(predictor variables, random noise, parameters)

The values of the parameters are estimated from the data and the model then used for information and/or prediction. Thus the black box is filled in like this:

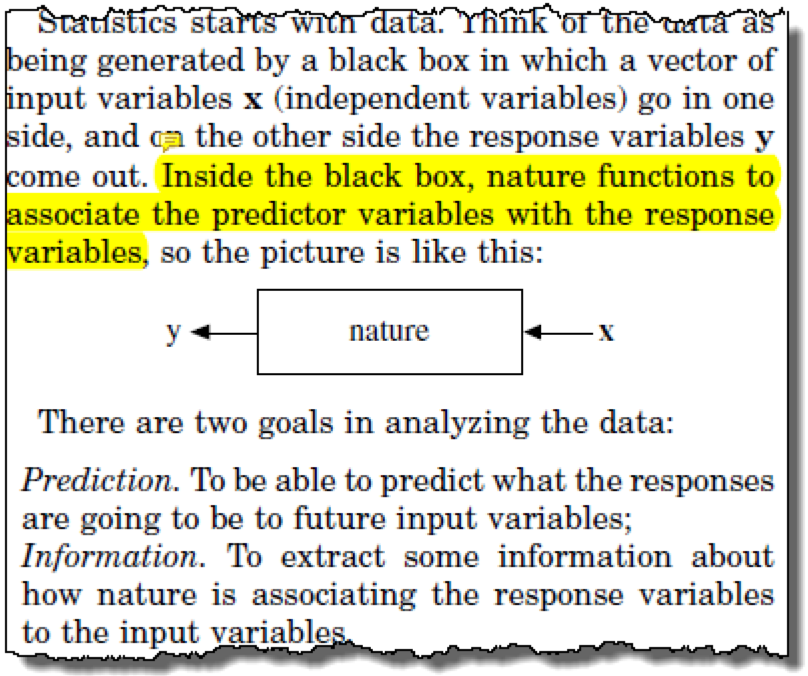


Model validation. Yes–no using goodness-of-fit tests and residual examination.

Estimated culture population. 98% of all statisticians.

Prediction, the aim of the algorithmic modeling culture

- Leo Breiman, 2001, “Statistical modeling: the two cultures”, *Statistical Science*, provides a good starting point.
  - *Algorithms* usually refers to machine learning methods.

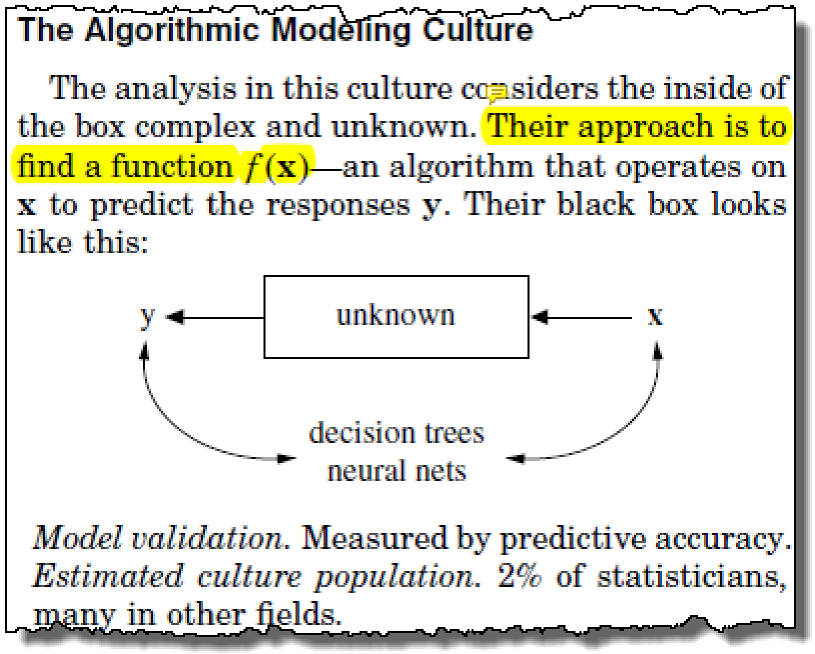


Statistics starts with data. Think of the data as being generated by a black box in which a vector of input variables **x** (independent variables) go in one side, and on the other side the response variables **y** come out. Inside the black box, nature functions to associate the predictor variables with the response variables, so the picture is like this:

y
nature
x

There are two goals in analyzing the data:

Prediction. To be able to predict what the responses are going to be to future input variables;
Information. To extract some information about how nature is associating the response variables to the input variables.



The Algorithmic Modeling Culture

The analysis in this culture considers the inside of the box complex and unknown. Their approach is to find a function $f(\mathbf{x})$—an algorithm that operates on **x** to predict the responses **y**. Their black box looks like this:

y
unknown
x
decision trees
neural nets

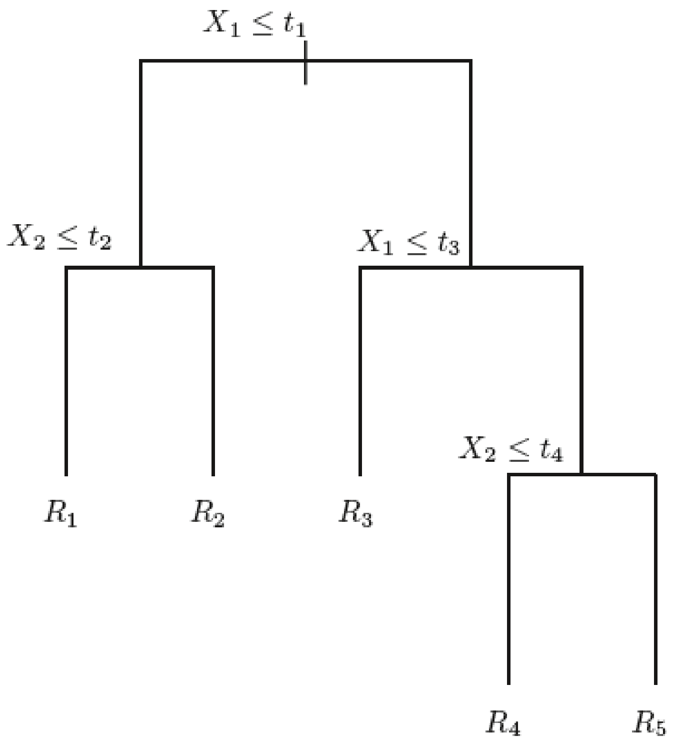
Model validation. Measured by predictive accuracy.
Estimated culture population. 2% of statisticians, many in other fields.

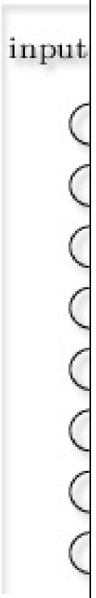
The difference between prediction and inference corresponds in practice to the difference between $\hat{Y}$ and $\hat{\beta}$.

- In <u>inference</u>, we estimate β, given a pre-specified function, *f*, for example a linear model:
  - $Y = \hat{\beta}_0 + \hat{\beta}_1 T + \hat{\beta}_2 X + e$
  - *But the goal is not to predict* $\hat{Y}$ for new data!
- In <u>prediction</u>, we let a supervised ML algorithm identify the relationship between *Y* and *X* by estimating *f*, for a specific function class, *F*, to predict $\hat{Y}$ for new data.
  - *But the goal is not to produce unbiased estimate(s) of* $\hat{\beta}$!
    - *Bias-variance tradeoff: can deliberately bias the model to reduce variance*
  - $\hat{Y} = f(X)$
  - In ML, $f(\quad)$, can be many things…

In ML, $f(X)$, can be m

- Trees, Neural nets, ensembles (a
- **The goal of all of them (supervis data.**





6 Available Models

The models below are available in `train`. The code behind these protocols can be obtained using the function `getModelInfo` or by going to the github repository.

Show 238 entries

Search:

Model	*method* Value	Type	Libraries	Tuning Parameters
Adaptive-Network-Based Fuzzy Inference System	ANFIS	Regression	frbs	num.labels, max.iter
Bayesian Regularized Neural Networks	brnn	Regression	brnn	neurons
Bayesian Ridge Regression	bridge	Regression	monomvn	None
Bayesian Ridge Regression (Model Averaged)	blassoAveraged	Regression	monomvn	None
Cubist	cubist	Regression	Cubist	committees, neighbors

In sum, ML infuses a shift from $\hat{\beta}$-problems to $\hat{Y}$-problems.

	Data-modeling culture (DMC)	Algorithmic-modeling culture (AMC)
Exemplifying question	What is the causal relationship between food supply and famines?	How well can famines be predicted from available data?
Goal	Estimating unbaised parameters for causal estimation, to populate the magnitudes of the edges of a DAG.	To develop and train an algorithm f for accurate prediction.
A key assumption	Assuming a DAG, a stipulated and interpretable statistical model such as $y_i = c_0 + \beta w_i + e_i$ produces unbiased estimates of the true causal quantity β.	The algorithm f can produce accurate predictions of Y from data source, D.
Limitation	Although the parametric model is interpretable, its statistical structure may be a poor representation of the causal system.	Although f produces accurate predictions, the model is a black-box restricting causal interpretations.
Quantity of interest	$\hat{\beta}$	$\hat{Y}$

Notes: a) in the equation $y_i = c_0 + \beta w_i + e_i$, the outcome is y_i and the treatment is w_i. The variable c_0 is the intercept and e_i is the residual.

So, what statistical culture should we follow?

Contents lists available at ScienceDirect

SSM - Population Health

journal homepage: www.elsevier.com/locate/ssmph

ELSEVIER

A scoping review on the use of machine learning in research on social determinants of health: Trends and research prospects

Shiho Kino [a,b,*], Yu-Tien Hsu [a], Koichiro Shiba [c], Yung-Shin Chien [a], Carol Mita [d], Ichiro Kawachi [a], Adel Daoud [e,f,g,h]

(c.f., with Kino et al 2021)

If the main goal of the sciences is to *explain* reality (nature, society), then there is one way forward…

Synthesizing the two cultures into a third: *the hybrid modeling culture (HMC)*.

Melting the distinction between $\hat{Y}$ and $\hat{\beta}$.

A hybrid modeling culture: a framework that unifies prediction and inference

- The ultimate goal of the hybrid culture, as in the data-modeling culture, is to advance applied science in explaining phenomena of interest.
- HMC synthesizes two elements, one from DMC and one from AMC:
    1. In kinship with DMC, the HMC stipulate a causal structure on "nature" to test a model (hence, theory-driven and hypothesis-testing oriented);
    2. In kinship with AMC, the HMC tends to mobilize the modeling power of machine learning for estimation.

    ➔ The hybrid culture melts the distinction between $\hat{Y}$ and $\hat{\beta}$. (Now, $\hat{Y}$-problems encompasses all types of problems. E.g., the $\hat{\tau}$-problem replaces the $\hat{\beta}$-problem in causal inference.)

Central practices of the hybrid-modeling culture

Table 2:Central practices of the hybrid-modeling culture (HMC)

	ML for causal inference	**ML for data acquisition**	**ML for theory prediction**
Exemplifying question	What is the causal relationship between food supply and famines?	Can food availability be measured from satellite images?	How well does the Malthusian theory of famines predict new famines? How does it compare to a Senian theory?
Goal	Imputing potential outcomes for causal estimation, to populate the magnitudes of the edges of a DAG.	Producing new indicators from digital sources, D, to populate the nodes of a DAG.	Comparing the predictive power of two or more theories' DAGs, $\hat{Y}_{G_1}, \hat{Y}_{G_2}, \dots, \hat{Y}_{G_k}$, for new realizations of an outcome, Y.
A key assumption	The algorithm f produces unbiased estimates of the true causal quantity τ, assuming a DAG.	The algorithm f can measure the true quantity of the variable of interest (X, W, Y) from a digital source, D.	The algorithm f is an appropriate representation of G_k to predict, Y.
Quantity of interest	$\hat{\tau} = \hat{Y}_i^1 - \hat{Y}_i^0$	$\hat{X}, \widehat{W}, \hat{Y}$	$\epsilon_{\hat{Y}_{G_k}} \approx Y - \hat{Y}_{G_k}$ or $\epsilon_{\hat{\tau}_{G_k}} \approx \tau - \hat{\tau}_{G_k}$

$$\hat{\tau} = \widehat{\mathrm{Y}}(1) - \widehat{\mathrm{Y}}(0)$$

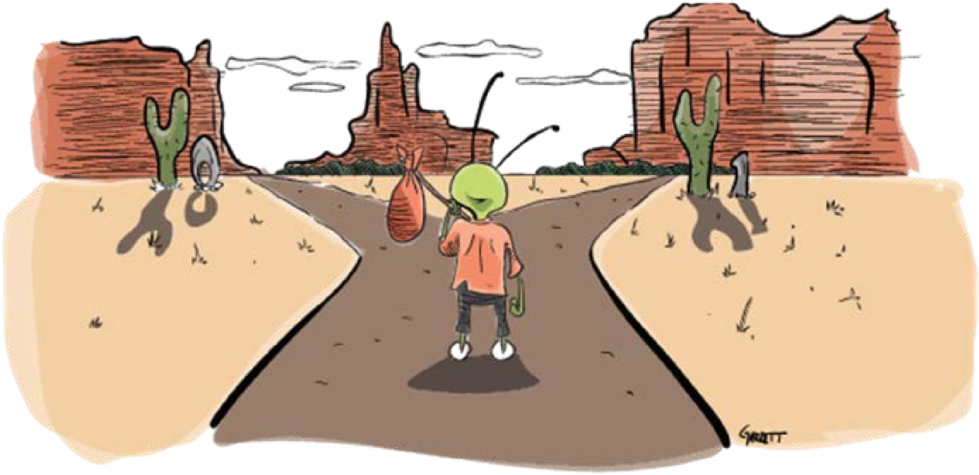
(Causal inference)

The fundamental problem of causal inference

- The potential outcome framework. We cannot observe a person's, *i*, two outcomes (Y_i): with an exposure (X=1) and without it (X=0). If we could then, we could calculate individual-level treatment (ITE) effect directly:

$$\tau_i = Y_i(1) - Y_i(0)$$

- Define potential outcomes as a missing data problem.



	X	Y(1)	Y(0)	τ
Jane	1	20	?	?
John	1	30	?	?
Joe	0	?	25	?
Jan	0	?	22	?

ML imputes potential outcomes

- Assuming ignorability
- Estimate ITE for all children, with and without IMF programs:

$$Y_i(1) = \hat{m}_1(x_i) \text{ and } Y_i(0) = \hat{m}_0(x_i).$$

$$\hat{m}_1(x) \text{ trained on treated and } \hat{m}_0(x) \text{ for control.}$$

- By imputing potential outcomes, we get to see the other previously hidden half of the data

	X	Y(1)	Y(0)	τ
Jane	1	20	?	?
John	1	30	?	?
Joe	0	?	25	?
Jan	0	?	22	?

ML imputes potential outcomes

- Assuming ignorability
- Estimate ITE for all children, with and without IMF programs:

$$Y_i(1) = \hat{m}_1(x_i) \text{ and } Y_i(0) = \hat{m}_0(x_i).$$

$$\hat{m}_1(x) \text{ trained on treated and } \hat{m}_0(x) \text{ for control.}$$

- By imputing potential outcomes, we get to see the other previously hidden half of the data

	X	Y(1)	Y(0)	τ
Jane	1	20	25	-5
John	1	30	32	-2
Joe	0	23	25	-2
Jan	0	24	22	+2

Natural disasters and mental health in social epidemiology

PCN Psychiatry and

EPIDEMIOLOGY

Estimating the Impact of Sustained Social Participation on Depressive Symptoms in Older Adults

Koichiro Shiba,[a,b] Jacqueline M. Torres,[c] Adel Daoud,[d,e,f] Kosuke Inoue,[g] Satoru Kanamori,[h] Taishi Tsuji,[i] Masamitsu Kamada,[j] Katsunori Kondo,[k,l] and Ichiro Kawachi[b]

Koichiro Shiba, Adel Daoud, Hiroyuki Hikichi, Aki Yazawa, Jun Aida,
Katsunori Kondo[8,9], Ichiro Kawachi[2]

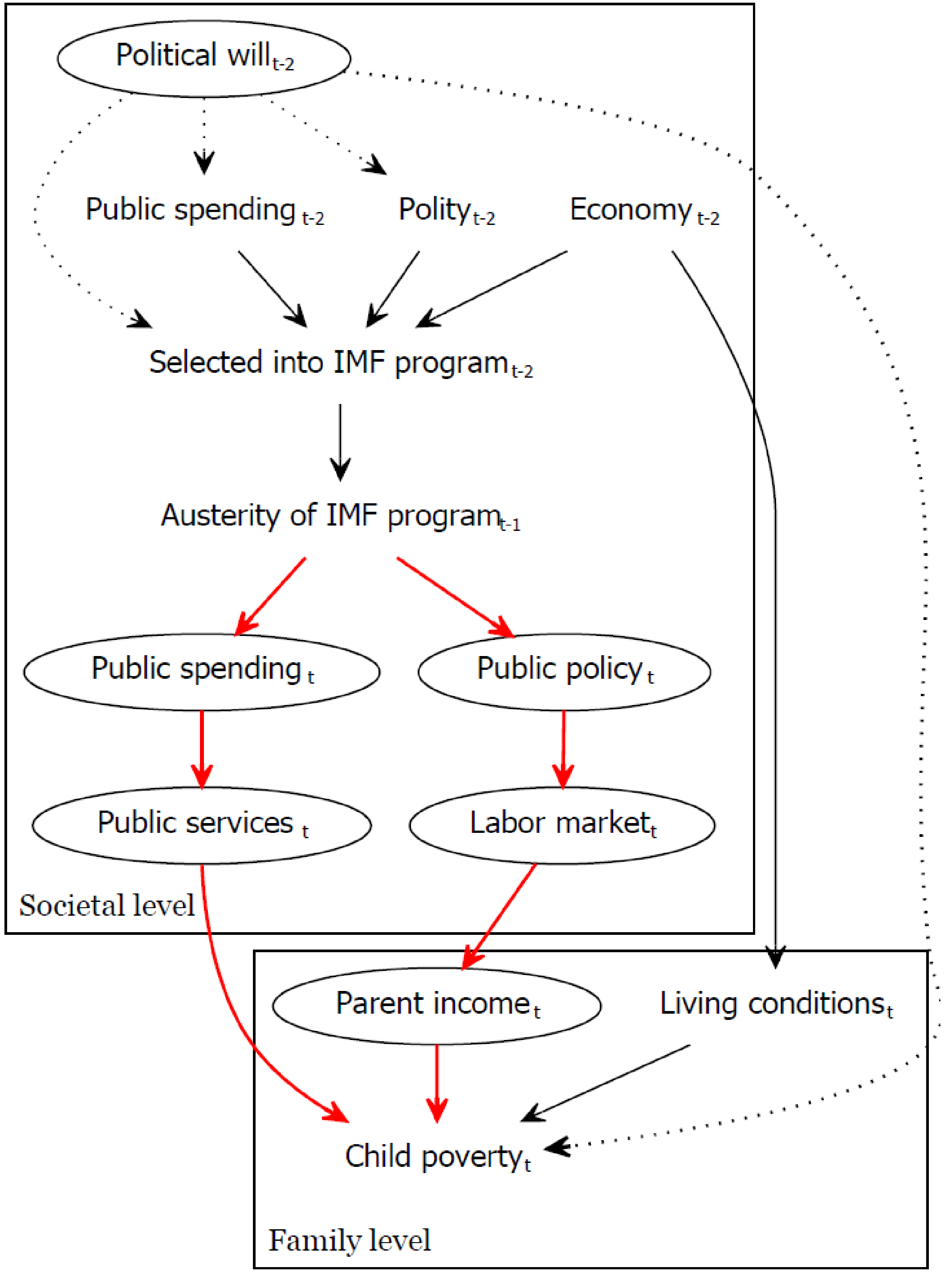
Katsunori Kondo, MD, PhD[9,10] and Ichiro Kawachi, MD, PhD[2]


PLOS ONE
Estimating Treatmen
on Child Poverty with
d Programs
AUTHORS
Adel Daoud, Fredrik Johansson
SUBMITTED ON
February 06, 2019
LAST EDITED
February 13, 2019
SOC ARXIV SocArXiv Papers
ELSEVIER
World Development
Volume 157, September 2022, 105924
IMF fairness: Calibrating the policies of the International Monetary Fund based on distributive justice
Adel Daoud a, b, c, Anders Herlitz d, e, S.V. Subramanian c, f
PNAS
Impact o
on child
Adel Daoud a,1, Elia
and Lawrence P. Ki
aCentre for Business Res
Sociology, University of
dDepartment of Sociolog
Hamilton 3240, New Zealand
Edited by Arjumand Siddiqi, Unive
arXiv > econ > arXiv:2012.14941
Help | Advanced
Econ
[Subm
The
des
Adel
arXiv > cs > arXiv:2202.09391
Help | Advanced
Computer Science > Artificial Intelligence
[Submitted
Coun
in the
Sourabh
erty
Personalized Public Policy Analysis In Social Sciences Using Causal-Graphical Normalizing Flows
Sourabh Balgi 1, Jose M. Peña 1, Adel Daoud 2,3
1 The Department of Computer and Information Science (IDA), Linköping University, Sweden
2 The Institute for Analytical Sociology (IAS), Linköping University, Sweden
3 The Department of Computer Science and Engineering, Chalmers University of Technology, Gothenburg, Sweden
sourabh.balgi@liu.se, jose.m.pena@liu.se, adel.daoud@liu.se

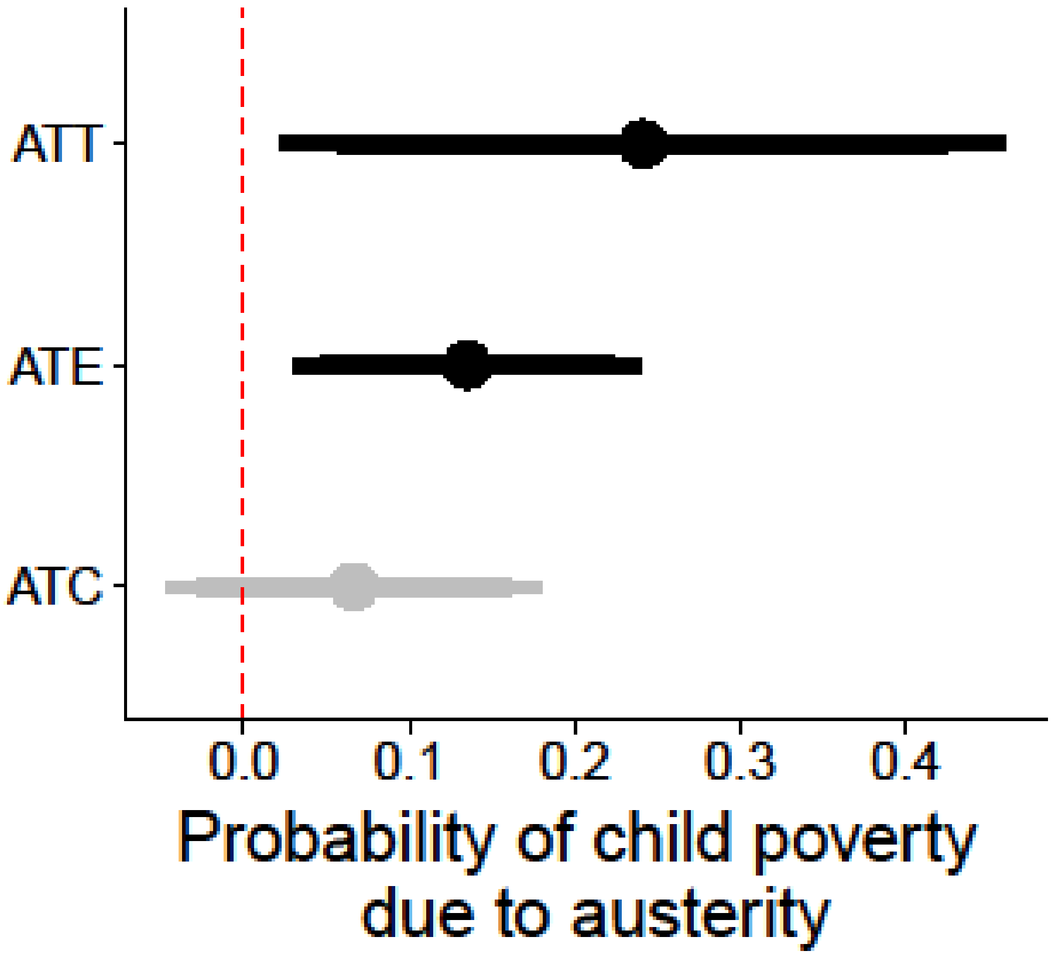
Conceptual framework (DAG): identifying IMF treatment effect on child poverty

- IMF affects child poverty via the red paths.
- By doing the following, (1) conditioning on observables and…
- ….(2) blocking the backdoor path SELECTION INTO IMF_{t-2} ← POLITICAL $WILL_{t-2}$ → CHILD $POVERTY_t$ …
- …we posit conditional ignorability. That is:

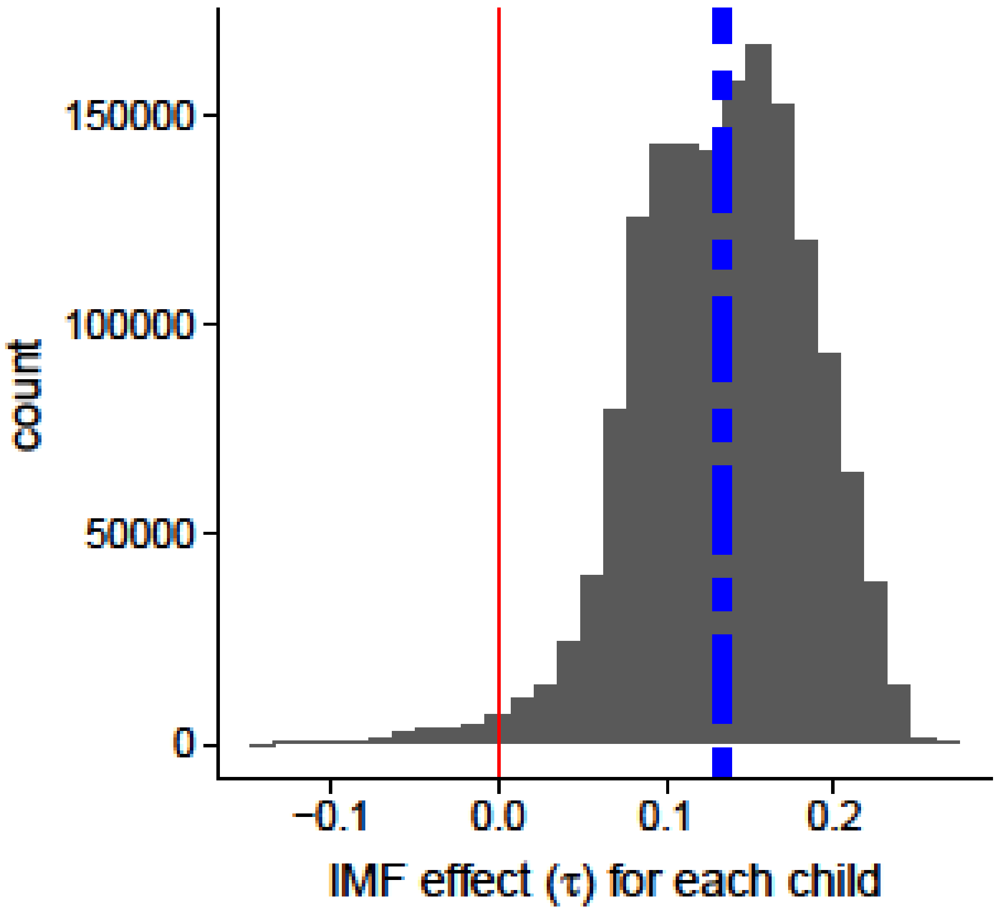
$$Y_i(1), Y_i(0) \perp\!\!\!\perp D \,|\, \mathbf{X}, ImR$$

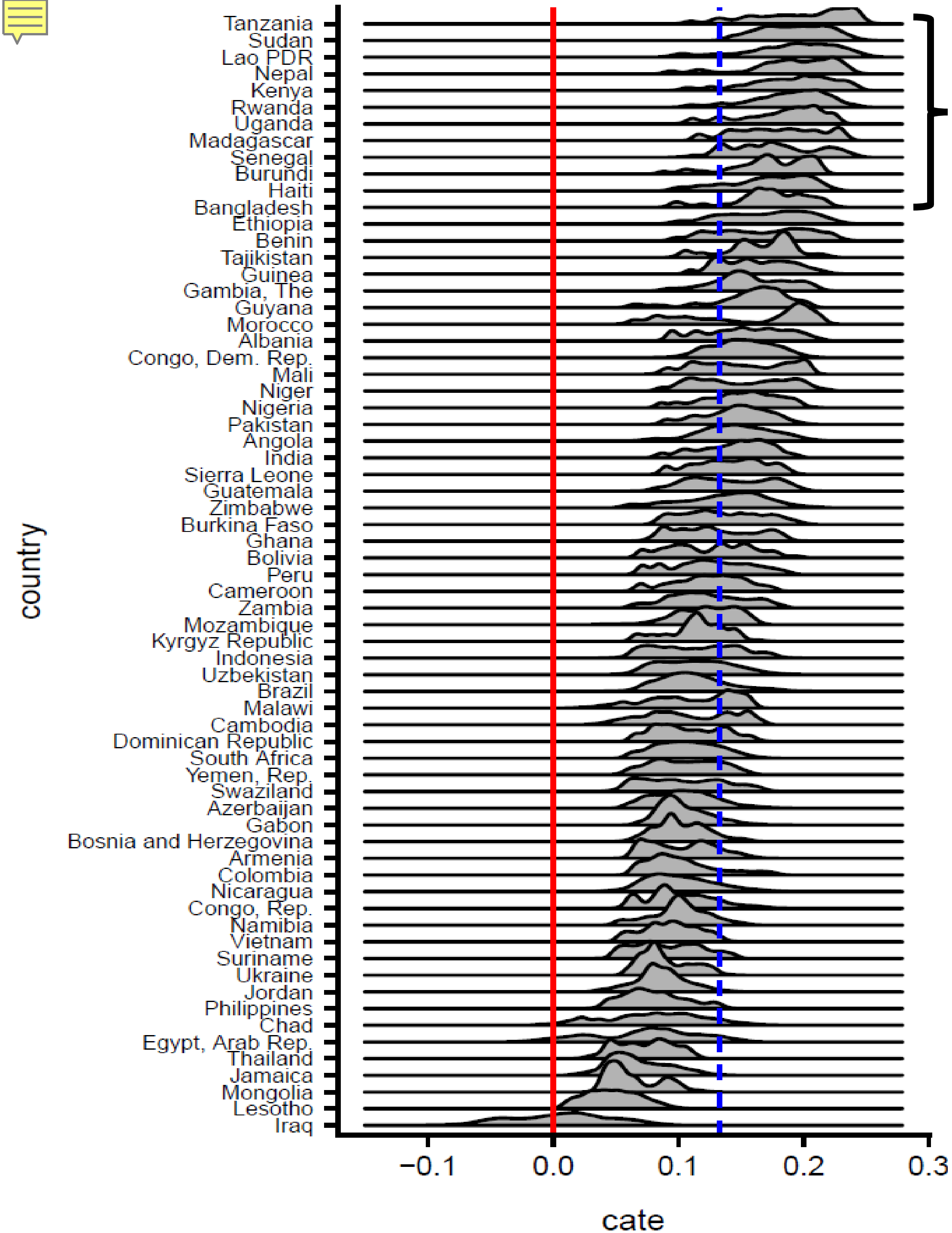


Average IMF impact

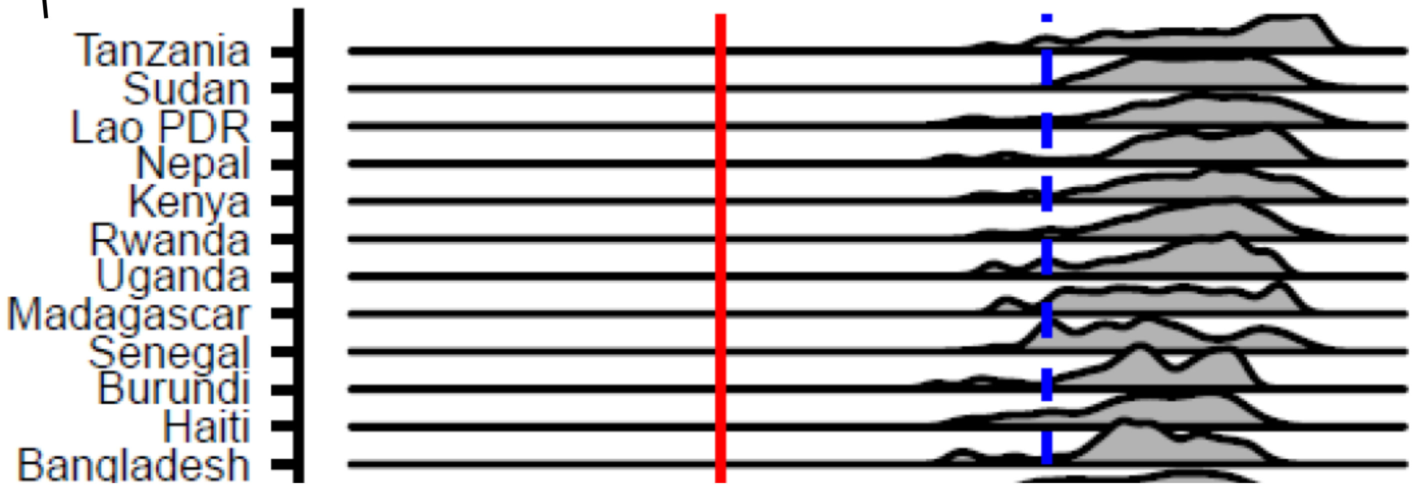


Impact heterogeneity by children



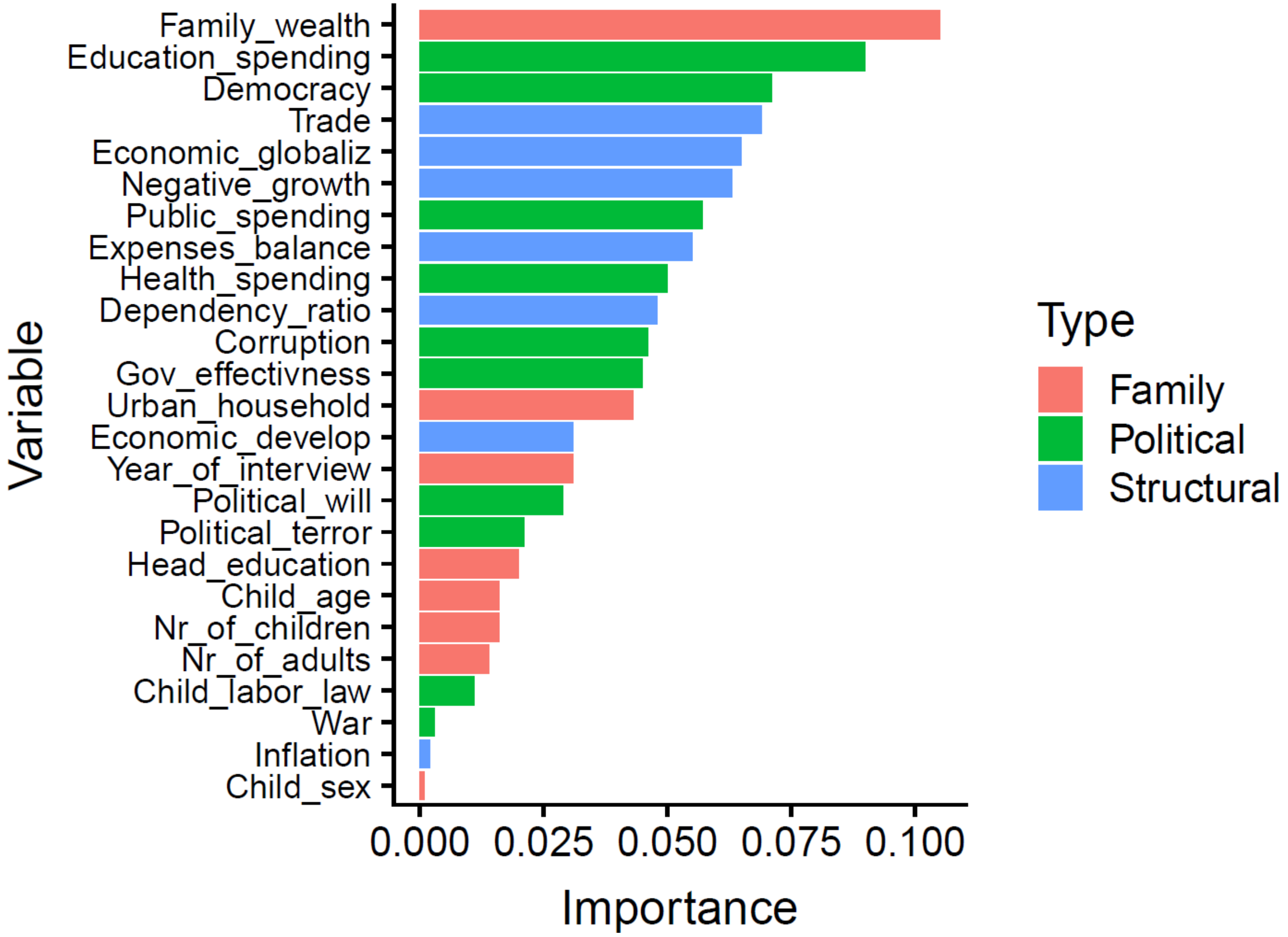


Impact by society



Turns out that Tanzania has one of highest IMF conditionality burden…

Which variables are predictive of impact heterogeneity?

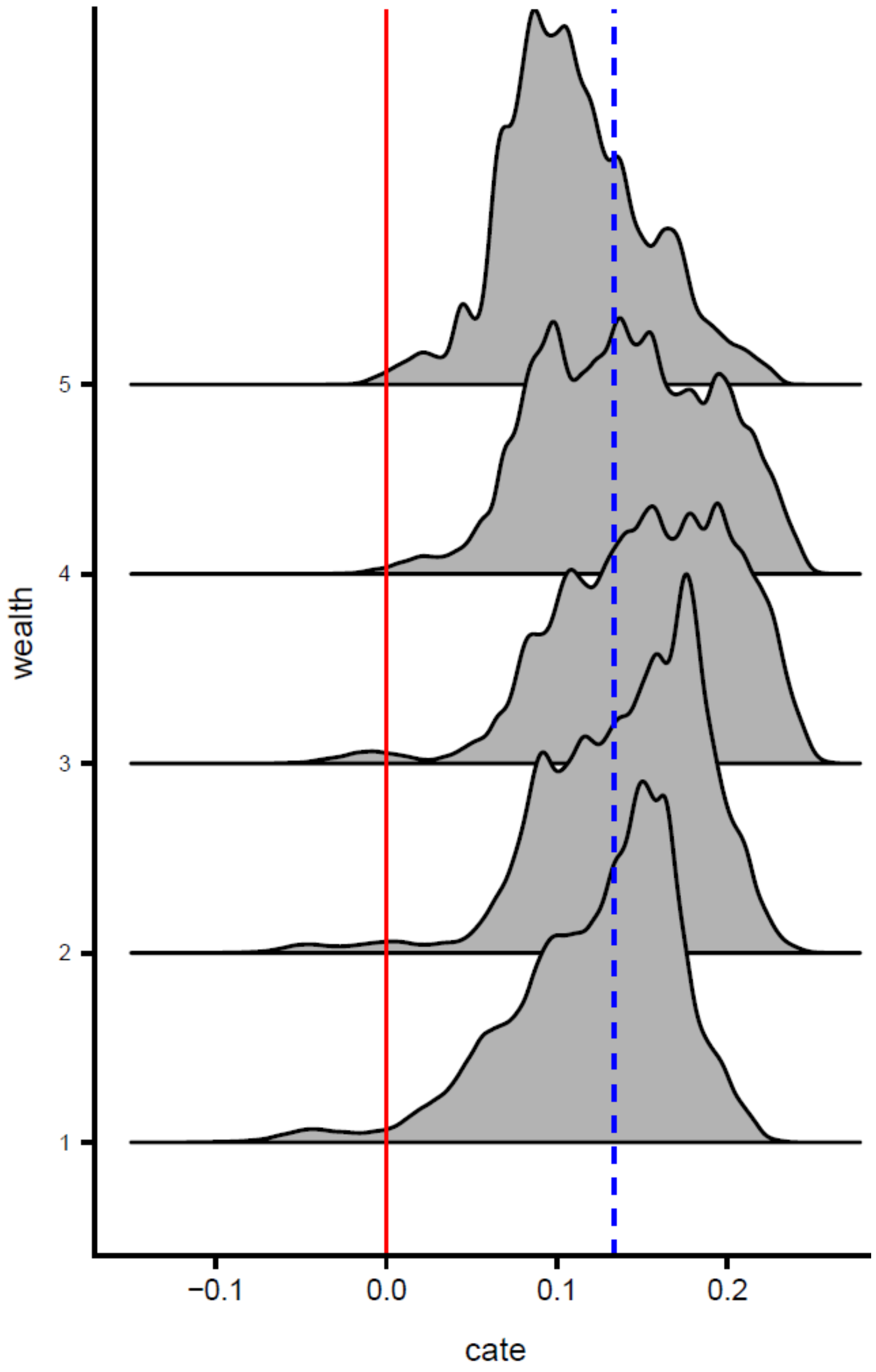


Impact heterogeneity by **family wealth**

- Children of the middle-social stratum are affected at least as hard by the macroeconomic shockwaves than children of the top and bottom strata (quintiles).

➔Top quintile families can protect their children better than families of the other social strata

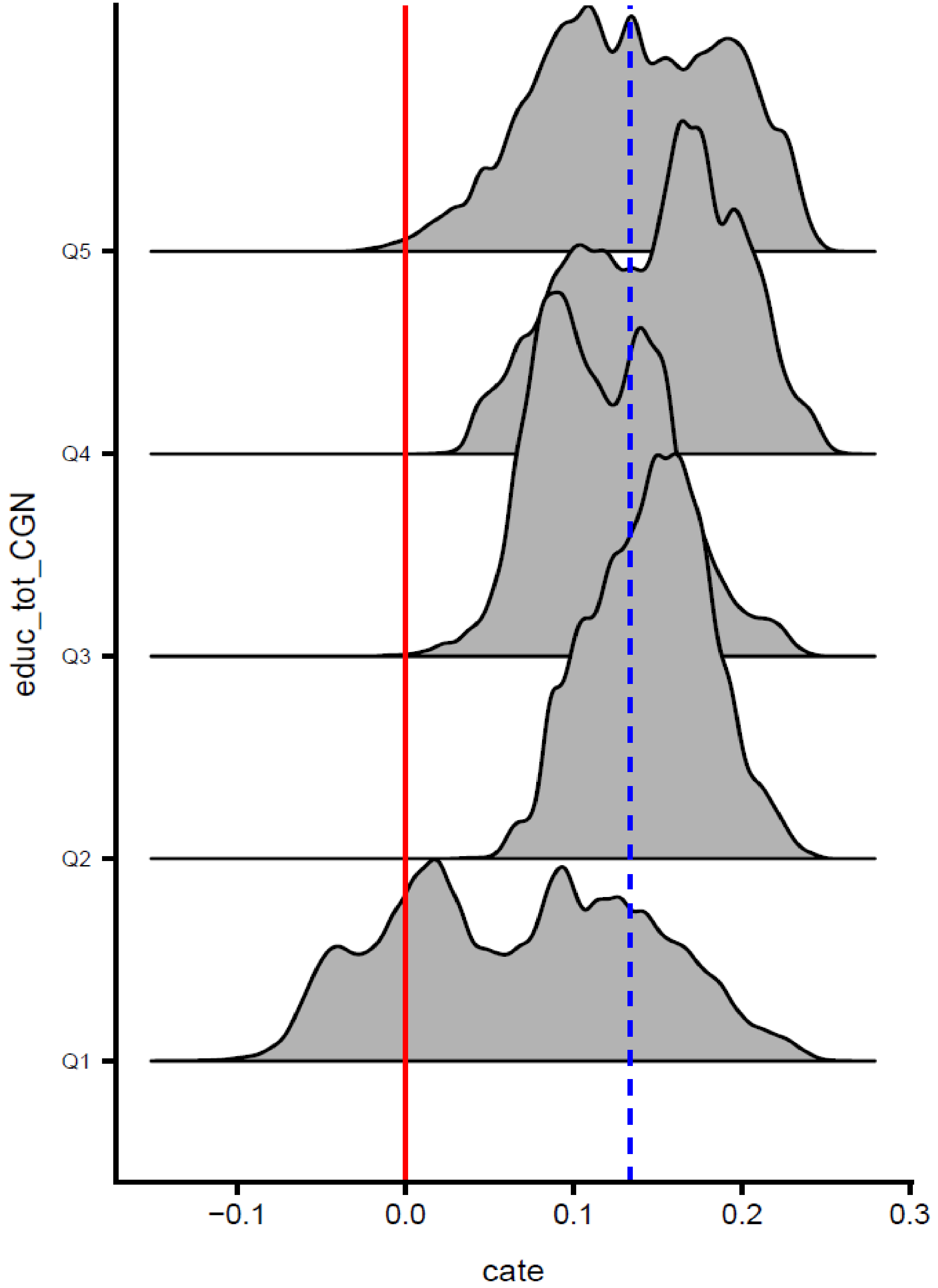
➔ Bottom quintile families are already relatively excluded.



Impact heterogeneity by **governments’ education spending**

- For those children residing in societies that spend most on education, are at a higher risk of falling into poverty during IMF programs.

➔ In the interest of balancing governments’ budget, the IMF aims to reduce public spending, including health and education expenses.



If one has tabular data, then analyzing CATE is straightforward.

Can we analyze CATE if we lack tabular data?

In the AI and Global Development Lab, we are

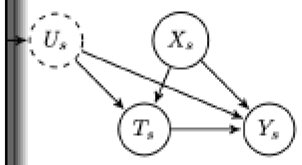
te images

Image-based Treatment Effect Heterogeneity

Connor T. Jerzak
Institute for Analytical Sociology
Linköping University
Email: connor.jerzak@liu.se
Website: ConnorJerzak.com

Fredrik Johansson
Data Science and AI Division
Chalmers University of Technology
Email: fredrik.johansson@chalmers.se
Website: fredjo.com

Adel Daoud
Institute for Analytical Sociology
Linköping University
Email: adel.daoud@liu.se
Website: AdelDaoud.se
AI and Global Development Lab: global-lab.ai



ANALYSIS

Existing Methods for TREATMENT EFFECT HETEROGENEITY:

- *Tabular CATE:* $\tau(\mathbf{x}) = \mathbb{E}[Y_i(1) - Y_i(0) \mid \mathbf{X}_i = \mathbf{x}]$
  - $\mathbf{x}$ – collected by researchers at *time* & *place* of intervention
  - *Potentially subject to missingness precisely for those w/ largest $\tau(\mathbf{x})$ (e.g., most vulnerable)*
- Flexible modeling of $\mathbb{E}[Y_i(1)|\mathbf{X}_i]$ & $\mathbb{E}[Y_i(0)|\mathbf{X}_i]$ (X-, S-, T-, etc. learners)
- Tree-based methods Athey et al. (2016)

Instead of relying on researcher-collected covariates

This project:

$\leadsto$ Decomposes effect heterogeneity using satellite data, $\tau(\mathbf{m})$–(no missingness!)

$\leadsto$ Allows neighborhood/geographical/historical heterogeneity reasoning

$\leadsto$ Massively enabling transportability analysis

Instead of focusing on flexibility in modeling

This project:

$\leadsto$ Develops a model prioritizing interpretability...

$\leadsto$ ... while incorporating uncertainty into an ML framework

— — —

Images unpack otherwise unseen moderators & sources of heterogeneity

STARTING FROM THE REAL WORLD PROBLEM

CONTEXT Uganda, Youth Opportunities Program Blattman et al. 2014

- Great geological diversity
- Diverse w.r.t. human settlements

TREATMENT Random assignment of youth team to business grant

OUTCOME Income; aggregate summary of skilled labor (2 year post)

X_i Age, gender, etc.

M_i Landsat satellite images of unit's neighborhood

How to make sense of heterogeneity based on neighborhood/geographic info in M_i?

Intuition for our solution:

"All models are wrong, some are useful" $\approx$ *Box in JASA, 1976*

=

Flexible enough, yet simple enough. Uncertainty included!

Observation :

$$\tau_i = Y_i(1) - Y_i(0)$$

$$\Longrightarrow$$

$$Y_i(1) = \underbrace{Y_i(0)}_{\text{(a) Flexible fxn of } \mathbf{M}_i} + \underbrace{\tau_i}_{\text{(b) Mixture model based on } \mathbf{M}_i}$$

- (a) Allows for *flexible prediction* of baseline outcome
- (b) Allows for *efficient summarization* of complex image effects via clusters
- (a)+(b) combined in variational inference framework $\rightsquigarrow$ *model uncertainty*

- For INCOME: likely no neighborhood/geographic heterogeneity via $\mathbf{M}_i$
- For HUMAN CAPITAL: some evidence isolated places in mountainous terrain $\rightsquigarrow$ smaller T_i benefit

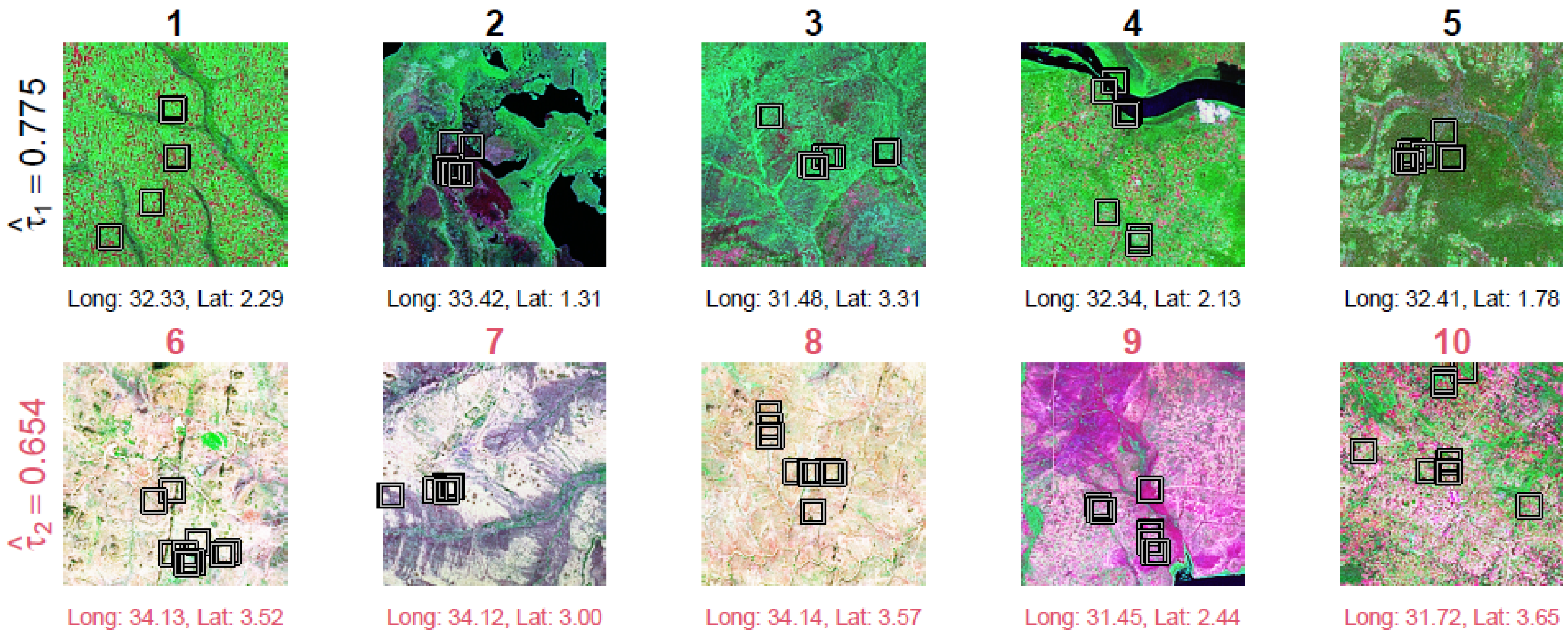


Figure: *Top*. High probability cluster 1 images. *Bottom*. High probability cluster 2 images.

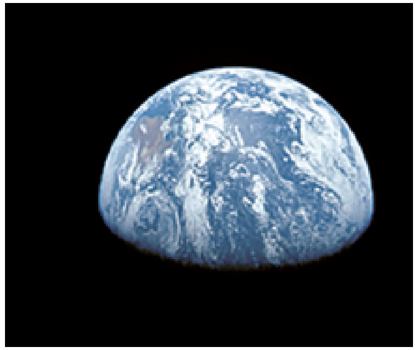
Details & Extensions

In paper, we further discuss:

- Validation via simulation
- Identifying sensitivity regions in image
- Causal regularization

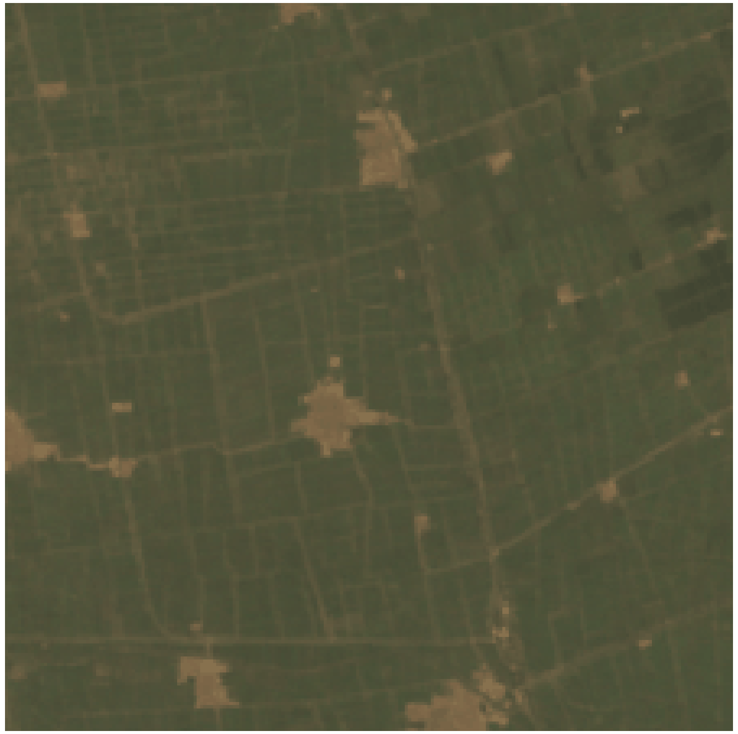
Ongoing work:

- Expand causal transportability via satellite images
  - Deploy fact that $\mathbf{M}_i$ universally available
  - $\mathbb{E}[Y_i(t)|\mathbf{M}_i]$ can be estimated even where no researcher-collected $\mathbf{X}_i$ collected
  - Gateway to **Global Causal Inference**?



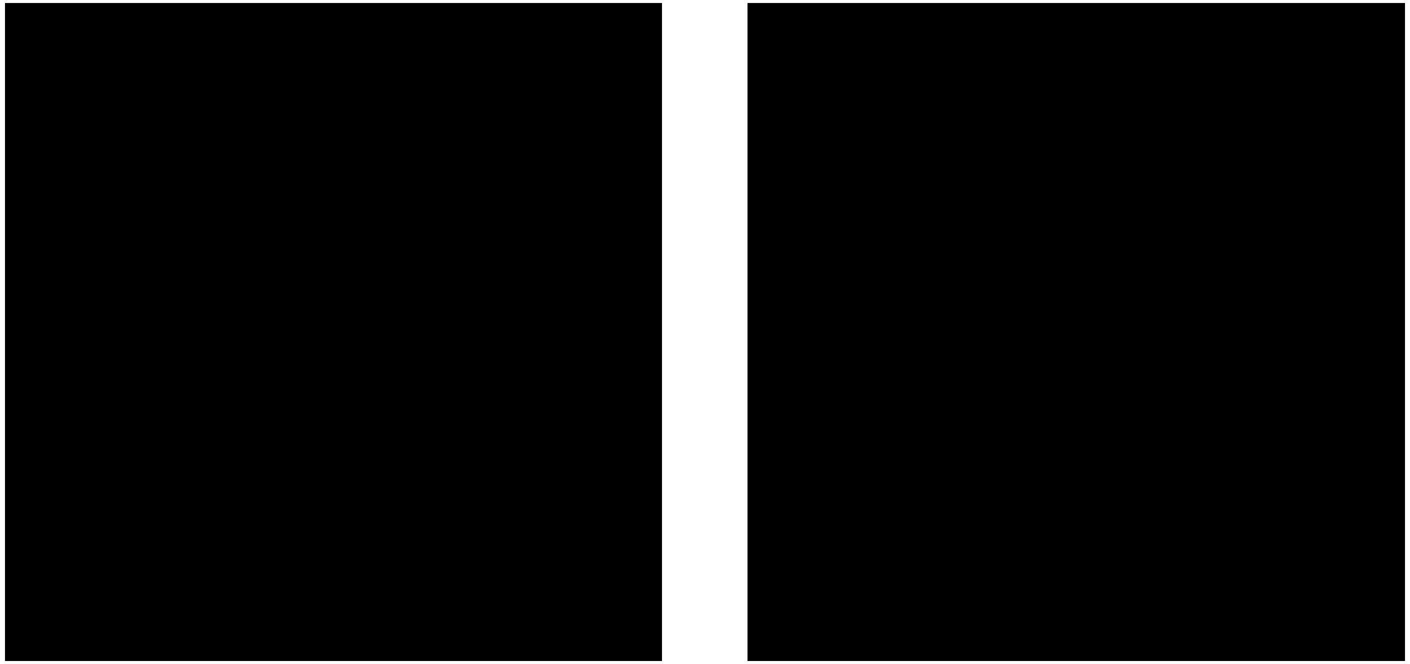
Identifying mechanisms (mediation) at the neighborhood level via videos

Egypt examples

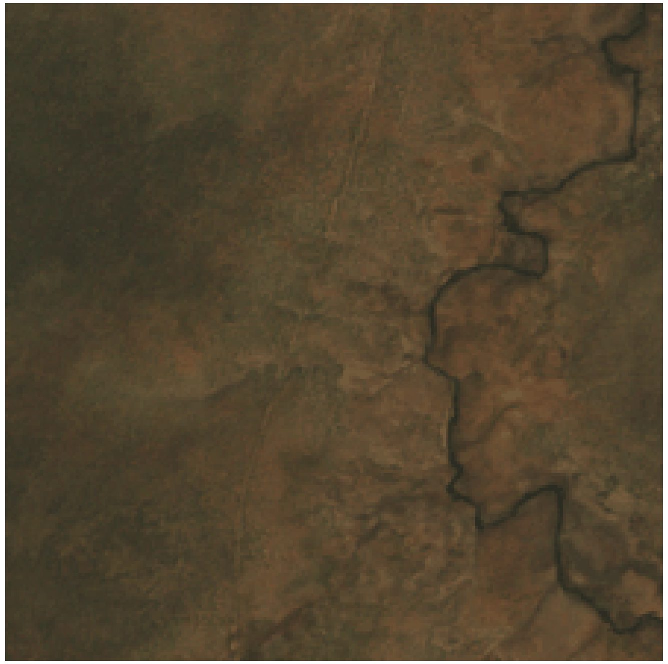


But there are data challenges

Corrupted-signal example from Ghana and Nigeria



Too-rapid-change example from Egypt



Conclusion

Social reality is inherently ***spatial, historical, dynamic****, but our data are often based on static researcher-devised constructs*

Satellite images bridge this gap

- $\leadsto$ learn about previously unobserved (evolving) neighborhood/geological/etc. factors important for some heterogeneity outcomes
- $\leadsto$ $\mathbf{M}_i$ can be harnessed in probabilistic ML models $\implies$ flexibility + simplicity = 😎

- Many exciting open questions:
  - Non-formalizable sources of uncertainty (e.g., pre-processing)
  - Where is $\mathbf{M}_i$ relevant for heterogeneity?
  - How do proposed models work w/ image data from healthcare?

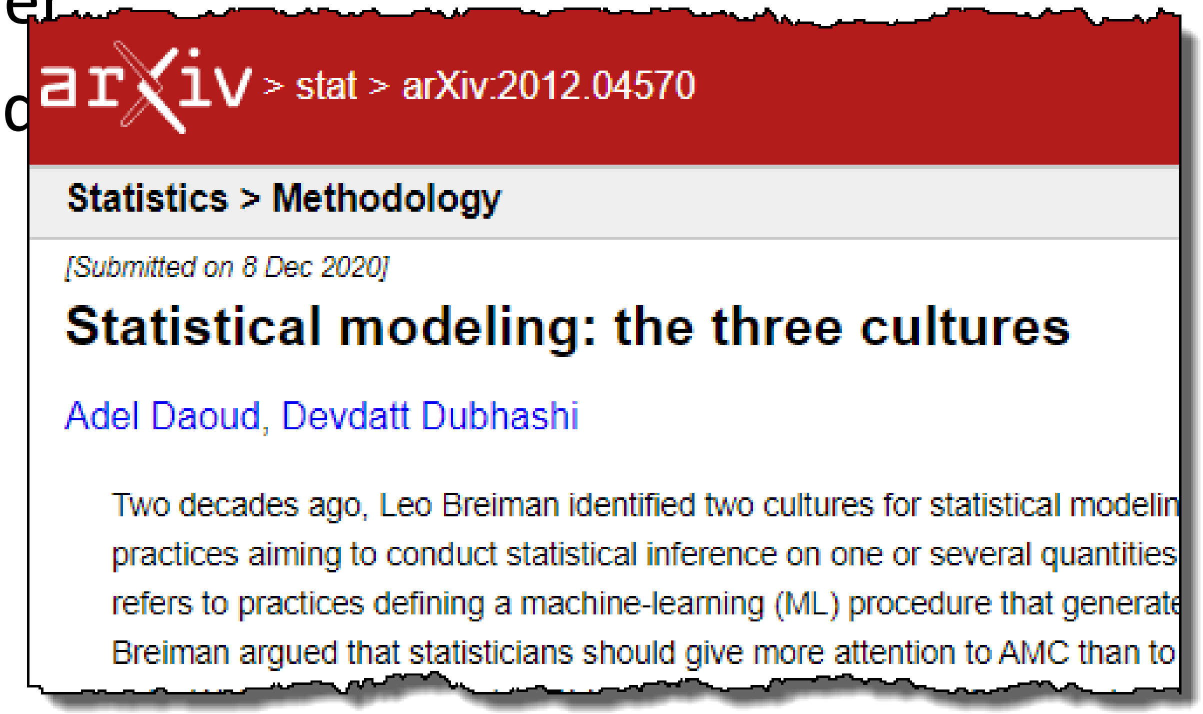
…and Beyond

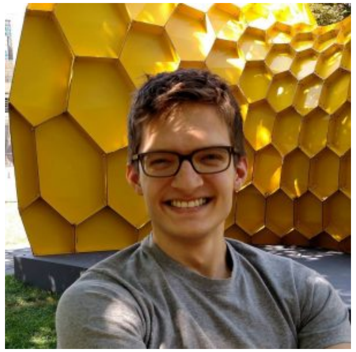
- Applied research on global Sustainable Development
  - Incorporate climate change, economic and ecological systems
- Methodological development for planetary wide health and social science
  - Statistical work on integrating causality and prediction
  - Moving beyond the qualitative and quantitative divide:
    - Create a framework for *qualitative potential outcomes* (e.g., thought experiments) to generate new hypotheses and spur new research.
    - Grounded theory for causal inference (analyzing CATE and ITE).
- Meta-theoretical development
  - Linking analytical sociology to (statistical) causal inference
  - Formulating an ontological and epistemological foundation for computational social science

Summary

- New methods available. Currently, social scientist operate mostly in methods from traditional statistics and less so from data science.
- New data sources combined with ML brings new opportunities to global health and poverty research and policy.
- Towards a new social-scientific frontier:

➔A mixture of research modes beyond
divide.











Adel.daoud@liu.se and global-lab.ai

Funded mainly by the Swedish Research Council (SRC)

More information at

- global-lab.ai (website of AI and Global Development Lab)
  - We are recruiting both CS and social or epi postdocs with strong skills in data science and causal inference.
- adeldaoud.se (my website) and adel.daoud@liu.se
- Additionally, check out *The Journeys of Scholars* at YouTube. A conversation-style podcast with top scholars, about "deciphering the pursuit of excellence."