Statistical modeling: the three cultures

Presented at SWE-REG conference
Machine learning methods in the medical and social sciences,
March 2020.

Adel Daoud, PhD,

Associate Professor, Department of Sociology, University of Gothenburg,

Associate Professor, Institute for Analytical Sociology, Linköping University,

Affiliated Associate Professor in Data Science and Artificial Intelligence for the Social Sciences, the Division of Data Science and Artificial Intelligence, Department of Computer Science and Engineering, Chalmers Technical University.

adel.daoud@sociology.gu.se and daoud@chalmers.se

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 modeling. The data modeling culture (DMC) refers to practices aiming to conduct statistical inference on one or several quantities of interest. The algorithmic modeling culture (AMC) refers to practices defining a machine-learning (ML) procedure that generates accurate predictions about an event of interest. Breiman argued that statisticians should give more attention to AMC than to DMC, because of the strengths of ML in adapting to data. While twenty years later, DMC has lost some of its dominant role in statistics because of the data-science revolution, we observe that this culture is still the leading practice in the natural and social sciences. DMC is the modus operandi because of the influence of the established scientific method, called the hypothetico-deductive scientific method. Despite the incompatibilities of AMC with this scientific method, among some research groups, AMC and DMC cultures mix intensely. We argue that this mixing has formed a fertile spawning pool for a mutated culture that we called the hybrid modeling culture (HMC) where prediction and inference have fused into new procedures where they reinforce one another. This article identifies key characteristics of HMC, thereby facilitating the scientific endeavor and fueling the evolution of statistical cultures towards better practices. By better, we mean increasingly reliable, valid, and efficient statistical practices in analyzing causal relationships. In combining inference and prediction, the result of HMC is that the distinction between prediction and inference, taken to its limit, melts away. We qualify our melting-away argument by describing three HMC practices, where each practice captures an aspect of the scientific cycle, namely, ML for causal inference, ML for data acquisition, and ML for theory prediction.

Subjects: **Methodology (stat.ME)**; Computers and Society (cs.CY)
Cite as: arXiv:2012.04570 **[stat.ME]**
(or arXiv:2012.04570v1 **[stat.ME]** for this version)

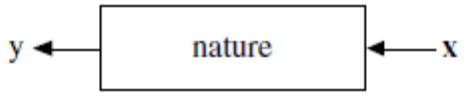
https://arxiv.org/abs/2012.04570

Thinking predictively, inferentially, and causally

Inference, the aim of the data modeling culture

- Leo Breiman, 2001, "Statistical modeling: the two cultures", *Statistical Science*, provides a good starting point.

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:

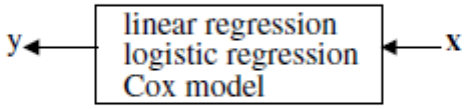
Prediction. To be able to predict what the responses are going to be to future input variables;
Information. To extract some information about how nature is associating the response variables to the input variables.

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)

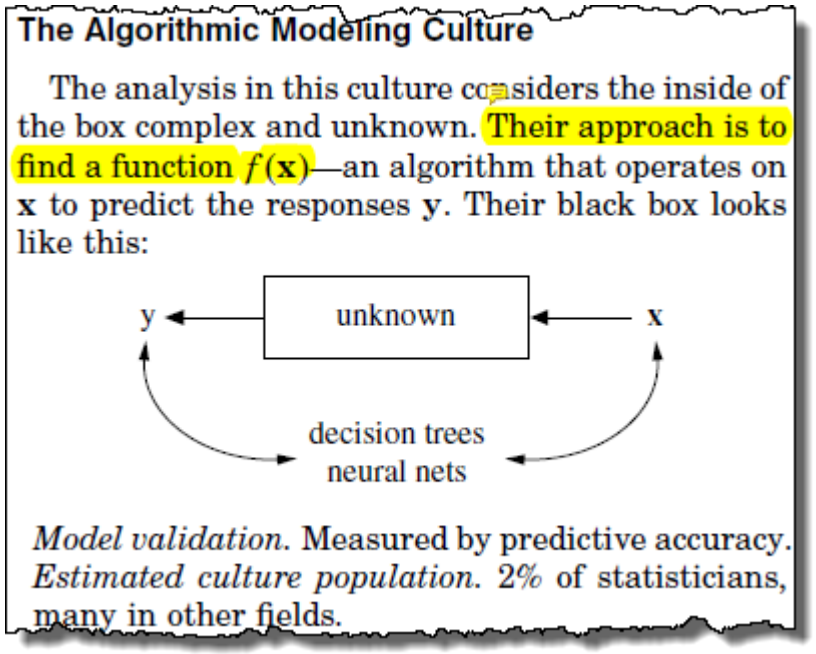
The values of the parameters are estimated from the data and the model then used for information and/or prediction. Thus the black box is filled in like this:



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.
- The algorithms usually refers to machine learning methods.



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 $\mathbf{x}$ to predict the responses $\mathbf{y}$. Their black box looks like this:

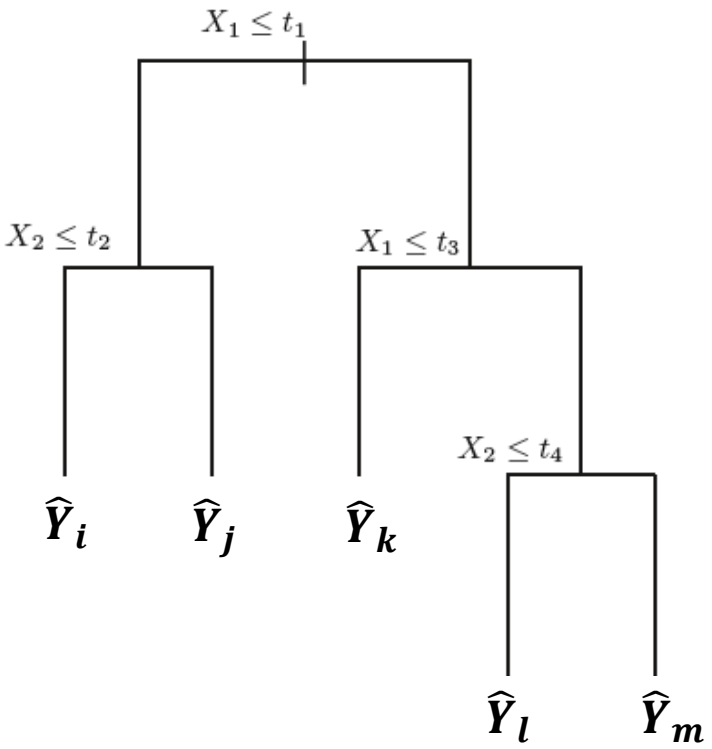
Model validation. Measured by predictive accuracy.
Estimated culture population. 2% of statisticians, many in other fields.

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

- **In inference**, we estimate β, given a pre-specified function, *f*, for example a linear model:
  - $Y = \widehat{\beta_0} + \widehat{\beta_1}T + \widehat{\beta_2}X + e$
  - *But the goal is not to predict* $\hat{Y}$ for new data!
- **In prediction**, 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 (supervi**



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:

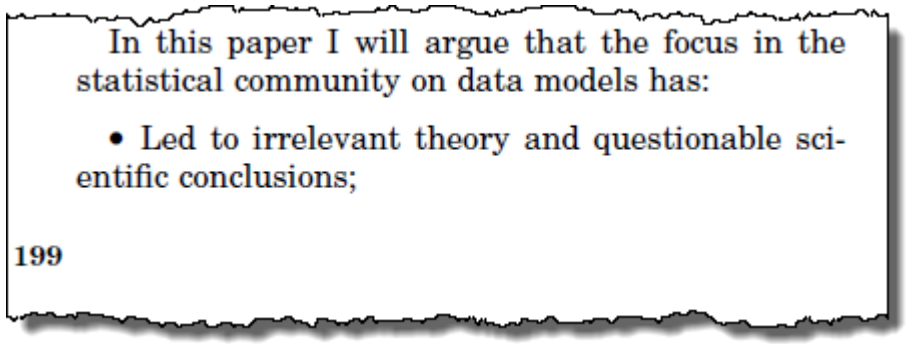
Model	*method* Value	Type	Libraries	Tuning Parameters
Adaptive-Network-Based Fuzzy Inference System	ANFIS	Regression	frbs	num.labels, max.iter
Bayesian Regularized Neural Networks	brnn	Regression	brnn	neurons
Bayesian Ridge Regression	bridge	Regression	monomvn	None
Bayesian Ridge Regression (Model Averaged)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

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

	Traditional (GLM) inference	**Machine learning (ML)**
Exemplifying question	What is the effect of economic crisis on child mortality?	What will the child mortality rate be next year?
Goal	Unbiased estimate	Accurate prediction
Limitation	Forces untestable assumptions	Rely on black-box models
Quantity of Interest	$\hat{\beta}$	$\hat{Y}$

So should we conduct more predictive and less inferential studies in the medical and social sciences!??

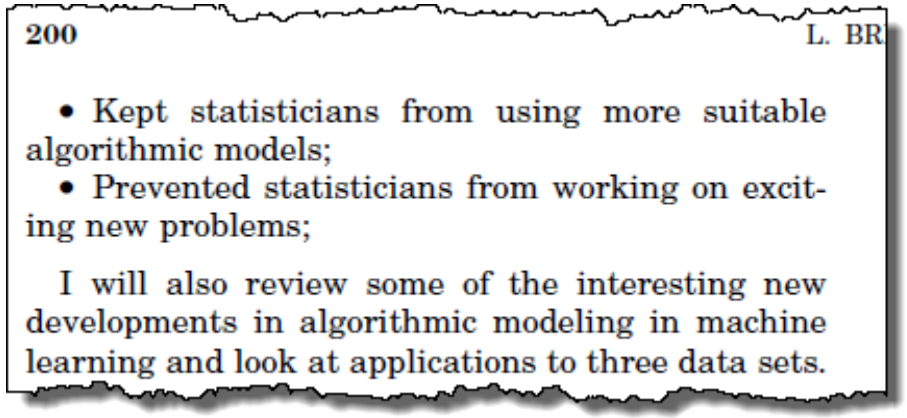
- Breiman wanted more scholars to endorse the algorithmic culture and its $\hat{Y}$-problems….
- Others are echoing that call. See for example the following overviews: Mullainathan & Spiess 2017 in **economics**, Cranmer & Desmarais 2017 in **political science**, Molina & Garip 2019 in **sociology**, and Yarkoni & Westfall 2017 in **psychology**, Wiemken and Kelley in **public health**.



In this paper I will argue that the focus in the statistical community on data models has:

- Led to irrelevant theory and questionable scientific conclusions;

199



200 L. BR

- Kept statisticians from using more suitable algorithmic models;
- Prevented statisticians from working on exciting new problems;

I will also review some of the interesting new developments in algorithmic modeling in machine learning and look at applications to three data sets.

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

I want to offer an alternative perspective….

…beyond the two cultures and towards a hybrid.

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, $\hat{Y}_{G_1}, \hat{Y}_{G_2}, \ldots, \hat{Y}_{G_k}$, 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	$\hat{\tau} = \hat{Y}_i^1 - \hat{Y}_i^0$	$\hat{X}, \hat{W}, \hat{Y}$	$\epsilon_{\hat{Y}_{G_k}} \approx Y - \hat{Y}_{G_k}$ or $\epsilon_{\hat{\tau}_{G_k}} \approx \tau - \hat{\tau}_{G_k}$

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

(Causal inference)

ML supports causal inference in at least three ways; or how $\hat{\tau}$ replaces $\hat{\beta}$.

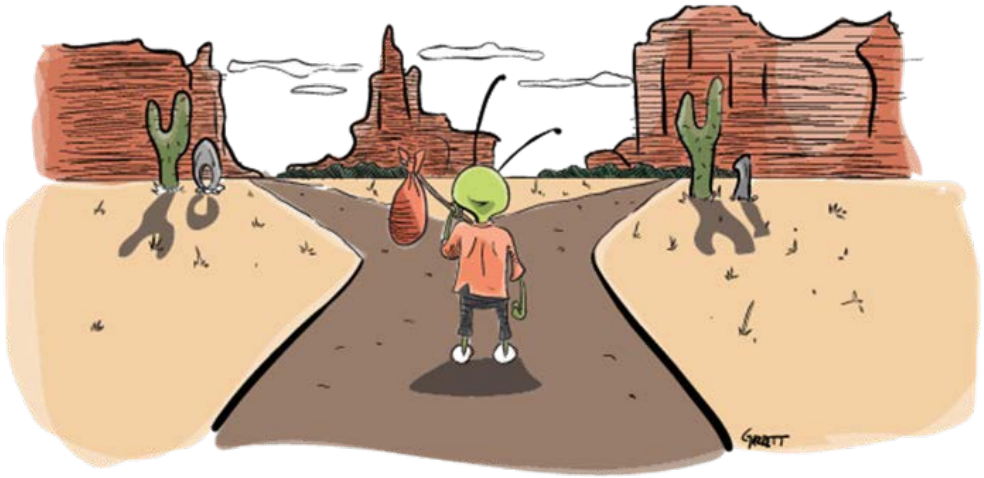
1. Impute potential outcomes, $\hat{Y}(0), \hat{Y}(1)$
2. Ignorability assumption (as-if random)
3. Treatment heterogeneity (+ functional form)

The fundamental problem of causal inference

- The potential outcome framework. We cannot observe an individual's, *i*, two outcomes (Y_i): with a treatment (T=1) and without it (T=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.



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

(1) ML imputes potential outcomes

- Assuming (conditional) ignorability
- Estimate "ITE" for all children, with and without treatment:

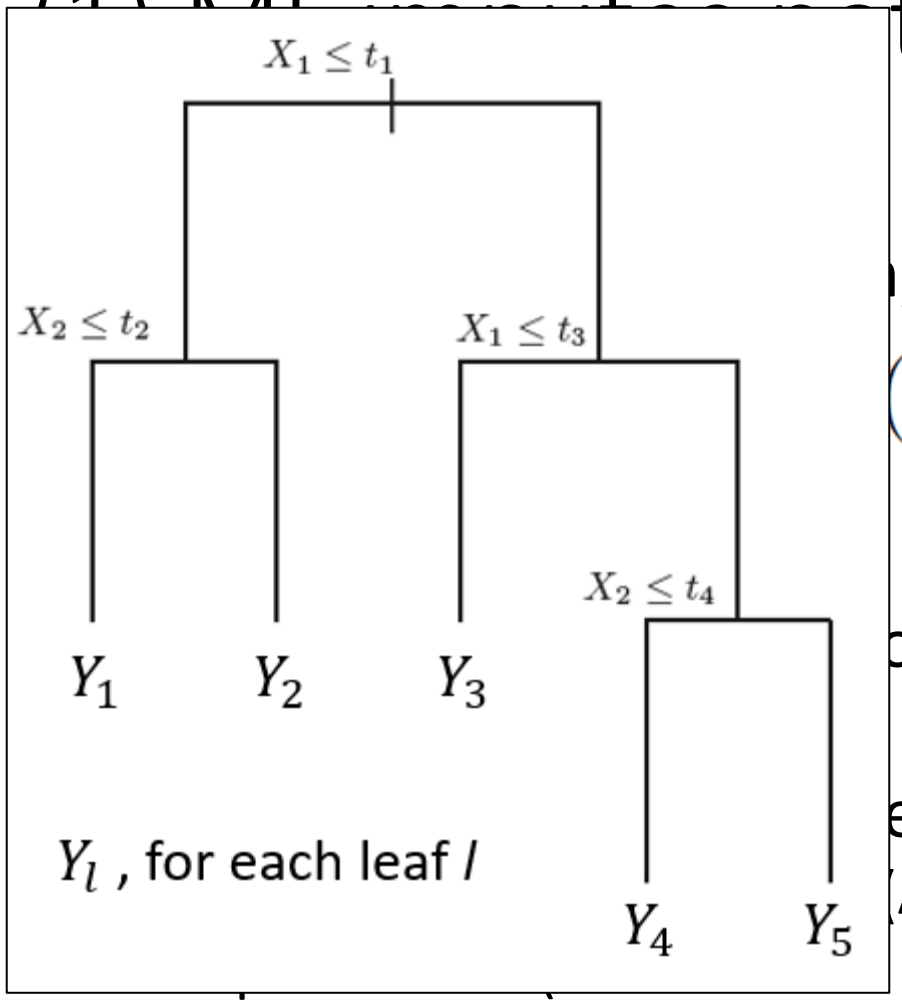
$$Y_i(1) = \hat{m}_1(x_i) \text{ and } Y_i(0) = \hat{m}_0(x_i).$$

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

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

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

(1) ML imputes potential ou



, with and with

(x_i) and $Y_i($

on treated an

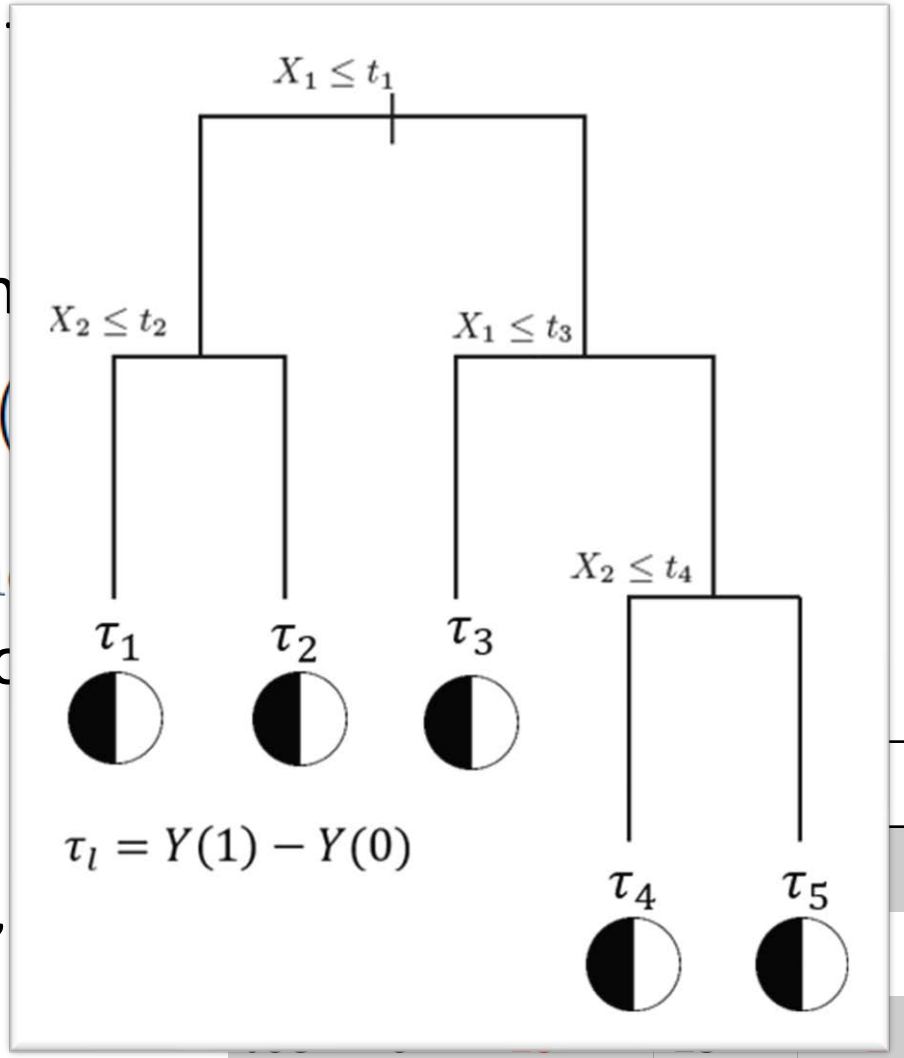
omes, we get to

es,

(Athey, Tibshirani,

et al)

- See Overview on meta-learners (Kunzel et al)



Jan	0	24	22	+2

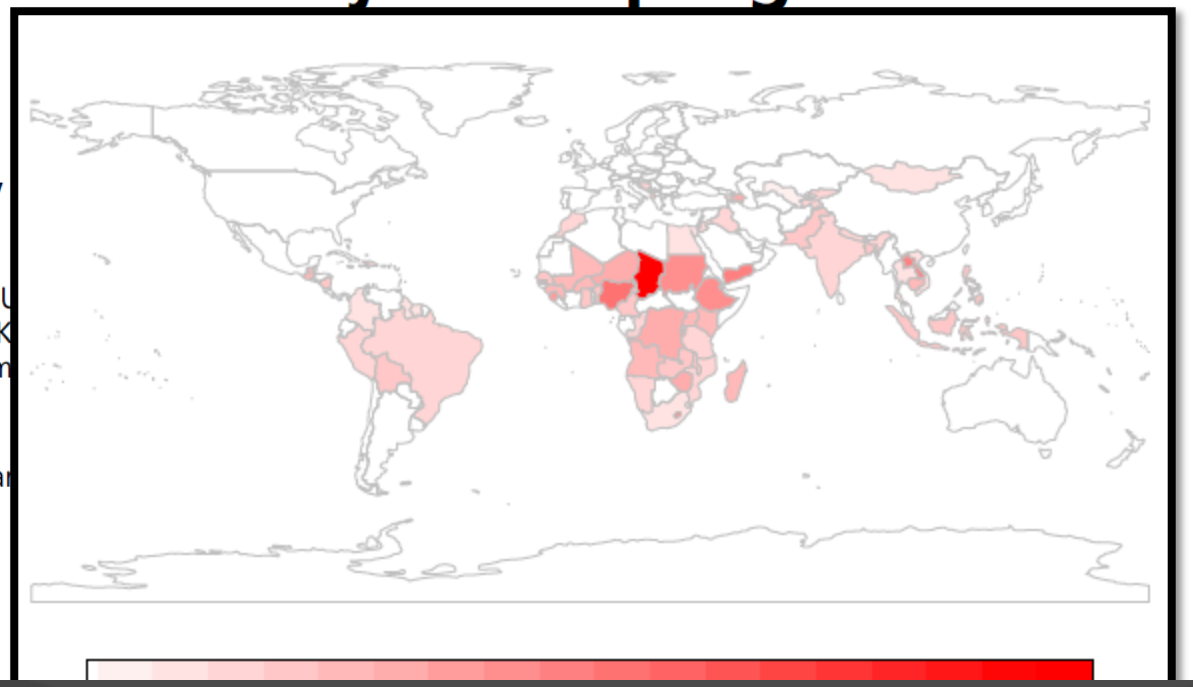
PNAS

Impact of International Monetary Fund programs on child health

Adel Daoud[a,1], Elias Nosrati[b], Bernhard Reinsberg[a], and Lawrence P. King[a,b]

[a]Centre for Business Research, Cambridge Judge Business School, U
Sociology, University of Cambridge, Cambridge CB2 3RQ, United K
[d]Department of Sociology, University of Amsterdam, 1018 WV Am
Hamilton 3240, New Zealand

Edited by Arjumand Siddiqi, University of Toronto, Toronto, ON, Ca
review October 20, 2016)



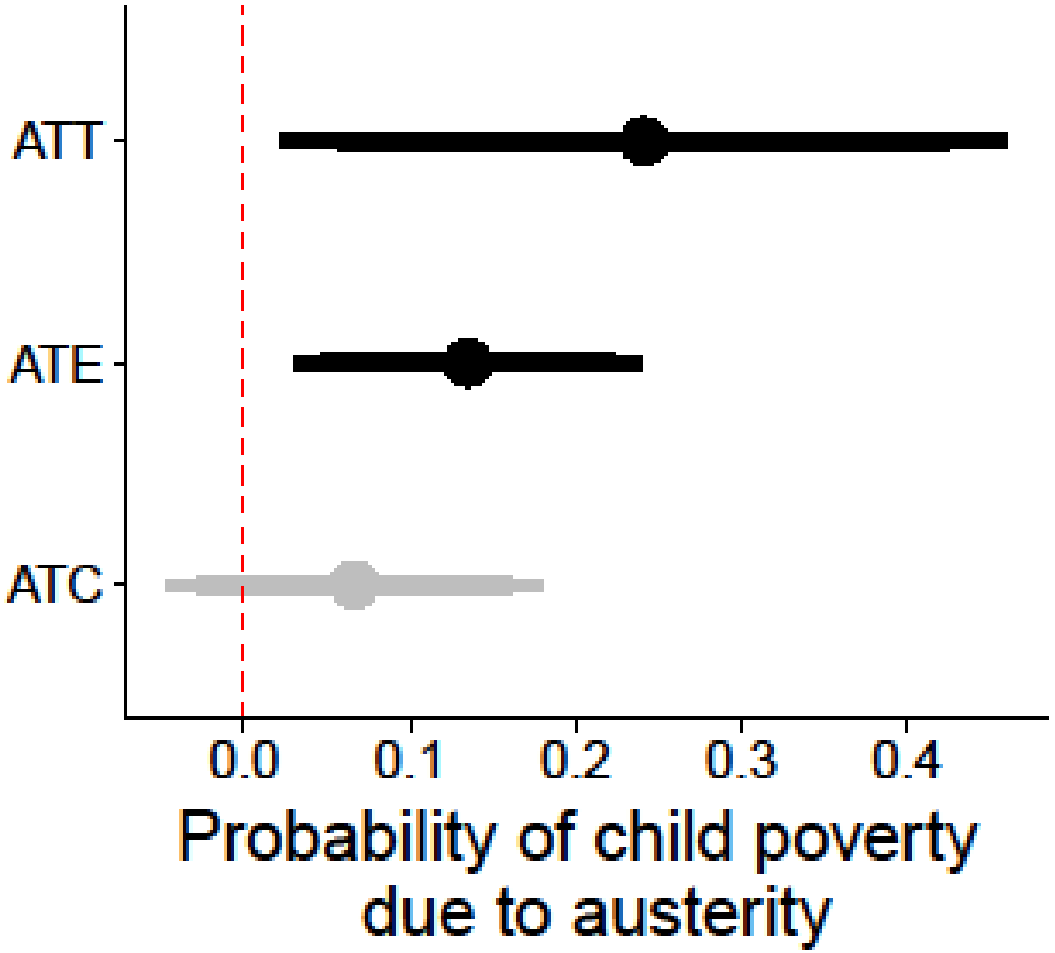
Estimating Treatment Heterogeneity of International Monetary Fund Programs on Child Poverty with Generalized Random Forest

AUTHORS
Adel Daoud, Fredrik Johansson

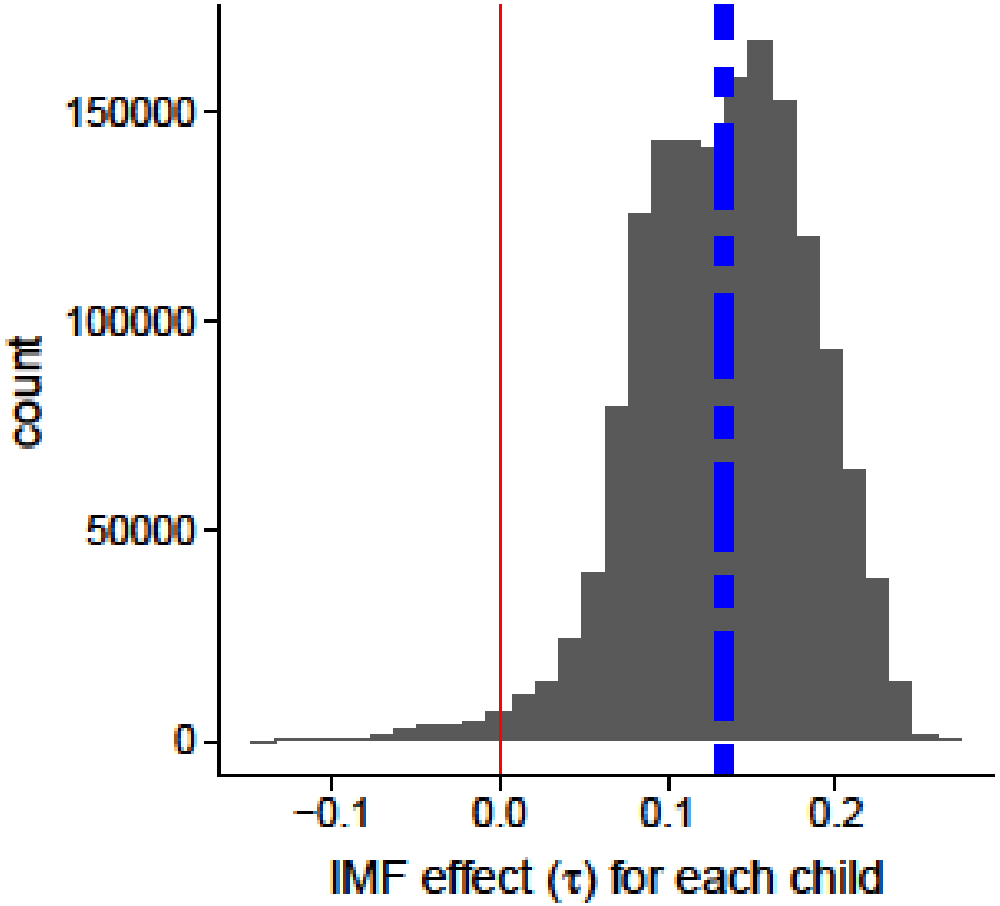
SUBMITTED ON
February 06, 2019

LAST EDITED
February 13, 2019

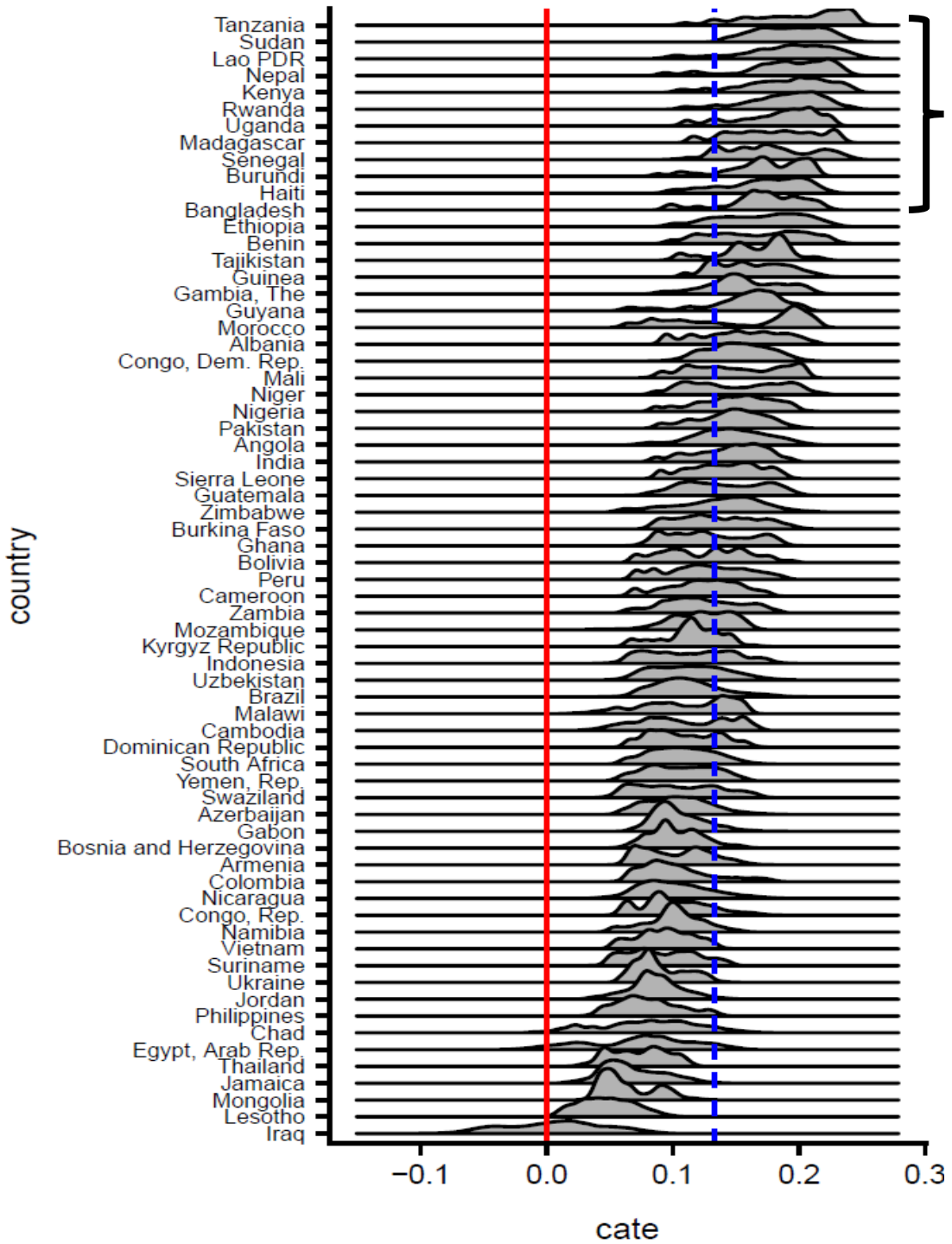
Average IMF impact

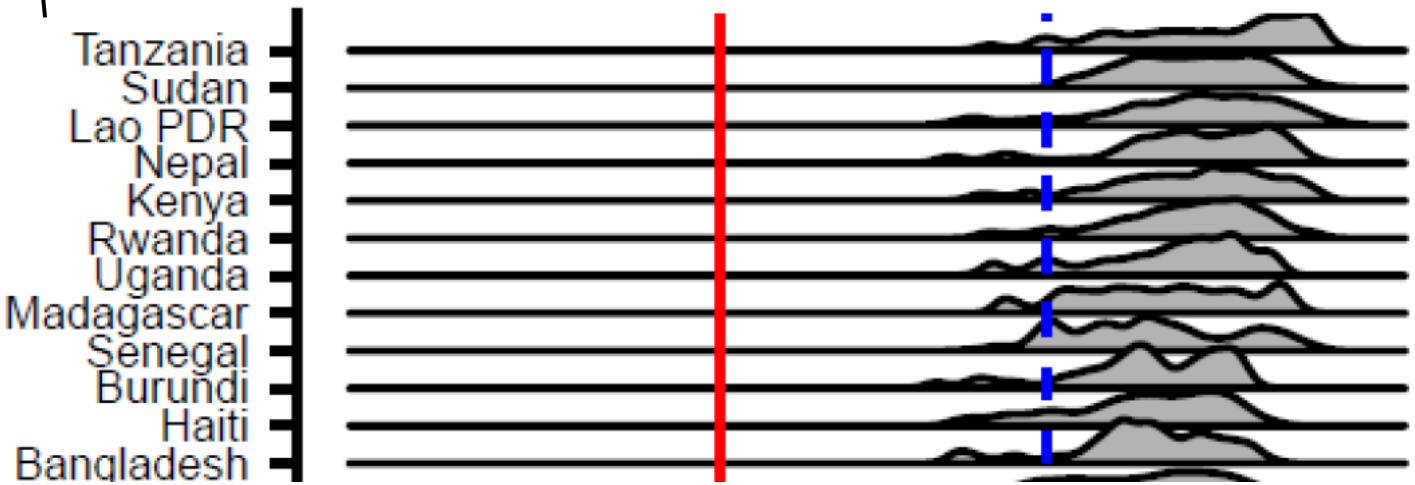


Impact heterogeneity by children



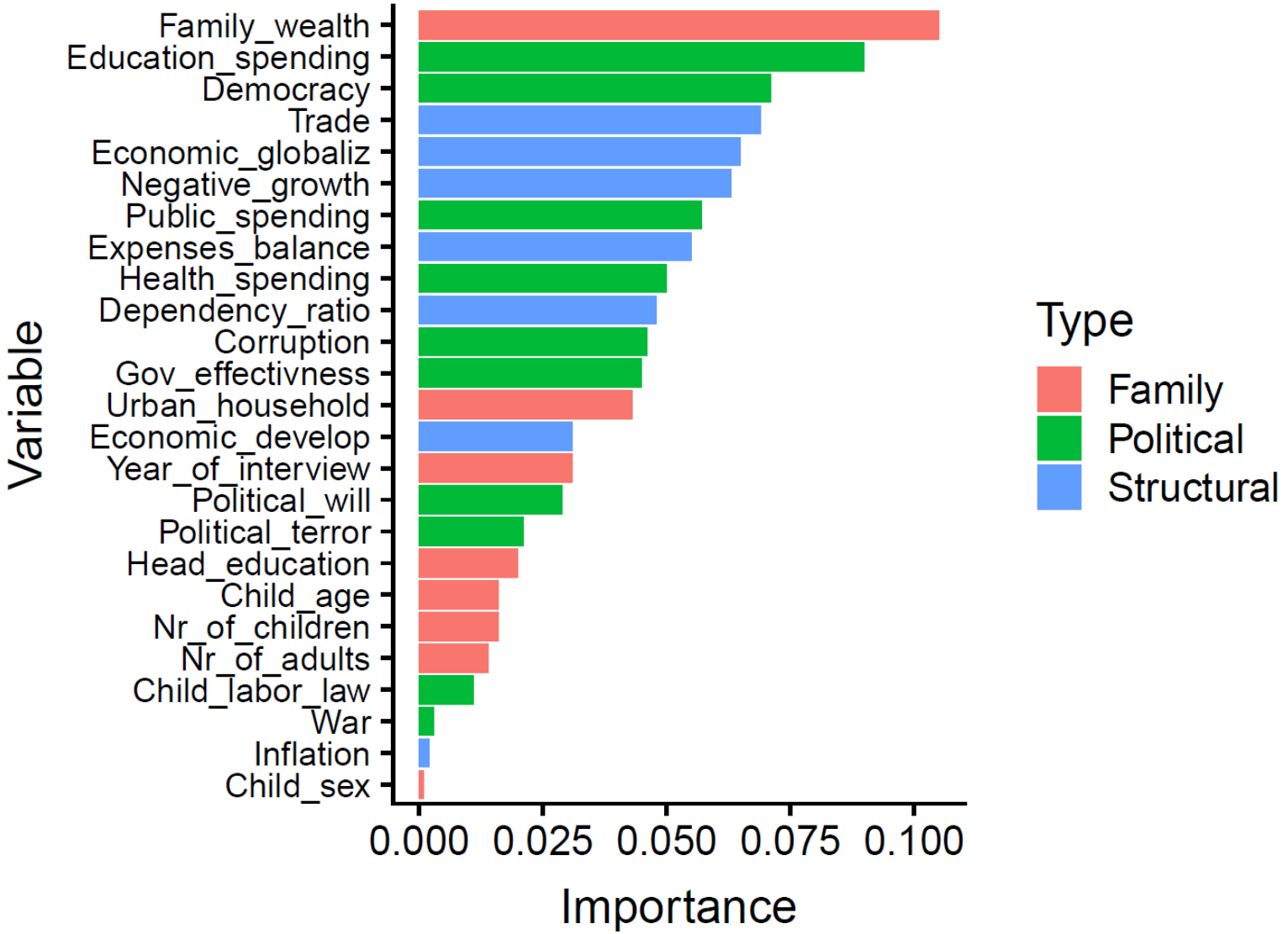
Impact by society





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

Which variables are predictive of impact heterogeneity?



In summary, causal inference in the hybrid culture translates to a specific form of $\hat{Y}$-problems: the old $\hat{\beta}$ has transformed to a $\hat{\tau}$.

Making sense of the new opportunities and pitfalls of machine learning…

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 modeling. The data modeling culture (DMC) refers to practices aiming to conduct statistical inference on one or several quantities of interest. The algorithmic modeling culture (AMC) refers to practices defining a machine-learning (ML) procedure that generates accurate predictions about an event of interest. Breiman argued that statisticians should give more attention to AMC than to DMC, because of the strengths of ML in adapting to data. While twenty years later, DMC has lost some of its dominant role in statistics because of the data-science revolution, we observe that this culture is still the leading practice in the natural and social sciences. DMC is the modus operandi because of the influence of the established scientific method, called the hypothetico-deductive scientific method. Despite the incompatibilities of AMC with this scientific method, among some research groups, AMC and DMC cultures mix intensely. We argue that this mixing has formed a fertile spawning pool for a mutated culture that we called the hybrid modeling culture (HMC) where prediction and inference have fused into new procedures where they reinforce one another. This article identifies key characteristics of HMC, thereby facilitating the scientific endeavor and fueling the evolution of statistical cultures towards better practices. By better, we mean increasingly reliable, valid, and efficient statistical practices in analyzing causal relationships. In combining inference and prediction, the result of HMC is that the distinction between prediction and inference, taken to its limit, melts away. We qualify our melting-away argument by describing three HMC practices, where each practice captures an aspect of the scientific cycle, namely, ML for causal inference, ML for data acquisition, and ML for theory prediction.

Subjects: **Methodology (stat.ME)**; Computers and Society (cs.CY)
Cite as: arXiv:2012.04570 **[stat.ME]**
(or arXiv:2012.04570v1 **[stat.ME]** for this version)

https://arxiv.org/abs/2012.04570