Online Appendix to:

Why Are the Affluent Better Represented Around the World?

Noam Lupu and Zach Warner

December 30, 2020

A.1 Data sources and coding rules

Table A1: Data sources

Variable	Source
Δ EMD	Lupu and Warner (Forthcoming) variable emd_diff.
Foreign cap. depend.	World Bank (2019), foreign direct investment, net inflows, balance of payments in current US dollars (variable BX.KLT.DINV.CD.WD). Gathered using the R package WDI and logged.
GDP (logged)	World Bank (2019), GDP per capita in constant 2010 US dollars (variable NY.GDP.PCAP.KD). Gathered using the R package WDI and logged.
HDI	Quality of Government Standard Dataset, January 2019 version (Teorell et al. 2019). Variable undp_hdi, originally provided by the UNDP's Human Development Report.
Income inequality	World Bank (2019), GINI index, World Bank estimate (variable SI.POV.GINI). Gathered using the R package WDI.
Trade openness	World Bank (2019), trade as a percentage of gross domestic product (variable NE.TRD.GNFS.ZS). Gathered using the R package WDI.
Age of democracy	Boix et al. (2013), data version 3.0, variable democracy_duration.
Disproportionality	Gandrud (2019), variable disproportionality. Gathered using the R package devtools via `http://bit.ly/Ss6zDO`.
Party institutionalization	The Database of Political Institutions (Cruz et al. 2016), version DPI2015, variable partyage.
Clientelism	V-Dem, data version 7.1 (Coppedge et al. 2017). Variable v2psprlnks, inverted so that higher values indicate more clientelistic and less programmatic linkages.
Corruption	V-Dem, data version 7.1 (Coppedge et al. 2017). Variable v2x_corr.

Table A1: Data sources (continued)

Variable	Source
Government ideology	Chapel Hill Expert Survey, 1999-2014 Trend File, version 1.1 (Bakker et al. 2015). Variable seat divided by the sum of seat for a given country-year gave the legislative proportion for a given party, while lrgen gave the party's ideology. Parties in government were chosen using the govt variable, values "in government" or ".5" (in government for part of the year). We then imputed missing years for which CHES data were available. We then supplemented with Manifesto Project data, version 2018b (Volkens et al. 2018), using variable ideology and manually selecting parties in government using secondary sources. We then supplemented with data from Baker and Greene (2011), updated through 2018 in the 8 January 2019 data version. Here again we used variable ideology and manually selected parties in government using secondary sources.
% female legislators	Scraped from the Inter-Parliamentary Union website, now available through Parline (Inter-Parliamentary Union 2019).
Civil society	V-Dem, data version 7.1 (Coppedge et al. 2017). Variable v2x_cspart.
Pol. donation restrictions	V-Dem, data version 7.1 (Coppedge et al. 2017). Variable v2eldonate.
Trade union density	Trade union density rate (percentage), downloaded from ILOSTAT (International Labour Organization 2019) on 27 April 2019.
Compulsory voting	V-Dem, data version 7.1 (Coppedge et al. 2017). Variable v2elcomvot, recoded into a binary variable by setting all values greater than 1 to 1, to reflect any legal requirement to vote.
Cross-cuttingness	Data from Selway (2011), August 2013 version, variable RalC.
Turnout	V-Dem, data version 7.1 (Coppedge et al. 2017). Variable v2elvaptrn, divided by 100 so as to indicate proportions.

A.2 Details of the dependent variable

Our dependent variable is the Earth Mover's Distance (EMD; Lupu et al. 2017) between legislators and the least affluent quintile of citizens minus the EMD between legislators and the most affluent quintile of citizens. The EMD is computed as a measure of distributional distance wherein the object is to minimize the amount of "work" required to transform one distribution into another. Given two histograms and a distance metric, the EMD evaluates every possible mapping that would shift one distribution until it was identical to the other, and then finds the minimum total distance data would have to be moved across all of these mappings.

The EMD has several desirable properties. Most notable among them is that it captures the entire distribution of data and not just summary statistics such as the mean or median. It also better captures non-normal distributions than competing measures such as the difference in probability density functions. Lastly, it can be used to study distributional distance in multiple dimensions—in this context, to evaluate congruence across multiple issue-areas simultaneously.

The data for the EMD come from Lupu and Warner (Forthcoming). As described in that paper, each country-year relies on only one legislator survey to avoid the potential for non-response bias to be exacerbated: if only certain kinds of legislators respond to requests for their opinions, then duplicating that sample may decrease the representativeness of the legislator sample. To avoid this, Lupu and Warner (Forthcoming) use only the survey for which the fieldwork was most proximate to the year of the observation. Where there are multiple such surveys, those from large cross-national projects are prioritized for greater comparability.

Mass surveys are then matched to these legislator surveys. Both mass and legislator responses are scaled so that "left" or "liberal" is -1 and "right" or "conservative" is 1. Affluence quintiles are then constructed using variables relating to ownership of durable goods, income, or occupation. In country-years where mass respondents are asked a battery of questions relating to their ownership of things like cars, housing, or electronics, multiple correspondence analysis is used to generate a factored index of affluence. Where these variables are not available, self-reported income are used instead. Where neither are available, the authors code occupation into categories (e.g., "worker" and "white-collar professional").

The EMD is then computed between the least affluent quintile and legislators, as well as between the most affluent quintile and legislators, within each country-year. Positive values indicate that poor respondents are underrepresented relative to the rich, while negative values indicate the opposite. As indicated in the text, we only use country-years for which both the legislator and mass samples each have at least 30 respondents, since fewer respondents may indicate a non-representative sample and an unreliable measure of affluence bias.

A.3 Models used in the main analysis

The following list gives the name and description for each model studied in our machine learning task, as given in the R package caret (Kuhn 2008). We chose these models for their diversity of underlying approach.

1. avNNet: Model Averaged Neural Network
2. cforest: Conditional Inference Random Forest
3. dnn: Stacked AutoEncoder Deep Neural Network
4. glm: Generalized Linear Model
5. glmboost: Boosted Generalized Linear Model
6. glmnet: glmnet
7. knn: k-Nearest Neighbors
8. mlp: Multi-Layer Perceptron
9. nnet: Neural Network
10. pcaNNet: Neural Networks with Feature Extraction
11. ppr: Projection Pursuit Regression
12. rf: Random Forest
13. treebag: Bagged CART

A.4 Model performance

Table A2: Predictive performance

Model	RMSE	Model	RMSE
avNNet	0.077	mlp	0.086
	(0.008)		(0.013)
cforest	0.074	nnet	0.078
	(0.009)		(0.009)
dnn	0.086	pcaNNet	0.079
	(0.012)		(0.009)
glm	0.078	ppr	0.082
	(0.009)		(0.010)
glmboost	0.077	rf	0.074
	(0.010)		(0.010)
glmnet	0.077	treebag	0.075
	(0.010)		(0.009)
knn	0.079		
	(0.008)		

Values in parentheses indicate standard deviations. Note that RMSE is on the scale of the dependent variable, which ranges over $[-1, 1]$.

A.5 Additional partial dependence plots

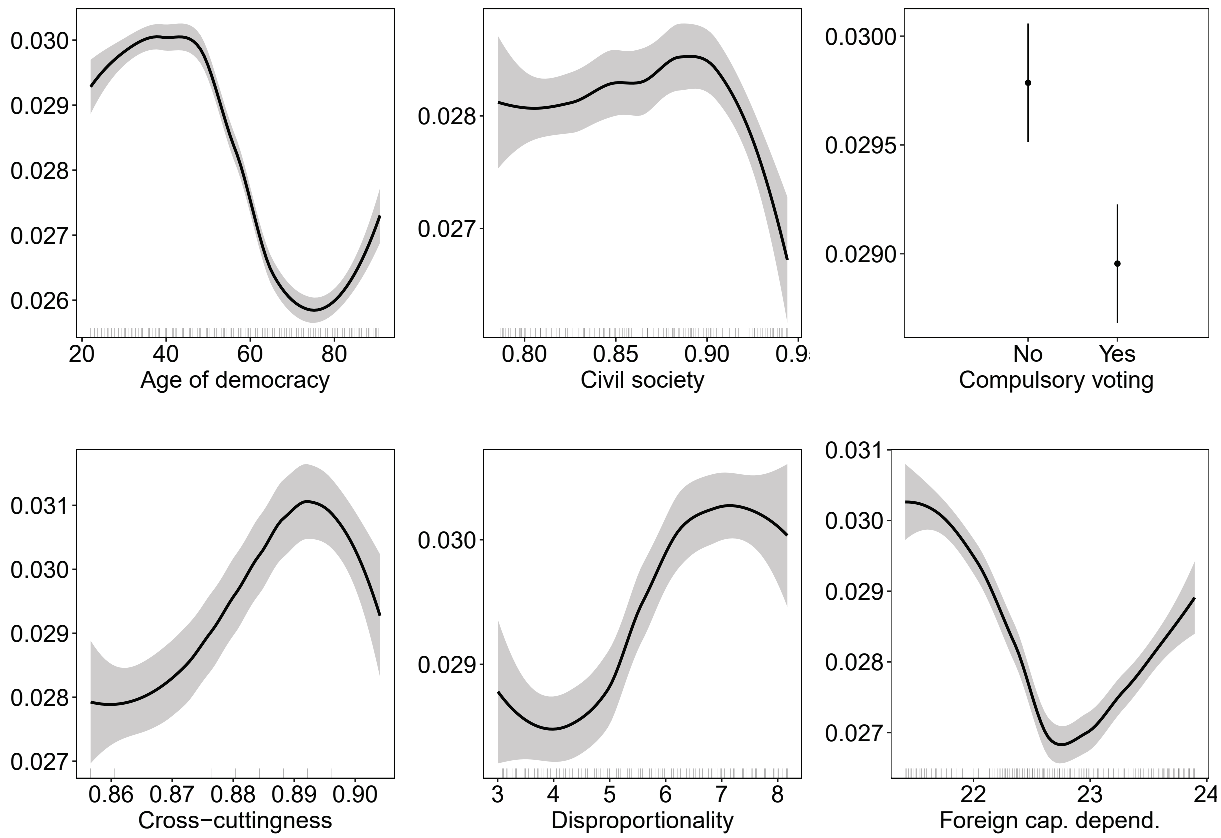

Figure A1: Partial dependence plots. Each panel provides the predicted change in unequal representation as a predictor is moved across its inter-quartile range. Lines represent loess fits, with 95% confidence intervals in gray, computed from random forest predictions across all imputation replicates. Rug plots are also provided along the x axis to indicate support in the underlying data for these predictions. Note the differing axes in each panel.

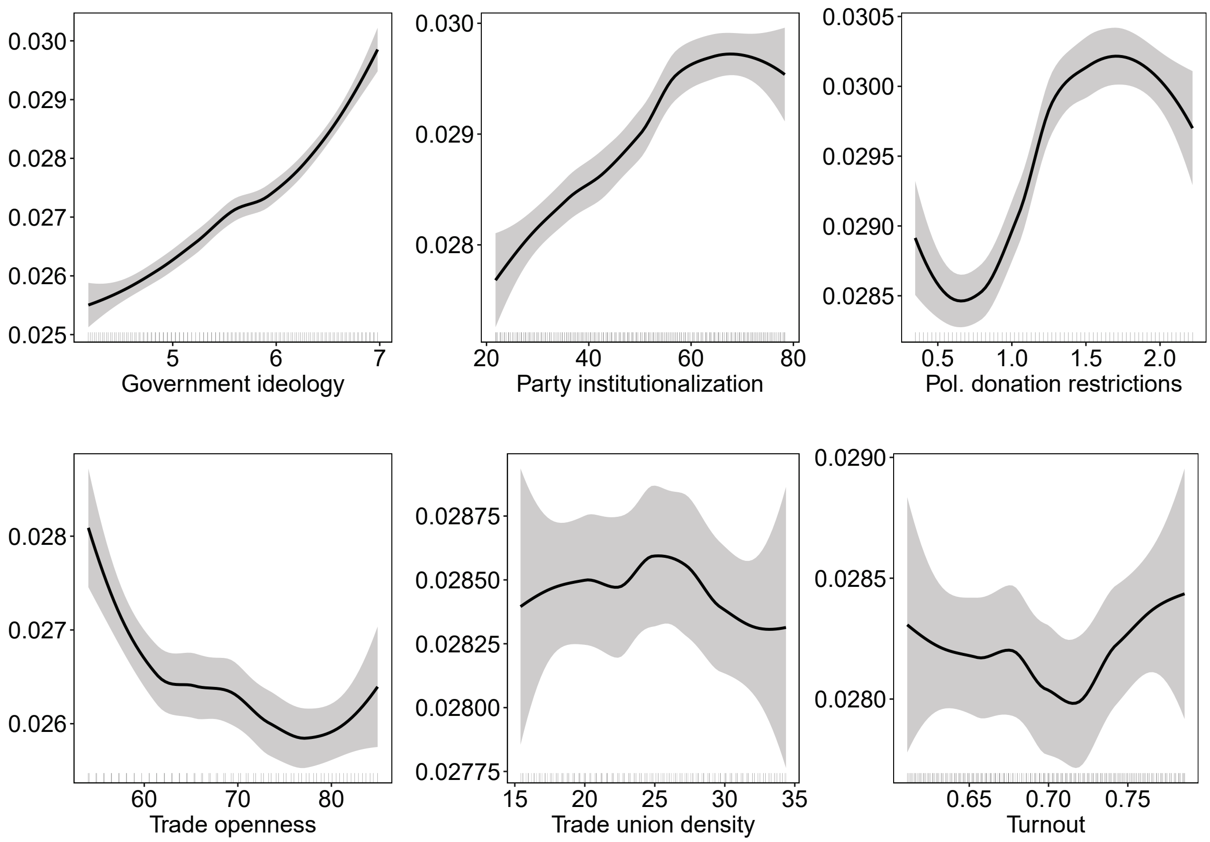

Figure A1 (continued): Partial dependence plots. Each panel provides the predicted change in unequal representation as a predictor is moved across its inter-quartile range. Lines represent loess fits, with 95% confidence intervals in gray, computed from random forest predictions across all imputation replicates. Rug plots are also provided along the x axis to indicate support in the underlying data for these predictions. Note the differing axes in each panel.

A.6 Listwise deletion

Table A3: Variable importance results under listwise deletion

Variable	Importance
Clientelism	100.00
% female legislators	97.50
Corruption	95.06
Party institutionalization	91.27
Income inequality	86.98
HDI	83.47
Foreign cap. depend.	66.47
GDP (logged)	59.43
Cross-cuttingness	56.47
Civil society	54.49
Turnout	52.67
Age of democracy	41.39
Pol. donation restrictions	39.10
Trade union density	35.56
Government ideology	21.16
Trade openness	11.07
Compulsory voting	0.00

Variable importance metrics are from the random forest model using listwise deletion instead of multiple imputation. Values are automatically scaled so that 100 indicates the most important variable and 0 indicates a variable that is not used for prediction. Note that disproportionality is dropped due to its high missingness; leaving it in causesbreaks the model fitting process.

References

Baker, Andy, and Kenneth F. Greene. 2011. "The Latin American Left's Mandate: Free-Market Policies and Issue Voting in New Democracies." *World Politics* 63 (1): 43-77.

Bakker, Ryan, Erica Edwards, Liesbet Hooghe, Seth Jolly, Jelle Koedam, Filip Kostelka, Gary Marks, Jonathan Polk, Jan Rovny, Gijs Schumacher, Marco Steenbergen, Milada Vachudova, and Marko Zilovic. 2015. "1999-2014 Chapel Hill Expert Survey Trend File.".

Boix, Carles, Michael K. Miller, and Sebastian Rosato. 2013. "A Complete Data Set of Political Regimes, 1800-2007." *Comparative Political Studies* 46 (12): 1523-1554.

Coppedge, Michael, John Gerring, Staffan I. Lindberg, Svend-Erik Skaaning, Jan Teorell, David Altman, Michael Bernhard, M. Steven Fish, Adam Glynn, Allen Hicken, Carl Henrik Knutsen, Joshua Krusell, Anna Lührmann, Kyle L. Marquardt, Kelly McMann, Valeriya Mechkova, Moa Olin, Pamela Paxton, Daniel Pemstein, Josefine Pernes, Constanza Sanhueza Petrarca, Johannes von Römer, Laura Saxer, Brigitte Seim, Rachel Sigman, Jeffrey Staton, Natalia Stepanova, and Steven Wilson. 2017. "V-Dem Country-Year Dataset v7.1." Varieties of Democracy (V-Dem) Project.

Cruz, Cesi, Philip Keefer, and Carlos Scartascini. 2016. "Database of Political Institutions Codebook, 2015 Update (DPI2015)." Inter-American Development Bank.

Gandrud, Christopher. 2019. "Gallagher Electoral Disproportionality Data." Available at http://bit.ly/Ss6zDO. Last accessed November 5, 2019.

Inter-Parliamentary Union. 2019. "Parline." Available at https://data.ipu.org/women-ranking. Last accessed November 5, 2019.

International Labour Organization. 2019. "ILOSTAT." Available at https://ilostat.ilo.org. Last accessed November 5, 2019.

Kuhn, Max. 2008. "Building Predictive Models in R using the caret Package." *Journal of Statistical Software* 28 (5): 1-26.

Lupu, Noam, Lucía Selios, and Zach Warner. 2017. "A New Measure of Congruence: The Earth Mover's Distance" *Political Analysis* 25 (1): 95 - 113.

Lupu, Noam, and Zach Warner. Forthcoming. "Affluence and Congruence: Unequal Representation Around the World." *Journal of Politics*.

Selway, Joel Sawat. 2011. "Electoral Reform and Public Policy Outcomes in Thailand: The Politics of the 30-Baht Health Scheme." *World Politics* 63 (1): 165 - 202.

Teorell, Jan, Stefan Dahlberg, Sören Holmberg, Bo Rothstein, Natalia Alvarado Pachon, and Richard Svensson. 2019. "The Quality of Government Standard Dataset, version Jan19." University of Gothenburg: The Quality of Government Institute.

Volkens, Andrea, Werner Krause, Pola Lehmann, Theres Matthieß, Nicolas Merz, Sven Regel, and Bernhard Weßels. 2018. "The Manifesto Data Collection." The Manifesto Project (MRG/CMP/MARPOR). Version 2018b. Berlin: Wissenschaftszentrum Berlin für Sozialforschung (WZB).

World Bank. 2019. "Data Catalog." Available at http://datacatalog.worldbank.org. Last accessed November 5, 2019.