Machine Learning Methods for "Small-n, Large-p" Problems: Understanding the Complex Drivers of Modern-Day Slavery

Rosa Lavelle-Hill (rlavelle-hill@turing.ac.uk)
The Alan Turing Institute https://orcid.org/0000-0002-1767-9828

Anjali Mazumder
The Alan Turing Institute

James Goulding
University of Nottingham

Gavin Smith
University of Nottingham

Todd Landman
University of Nottingham

Article

Keywords: slavery, survey data, statistics

Posted Date: March 4th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-296275/v1

Version of Record: A version of this preprint was published at Humanities & Social Sciences Communications on November 17th, 2021. See the published version at https://doi.org/10.1057/s41599-021-00938-z.

Machine Learning Methods for “Small-n, Large-p” Problems: Understanding the Complex Drivers of Modern-Day Slavery

Affiliation

40 million people are estimated to be in some form of modern slavery across the globe. Understanding the factors that make any particular individual or geographical region vulnerable to such abuse is essential for the development of effective interventions and policy. Efforts to isolate and assess the importance of individual drivers statistically are impeded by two key challenges: data scarcity and high dimensionality. The hidden nature of modern slavery restricts available datapoints; and the large number of candidate variables that are potentially predictive of slavery inflates the feature space exponentially. The result is a highly problematic “small *n*, large *p*” setting, where overfitting and multi-collinearity can render more traditional statistical approaches inapplicable. Recent advances in non-parametric computational methods, however, offer scope to overcome such challenges. We present an approach that combines non-linear machine learning models and strict cross-validation methods with novel variable importance techniques, emphasising the importance of stability of model explanations via Rashomon-set analysis. This approach is used to model the prevalence of slavery in 48 countries, with results bringing to light the importance predictive factors - such as a country’s capacity to protect the physical security of women, which has previously been under-emphasized in the literature. Out-of-sample estimates of slavery prevalence are then made for countries where no survey data currently exists.

Slavery was not eradicated with its ‘abolition’ in the 19th century; instead it has morphed in form, and continues as a phenomenon in all countries in the world [1]. Recent estimates have put the number of enslaved people at over 40 million [2], translating to 5.4 victims from every 1,000 people world-wide. Modern slavery is highly challenging to identify and measure, defined as “any situation of exploitation that a person cannot refuse or leave because of threats, violence, coercion, deception, and/or an abuse of power” [2] - situations that include forced labour, debt bondage, forced marriage, sexual exploitation, bonded labour, and human trafficking [1]. In order to address this global challenge and meet United Nations Sustainable Development Goal (SDG) 8.7, it is crucial to not only to find new ways to measure prevalence [3, 2, 4], but to understand better the conditions which allow its continued perpetuation. We must isolate and address the core, driving factors that are leaving particular individuals, regions, or countries at risk.

Estimating Slavery Prevalence

Estimating the prevalence of slavery, of course, remains a central task. National and regional estimates not only highlight the extent of the issue, but serve as dependent variables in analyses which attempt to model, and hence find explanations of, slavery’s underlying causes [5]. However, the hidden nature of modern-slavery makes it an intrinsically difficult to measure [6] and there remains much uncertainty around the *true* number of people enslaved [7]. In recent years, the Walk Free Foundation (WFF) has made valuable progress in this area, providing estimates of slavery incidence across 48 countries in 2016 and 2018, based upon surveys from the Gallup World Poll (GWP). These estimates have been further extrapolated out-of-sample to countries where no GWP survey data existed using theoretically-driven risk models [8, 2]. While this has proven a highly beneficial exercise [9, 10], methods used to produce such estimates are not without their limitations [7, 10, 11, 12, 13, 14, 15]; and despite recent advances in Multiple Systems Estimation (MSE) [3, 16, 17, 18], assessment of national slavery prevalence to any degree of accuracy remains an active research area.

A further problem in modelling slavery is that, again due to its hidden nature, there exists an extensive range of candidate independent variables to investigate, while at the same time a shortage of prevalence estimates to regress against. Therefore, aggregated data used to model slavery prevalence in a population will likely be “small *n*, large *p*” in nature [19]. This issue is symptomatic of many “wicked problems” [20] facing computational social sciences, with too few experimental units, *n*, available to researchers in comparison to the vast number of potential driving factors, *p*, to be modelled. In such situations, dangers of overfitting and multi-collinearity can render traditional statistical approaches inapplicable - and as a result, investigation of the factors underlying modern-slavery remains a challenging statistical task.

Identifying Drivers of Modern-day Slavery

Due to the "small *n*, large *p*" problem, prior research exploring the factors underlying slavery has predominantly been theoretically driven, with social science literature reporting individuals most vulnerable to exploitation as being: economic migrants; political asylum seekers; illiterate [21] orphaned children [22]; homeless people [23]; and the jobless [8]. All categories pertain to those who either have poor well-being [8] or acute poverty and/or significant debt, and hence seek a better way of life [21]. Children, estimated to make up a quarter of all victims [1], have also been identified as being particularly vulnerable to coercion through fake promises of work, food, or 'western' lifestyles [24]. It is well understood that "trafficking thrives better on willingness", with traffickers targeting areas where individuals are most vulnerable, desperately poor and with few options available [24].

At a national level, higher slavery incidences have been linked to both low GDP and higher levels of corruption [25]. Poverty neither equates to slavery, nor makes it inevitable and there are many poor areas where slavery is rare [26]. However, poverty can leave people in desperate circumstances, and is a well-established factor in making individuals easier for traffickers to coerce [27]. Slavery and human trafficking in turn undermines local and national economies, lowering GDP further [28], making cause and effect harder to discern. The impact of armed conflict on slavery is also well reported in the literature, with its effects being both direct and indirect. Men, women, and children have been abducted and forced to serve as soldiers, porters or smugglers of looted goods, forced labourers in military camps, and sex slaves for militia officers [21, 29]. It is estimated that over 120,000 children have been used in armed conflicts in Africa alone [29], with battle deaths also leaving children orphaned and vulnerable [21]. Destruction of infrastructure and societal systems leaves populations isolated and unprotected, making conflict-torn areas easy targets, where crimes are rarely investigated.

At higher regional levels, areas such as Africa and Asia have been linked to higher prevalence of slavery [2]. This is thought to be caused by a number of factors including: geographic position with regards to trade routes and smuggling channels [24]; occurrence of natural disasters destroying livelihoods and displacing populations; diseases such as HIV and AIDS increasing the number of orphaned children [30, 22]; and desertification and rising sea levels causing famine [31]. The sheer range of factors emphasizes the complexity of slavery as a social problem, stemming from the multitude of possible causes interacting with one another.

Quantifying the Effects of Drivers

Despite a rich literature considering the drivers of modern slavery, most findings stem from qualitative case data, victim/survivor interviews and small scale surveys. While a depth of insight is obtainable from such methods, findings are necessarily drawn from small samples, limiting generalizability to specific regions, industries, and forms of slavery. Few of the predictors hypothesised, and detailed in the previous section, have been studied statistically, restricting our understanding of their impacts and interactions. As a consequence, not enough is known about the extent to which any individual driver engenders slavery incidence; nor how factors combine to allow the phenomena to perpetuate, limiting the formation of informed national policy.

There are a few recent exceptions; researchers have used national prevalence estimates from GWP surveys [2] to statistically consider relationships between slavery and specific phenomena such as fishing [32] and globalisation [33]. The WFF have also established a theory driven Vulnerability Model [34], to summarize factors impacting national prevalence. The WFF framework groups 23 national-level risk variables into five major dimensions: Governance Issues; Lack of Basic Needs; Effects of Conflict; Inequality; and Disenfranchised Groups [34], and allows a vulnerability score to be formulated for each country using its scores on each dimension. However, the extent to which individual factors predict slavery is not reported, nor is there indication of how factors relate to one another. The WFF framework currently drops variables when collinearities occur (e.g. Gender Inequality Index [34]) and the framework's combined vulnerability score (which considers all 5 dimensions together) shows only moderate correlation with country prevalence estimates (r=0.33) [8].

Other quantitative studies have predominantly restricted themselves to linear hypothetico-deductive approaches [32, 33], focusing on single or small sets of independent variables. While this is an understandable situation given limited data, it has left an open challenge of assessing the problem domain via an inductive and computational approach. Exploration of slavery drivers requires machinery that can accommodate the high-dimensional and non-linear nature of modern slavery, while also accounting for interactions and collinearities between explanatory variables. In this work, we extend WFF's valuable ground-work in the field, introducing a computational model able to directly map relationships between vulnerability indicators and national slavery prevalence. Our explicit target is to quantify the importance of individual factors in predicting slavery prevalence, despite a "small *n*, large *p*" context. Central to the approach is the first application of Rashomon-set analysis [35] to the issue, a technique which allows assessment of the stability of model explanations and interactions across individual variable importances.

Overcoming the Limitations of Traditional Approaches

Studies capitalising on the national prevalence estimates represent an important step forward in understanding the statistical relationships between variables that may affect slavery. Yet, the insights that can be gained from the linear regression/correlation analyses predominantly used are limited by three key factors to overcome: issues of (1) multicollinearity; (2) assumed linearity and (3) overfitting. These challenges respectively impact on variable importance analyses, overall explanatory-power of different models, and the generalisability of those models:

Multicollinearity: In a traditional regression model predictors need to be independent if β coefficients are to be used as reliable estimates of variable importance [36]. Due to multicollinearity the WFF, for example, excludes several variables from its factor analyses despite "conceptual gaps that were potentially addressed by their inclusion" [34]. Omitted variables in this instance included: political/civil rights, wages, literacy, child mortality, corruption, GDP, government effectiveness, and gender inequality [34], many of which have been highlighted as important in the literature. As will be shown, omitting variables with partial collinearities can not only impact model accuracy, but can alter the explanations for slavery produced. Such variables can be involved in non-linear interactions, interpretation of which may offer benefits to policy. Recognising and understanding the effects of such collinearities in a model, rather than simply discarding them, is hence a key focus.

Assumptions of linearity: Traditional regression analyses do not readily reveal non-linear dependencies, tipping-point thresholds, nor sub-population effects despite such aspects being important issues when forming and implementing policy interventions. If addressing some key factor (e.g. education of women) predominantly impacts a particular context only (e.g. a specific age group), its importance can be lost within linear models - and insufficiently reflected via analysis of β coefficients. In domains such as slavery this is an important consideration, with the literature clearly identify variation in the relevance of particular drivers in different geographical regions. In Northern African countries, for example, relative AIDS prevalence is suggested as a strong candidate predictor of slavery, due to the number of children orphaned by the disease [22]; in Western countries this relationship would be completely unanticipated, given the vastly more-developed social and health services. Standard regression approaches are not able to delineate non-linear contexts of this nature, despite their common occurrence in the domain.

Overfitting: Performing traditional regression analysis on small data carries significant risk of model over-fitting [37]. This is a critical issue, with models that appear to provide good 'fit' to data, offering no guarantees of out-of-sample generalizability in reality. Such cases undermine the efficacy of any explanations emerging from models; and this adds risk to any real-world interventions formed from them. This issue is exacerbated in high-dimensional settings where multiple variables are modelled simultaneously [37]. Partly, this may be why hypothetico-deductive approaches remain so prevalent in the social sciences, with studies tending to focus on a few key hypotheses in order to prevent the increase of family-wise error rates. While this does indeed help to prevent overfitting, the ability to uncover new, and perhaps unexpected, predictors is often lost. Furthermore, the ability to assess relative variable importances across a wide range of features is foregone, with interactions between variables left unmodelled - a situation which is particularly problematic to settings such as slavery, where driving factors are unlikely to act in isolation.

In response to issues of multicollinearity, non-linearity and overfitting, the field of machine learning has sought to develop alternative ways to guard against model overfitting, while accommodating both non-linear and multivariate data. This has produced a rich array of *cross-validation* techniques, methods which support inductive experimental setups whilst defending against p-hacking and 'procedural overfitting' [37]. Cross-validation aims to find model parameterisations that maximise generalizability, rather than minimizing regression residuals (whilst additionally reducing the influence/bias of the researcher in model specification and model selection phases). Due to their inductive nature, such frameworks permit discovery of potential new predictors - and crucially allows comparative assessment of the importance of variables, even in non-linear settings.

Cross-validation was, however, designed with large datasets in mind. If *stable* model interpretations are to be found using smaller datasets, we require additional considerations and alterations to a typical machine learning methodology. The steps required to achieve this our outlined in detail in the Method section, and centre on integration of feature compression/selection directly into the modelling process, rather than being undertaken *a priori* as has been common in the field. Reduction of the variable space at some point is unavoidable, if the "large p" problem [38] is to be handled. A set of k, potentially latent, variables must be identified, where $k < p$. Unlike methods such as *partial least squares*, which identify such latent variables via parametric modelling assumptions, our approach treats k as a hyperparameter to be identified as part of model class exploration. To traverse the model space, M, we employ full, leave-one-out cross validation (LOOCV), allowing model parameters to be optimized, whilst using all of the "small n" data. This further ensures that insights extracted from fitted models will generalize to the whole dataset as much as possible.

With a best performing compression strategy and model parameterization, $\hat{M} \in \mathcal{M}$, having been identified via this method, a final issue remains. The potential for multicollinearlity in the data increases the likelihood that multiple *equally* well-performing models exist - models which have (almost) equivalent prediction accuracy to $\hat{M}$, but which leverage their independent variables in different ways. To interpret a single model, and assume its internal predictive mechanisms are fully representative of the phenomena being analysed, is not credible in "small *n*" situations. To remedy this 'instability', and to increase confidence in the driving factors isolated, we therfore employ a *Rashomon-set analysis* [39, 40, 41]. All well-performing model solutions (i.e. those within an 'epsilon-threshold' of the best model's prediction accuracy) are fit to the full dataset. Variable importance methods can then be applied to all models in this Rashomon set, allowing analysis of variable importance volatility and shedding light on both masking and interaction effects across predictors. Together, these steps allow for meaningful, potentially novel insights to be extracted via a non-linear, inductive approach even in "small *n*, large *p*" settings.

Study Overview

In this study, we apply the "small *n*, large *p*" workflow described in the previous section to model the prevalence of slavery over 70 country-year datapoints. Data is drawn from published and freely available datasets, and is comprised of 106 independent variables gathered from open-source data (covering a range of economic, socio-demographic and contextual indicators) with a single dependent variable depicting slavery prevalence (derived from GWP surveyed data) [33]. Data modelling occurs via a strict LOOCV approach, seeking model parameterizations and identification of latent variables/factors, that minimize mean absolute error (MAE) on hold-out data. The primary goal of the study is to understand better the driving factors of modern-slavery, whilst avoiding assumptions of traditional linear methodologies predominately used in the field to date. Use of this inductive methodology provides the opportunity for new predictive features to emerge. We evaluate the utility of this approach via comparison to a (i) pre-selection of features via theory, (ii) use of full, raw feature sets with contemporary statistical methods. As well as allowing insight into drivers of slavery, our final model is also used to generate new out-of-sample estimates of slavery prevalence for countries where no survey data exists.

Results

Model Selection

The model selection phase of our analysis considered a range of candidate features spaces: the full feature set (all 106 national indicator variables), a subset of 34 variables reported in the literature as relating to modern-slavery, and a range of compressed feature spaces (expressing between 2

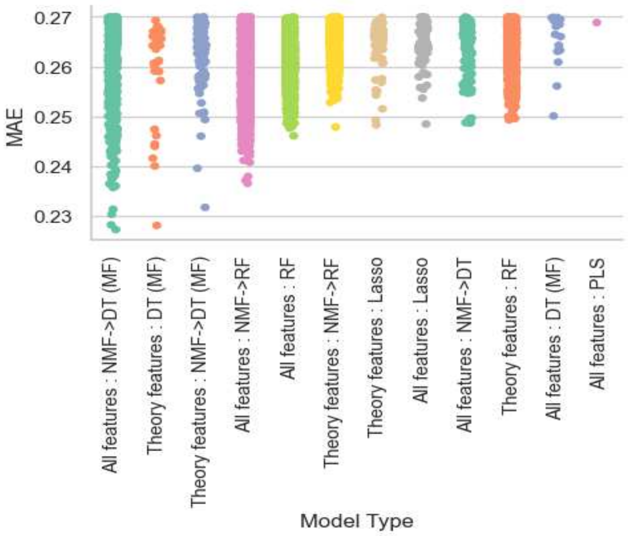


Figure 1. The LOOCV performance of all models where MAE was less than 0.27.

Table 1
LOOCV performance of the different pipelines

Feature Selection	Feature Compression	Model Class	MAE
Full feature set	NMF	Linear Regression	0.266
Full feature set	NMF	Decision Tree	0.241
Full feature set	**NMF**	**Decision Tree (MF)**	**0.227**
Full feature set	NMF	Random Forest	0.237
Full feature set	Partial Least Squares (PLS) Regression		0.269
Full feature set	No compression	Lasso Regression	0.248
Full feature set	No compression	Decision Tree	0.322
Full feature set	No compression	Decision Tree (MF)	0.250
Full feature set	No compression	Random Forest	0.246
Theory-based	NMF	Linear Regression	0.291
Theory-based	NMF	Decision Tree	0.282
Theory-based	NMF	Decision Tree (MF)	0.232
Theory-based	NMF	Random Forest	0.248
Theory-based	Partial Least Squares (PLS) Regression		0.284
Theory-based	No compression	Lasso Regression	0.248
Theory-based	No compression	Decision Tree	0.293
Theory-based	No compression	Decision Tree (MF)	0.228
Theory-based	No compression	Random Forest	0.249

Note: MAE = Mean Absolute Error. NMF = Non-negative Matrix Factorisation (chosen for it's interpretability). 'Theory-based' refers to feature selection based on the literature (N=34). MF refers to allowing the 'max features' hyper-parameter to be tuned, so that the decision tree could not select from all components at each split. A Lasso regression (L1 norm regularisation) was used instead of a linear regression when no feature compression was used due to the the high number of independent variables.

Table 2
The Rashomon set

Model Name	MAE	Features	Model Class	K Components
Best Model	0.227	All features	NMF->DT (MF)	6
Rashomon 1	0.228	Theory-selected	DT (MF)	
Rashomon 2	0.228	All features	NMF->DT (MF)	6
Rashomon 3	0.230	All features	NMF->DT (MF)	6
Rashomon 4	0.231	All features	NMF->DT (MF)	6
Rashomon 5	0.232	Theory-selected	NMF->DT (MF)	5

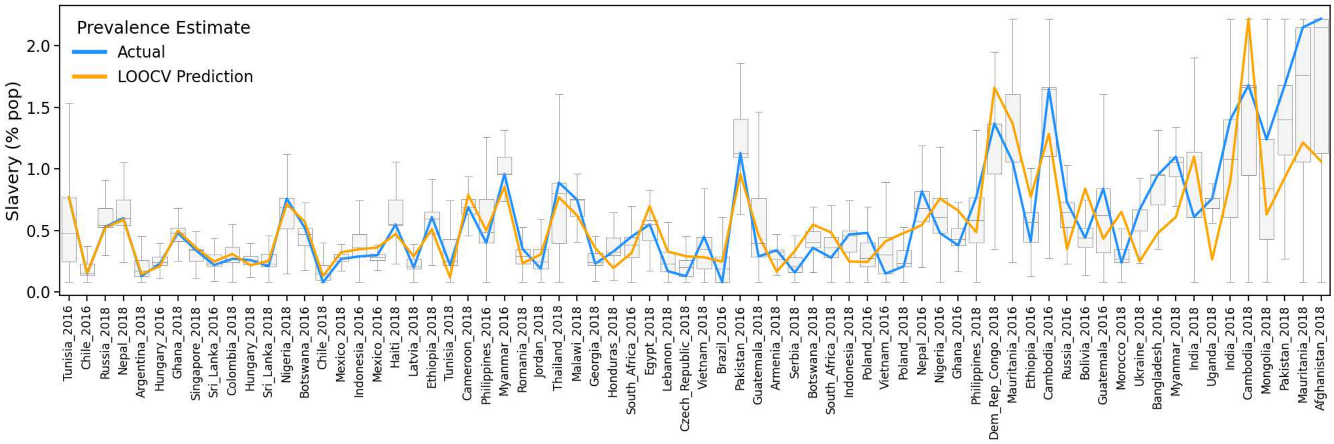


Figure 2. The predictions of slavery prevalence (individuals enslaved as a % of the population) made by the best model using leave-one-out cross validation (LOOCV), compared to the 'actual' prevalence as estimated using the Gallup World Poll (GWP) survey data. The grey box plots illustrate the distribution of 10,000 bootstrapped LOOCV predictions (using the full pipeline NMF->DT) to help illustrate the uncertainty associated with our model's predictions. The box shows the quartiles of the bootstrapped predictions while the whiskers extend to show the rest of the distribution, except for points that were determined to be "outliers" (using a function of the inter-quartile range) which are not plotted. The x axis is ordered by the MAE.

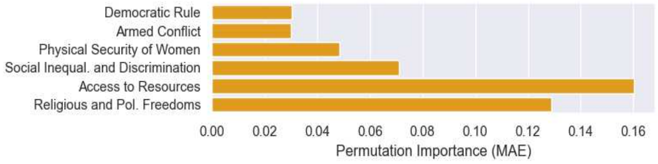


Figure 3. The permutation importance [42] of the NMF components (identified latent variables) in the best performing model.

and 8 components). This meant that as well as testing different model classes, the utility of feature-selection based on theory and different feature compression using Non-negative Matrix Factorisation (NMF) could also be tested. NMF was selected to identify latent variables due to ease of interpretation in comparison to other compression methods, such as PCA. NMF parameterizations (k) were explored as part of the grid search, meaning the interactions between feature compression and model class parameterizations were also captured. This allowed identification of optimal latent variable compression for each predictive mechanism/pipeline. The model classes explored ranged from linear and regularised regression based models, to non-linear approaches such as decision trees and random forest models (also chosen due to high interpretability). Partial least squares was also examined due to its internal identification of latent structures. The best preforming parameterization for each of these modelling strategies pipelines is detailed in Table 1, with Figure 1 illustrating the distribution of performances for all models with MAE less than 0.27.

The best performing model used the full feature pool ($p = 106$), compressed the feature space via latent NMF components ($k = 6$), then used a (non-linear) decision tree model with a restricted number of 'max features' (MF) available at each split (to help prevent the tree from getting stuck at a local minimum). For the full set of model parameters see Appendix B. The best model's performance was a significant improvement on both the mean (MAE= 0.366, p<0.001) and median (MAE= 0.349, p<0.001) baseline predictions analysed using a Wilcoxon Signed-Rank t-test.

Predictions using leave-one-out analysis compared to the 'actual' slavery prevalence estimations can be viewed in Figure 2, accompanied by the distribution of 10,000 bootstrapped predictions to illustrate the associated uncertainty. The graph highlights that the model was less effective at predicting a subset of countries, and in particular those with the highest prevalence of slavery, which it tended to underestimate[1].

Model Interpretation

In the model interpretation phase, the best model, $\hat{M}$, was fit to the data, and resulting NMF components and variable loadings analysed to determine component themes (see Appendix, Figure A3 for a breakdown of the model). Components' relative importances were then compared us-

[1] In this particular analysis we chose not to add a weighting to the high estimates, but recognise that this might be appropriate when using the model as a tool to identify areas at high risk.

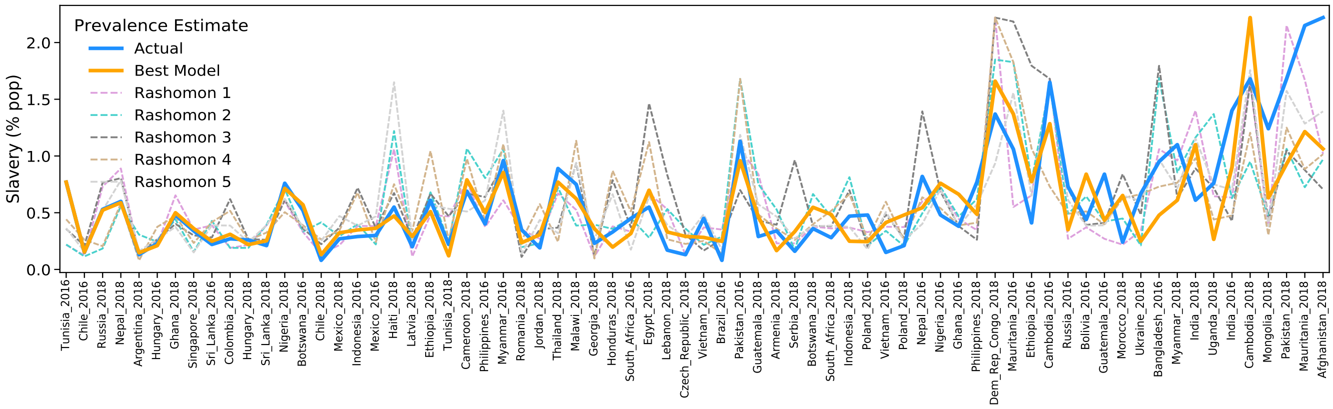


Figure 4. The predictions of slavery prevalence (individuals enslaved as a % of the population) made by the best model and the five other well performing models in the Rashomon set (see Table 2). The x axis is ordered by the MAE between our best model and the 'actual' prevalence as estimated by the GWP survey data.

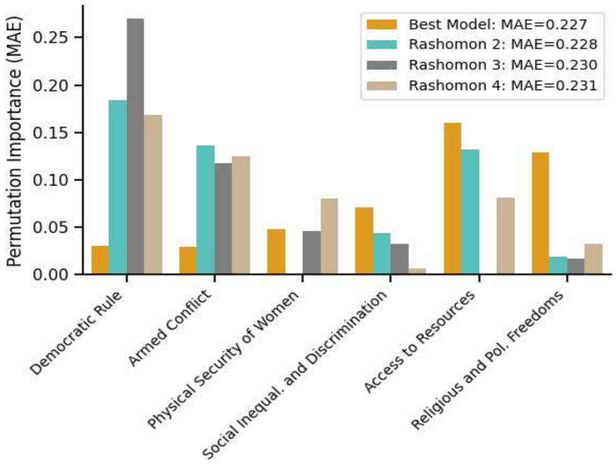


Figure 5. Comparison of variable importance ratings all models in the Rashomon set, omitting those in the same model class as the best model (NMF->DT (MF).

355 ing permutation importance [42]. Finally, a group of similarly performing models that constructed a *Rashomon set* [39, 40, 41] were utilised to investigate whether insights converged across multiple well performing models, or whether alternative explanations exist.

360 For $\hat{M}$, the NMF stage reduced the 106 different variables to 6 latent components, interpretable as: Democratic Rule, Armed Conflict, (lack of) Physical Security for Women, Social Inequality and Discrimination, Access to Resources, and Religious and Political Freedoms (specific variable loading's can be viewed in Appendix, Figure A2). Interestingly, the component themes that emerged closely matched the vulnerability factors used by the WFF in their vulnerability model, with one notable exception: the addition of a component focusing specifically on the *physical security of women*. The component permutation importance for $\hat{M}$ was also calculated, with Figure 3 providing a comparison of the components' importance, and illustrating that access to resources was identified as the most important predictor of national slavery in the model.

375 This, however, only reflects the viewpoint of a single model. To test the stability of the insights produced, further analysis was performed using a Rashomon set approach [39]. Risk of model instability is relatively high in "small n, large p", and can be exacerbated when non-parametric approaches such as decision trees [43] are employed to handle multicollinearity between the components. To deal with this it is therefore necessary to examine whether other well per-

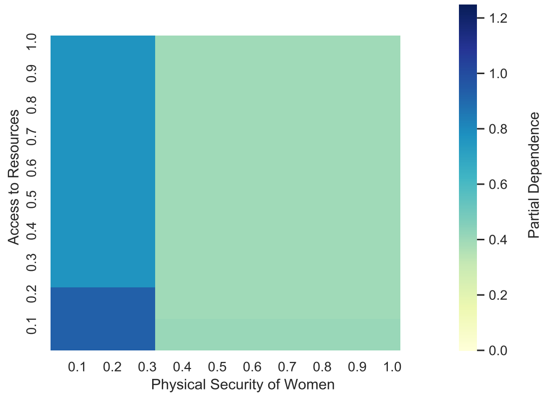


Figure 6. A heat map illustrating that (for the best performing model) the partial dependency of the prevalence prediction is especially high when both Access to Resources and Physical Security of Women are low. Here, a lower score for Physical Security of Women indicates less security.

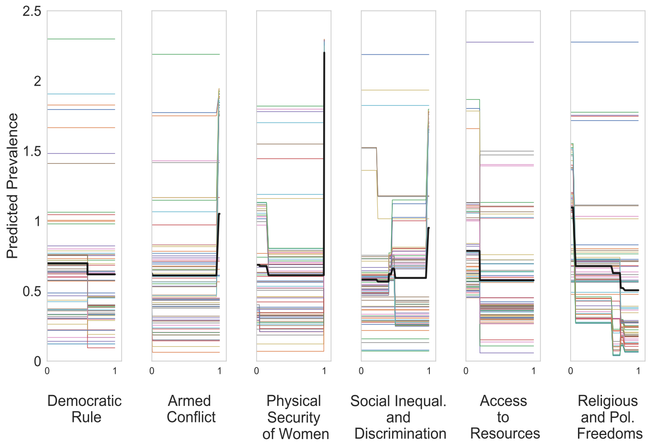


Figure 7. Individual Conditional Expectation (ICE) plots to illustrate the non-linear effects of components on the best model's predictions. Each coloured line is a country-year data point with some jitter applied. The thick black line is the average partial dependency of the component on the prevalence prediction.

forming models gave converging or diverging explanations (referred to as the *Rashomon effect* [44]). In this case, the Rashomon set was defined as any model whose MAE performance was within 2.5% of the best performing model (< 0.233)[2]. Models retained in the Rashomon set are outlined in Table 2, and feature: three other models with the same pipeline as the best model (using NMF->DT (MF) on all the features); a DT (MF) using just the theory features, and a final NMF->DT (MF) which used only theory-selected features. The disagreements in predictions made across the Rashomon set can be viewed in Figure 4.

Despite, NMF models utilizing different random seeds for initializing coordinate descent and different coefficients for the regularisation terms (alpha), K=6 emerged as optimal across all pipelines on the full feature set. Corresponding NMF themes remained predominantly stable across each model (see Appendix, Figure A4), providing strong evidence for the applicability of the six latent variables identified.

Within model's sharing the same class as $\hat{M}$ (NMF->DT (MF), all features), the permutation importances of components were compared across models, in order to understand how other competing solutions compared to the explanation provided by $\hat{M}$. This comparison can be viewed in Figure 5. The graph illustrates the different predictive strategies models used following feature reduction. The variation across models highlights the risk of over-interpreting insights from a single model. For example, in the best performing model, $\hat{M}$, Democratic Rule and Armed Conflict are relatively unimportant predictors of slavery prevalence, yet are the most important in Rashomon models, M_2, M_3 and M_4.

Subsequent analysis of components (see Appendix, Figure A5) indicate that, despite their stability, some multicollinearity does remain between them. In particular, Religious and Political Freedoms is negatively correlated with Armed Conflict (r=-0.62); and (lack of) Physical Security for Women is negatively correlated with Access to Resources (r=-0.44). The effect of the latter relationship is particularly apparent in Figure 5, with Rashomon models, M_2 and M_3 requiring only one of those components to perform well. This suggests that there may additionally exist non-linear relationships between the two components, which cannot be fully captured in the correlation coefficient.

To understand the unanticipated relationship between Access to Resources and the Physical Security of Women more fully, partial dependency plots were analysed to look for nonlinear dependencies between the variables. Figure 6 illustrates that the physical security of women is only predictive of slavery in contexts where access to resources is low. In other words, women are particularly vulnerable to being exploited in areas where there is poor access to fuel, electricity, piped/clean water, sanitation, and education (see variable loadings of the Access to Resources component in Appendix, Figure A2). To further highlight the non-linear interactions behind the Physical Security of Women component, Individual Conditional Expectation (ICE) graphs shown in Figure 7 indicate that when a (lack of) physical security for women is extremely high, this dramatically increases the partial dependency of the prevalence prediction - more than any other component. The important role that Physical Security of Women can play in the prediction of slavery prevalence is only made available here using a methodology that allows for non-linear interactions of this nature to be modelled.

Predicting prevalence for countries with no survey data

As a final stage of analysis we present new estimates for countries where no GWP survey data has been collected, projecting the best performing model out-of-sample in order to generate new estimates of prevalence. The model was fitted to the 70 country-year data points over the years 2016 and 2018 for which survey data were available. Predictions are output for the 172 countries in 2018 for which no survey data exist. Figure 8 shows the estimates produced compared to the estimates made by the model used in the 2018 edition of the Global Slavery Index (GSI) [2]. The GSI used a hierarchical Bayesian linear model with additional adjustments (see [2, 8] for full methodological details).

These out of sample predictions constitute the best predictions within the parameters of our grid search, noting that

[2] While selecting an epsilon-values of this nature when generate Rashomon set's remains a subjective task, Figure 1 illustrates the visible gap in performance between the set of models included in comparison to competitors.

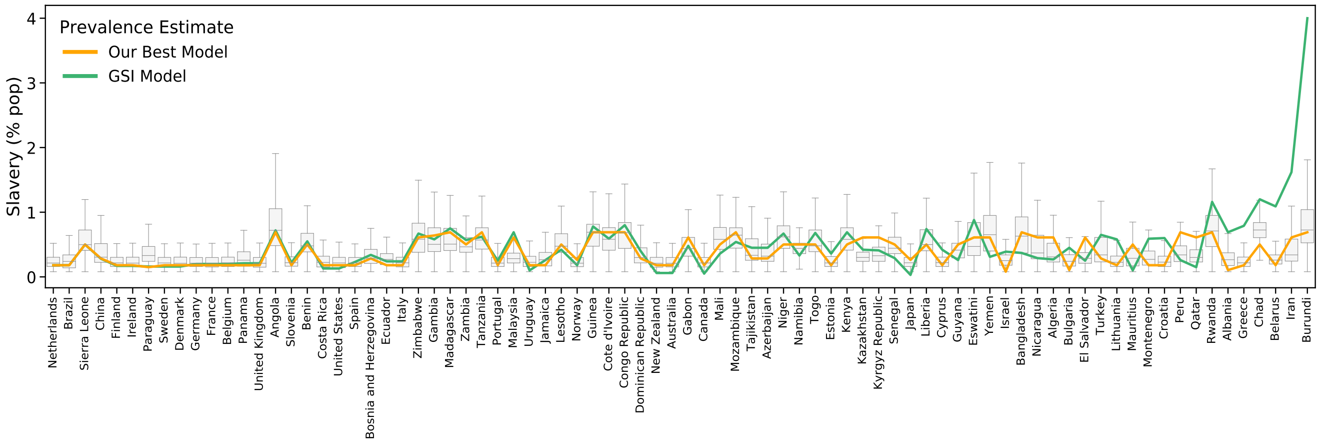


Figure 8. The predictions made for the 2018 prevalence of slavery (individuals enslaved as a % of the population) in countries where no GWP survey data exists. The estimates made by the best performing model are compared to those made by the GSI. The grey box plots illustrate the distribution of 10,000 bootstrapped LOOCV predictions (using the full pipeline NMF->DT) to help illustrate the uncertainty associated with our model's predictions. The box shows the quartiles of the bootstrapped predictions while the whiskers extend to show the rest of the distribution, except for points that were determined to be "outliers" (using a function of the inter-quartile range) which are not plotted. The x axis is ordered by the disagreement between our model's predictions and the predictions made by the GSI model. The countries displayed are those which had <10% missing independent variable data. For all 172 out-of-sample predictions and information on missing data see Appendix, Table A1.

that our parameter and model choices were influenced by the goals of the investigation - improved *understanding* of the predictors of slavery. It is recognised that the goals of explanation and prediction do not always align [45, 46, 37]. Model selection and parameter searches in our implementations were guided by a need for model interpretability (e.g. use of linear latent variable structures), and it remains possible that predictions could be improved through the use of other models (i.e. black box methods) or data sampling techniques (such as taking the mean of the bootstrapped predictions). Nevertheless, considering the estimates produced by computational models with well-established and understood mechanisms, establishes not only a) interpretable estimates, but also b) whether different models with different features diverge/converge in their predictions, and c) what further data/features might improve prediction.

Discussion

This study, for the first time, presents an inductive machine learning methodology and contemporary variable importance analysis to understand better the complex predictors of modern slavery in the real world. By evaluating multiple different pipelines the utility of theory-driven feature selection versus guided feature compression (to help deal with the "large p" problem) has been tested, as well as comparing the performance of non-linear models versus their more traditional linear counterparts. Generally, and with but few exceptions, models accessing all features performed better than those using features selected from the qualitative literature. Given the generous selection of variables entered into the theory-selected feature pool this highlights the potential of an inductive methodology to uncover novel predictors, even in "small *n*, large *p*" contexts. Model class comparisons unequivocally showed that non-linear models gave better predictions than their linear counterparts, supporting the notion that our models are capturing new non-linearities that have not been analysed before. Finally, allowing guided feature compression, parameterizable as part of a grid search, uniformly outperformed traditional approaches that incorporated latent structures, such as partial least squares.

The majority of the models in our Rashomon set benefited from identification of latent components. This broadly validates the utility of summarising predictors of slavery into *k* latent factors (first done by the WFF in the construction of the Vulnerability Model [34]). Whilst the final model decided on by the WFF consists of 5 factors, the naturally occurring solution (based on eigenvalues greater than 1) actually consisted of 6 factors [34], corroborating our bottom-up findings. The present study, however, extends the WFF's initial work by constructing components directly shaped by the outcome variable, slavery prevalence.

Importantly, our analysis highlights the additional contribution of a new component, *Physical Security of Women*, reflecting the onus on a country's law enforcement to protect woman from domestic violence, rape and sexual assault, marital rape, shame/honour killings or femicides [47]. The

Physical Security of Women has been previously been overlooked, due to the non-linear effect it has on the prediction of slavery, in combination with its interactions with other components. Most notable is the finding that not only is there is a greater vulnerability to women's physical security in areas lacking access to resources, but this in turn is a particularly strong indicator of slavery occurrence (including the sexual exploitation and forced marriage of women).

In previous models, gender inequality features have had to be removed because of issues with multicollinearity [34]. Our methodology, which does not require the removal of correlated variables, demonstrates that gender inequality is likely a core piece of the puzzle in predicting national slavery figures. In addition to the physical security of women being a predictive component, variables depicting either the reporting or prevalence of rape also load highly onto two other components (see variable loadings in Appendix A2). This highlights that when explanation is the goal, it can be important for the researcher not to remove features that are correlated *a priori*, and instead navigate issues of multicollinearity within the modelling so that a more complete explanation can be achieved.

Rashomon set analysis highlighted that despite the latent components found to be stable, there existed a high degree of variability in the importance ratings models assigned to each. All components were used to differing degrees in each Rashomon model, emphasizing the care that must be employed when using machine learning not just for prediction, but to understand underlying factors. There is danger in focusing on a single model for interpretation - even if the model performs well, misleading explanations can be drawn, particularly in the presence of multicollinearity [40] and when there is degree of uncertainty around the accuracy of the data (see limitations below).

Finally, this study leveraged machine learning to generate new estimates of slavery prevalence. These estimates are of course useful unto themselves in understanding both the extent of slavery, and surrounding uncertainty. Additional insights, however, can also be obtained by reflecting on the differences between the estimates produced by different models/methodologies. In contrast to the data driven approach used here, the WFF approach we considered in comparison, involves multiple adjustments and does not solely reflect the information held in independent variable data [2]. This prompts the question of what additional information or intuition is not currently being captured in data and what would be valuable data for researchers to have. Understanding these data gaps, as well as when and why certain estimates don't align with experts' expectations, will be crucial for advancing the measurement of modern-day slavery forward. Future efforts might focus on how human judgement and computational modelling can be combined to build prediction models - harnessing the advantages of a data driven approach (objectivity, quantification, and out-of-sample predictions) combined with expert human judgement and additional contextual information.

In this study, and while key limitations of traditional regression approaches were mitigated against, several limitations remain inherent to the data. The dependent variable, although derived from survey data collected using a representative sampling methodology [48], nonetheless represents an estimate rather than a direct measure of slavery. Consequently, prevalence estimates must be considered more a reflection of slavery risk across the given the population, rather then corresponding to formal incidence numbers. Further, the estimates we produce are not able to offer indication of the breakdown in the typology of slavery occurring within a country. Recent work has focused on establishing proxy indicators for specific types of vulnerability/exploitation, or exploitation within specific industries, using small scale surveys married to digital traces such as mobile phone records [49], satellite imagery [50, 51, 52, 53], and vessel data [54]. Such approaches can help to bolster the data in this domain without putting vulnerable individuals at further risk.

This study applied machine learning methods to the context of modern slavery to better understand the drivers of this ongoing phenomenon. Using rigorous cross validation approaches, careful consideration of model stability, and an in-depth variable importance analysis, stable predictive components emerged using data characterised as "small n, large p". Notably, a novel predictive component was found, in the *Physical Security of Women*, which was shown to predict prevalence non-linearly, and in association with other components (in particular Access to Resources). Such non-linear relationships, combined with the challenges traditional methods face when dealing with multicollinearity, may well be the reason that this component, and its in its ability to support quantitative prediction of slavery, has been previously overlooked. The findings make the case for further exploration of data-driven, inductive approaches and out-of-sample prediction to complement existing methodologies being used to study complex social problems such as modern-day slavery.

Method

Data

Selecting the Dependent variable. A single dependent variable was used. This was the GSI's country level prevalence estimates derived from Gallup World Poll survey data (provided by the ILO and WFF) and converted to a percentage of the population. This included prevalence estimates for 48 unique countries over 2016 and 2018[3]. For 22 countries

[3] The 2018 GSI in actually includes estimates of prevalence for a total of 167 countries. However, the majority of these were produced using a risk model, and not estimated from survey data. The reliability and practical usefulness of these extrapolated estimates

there were estimates for both 2016 and 2018, giving a total of 70 data points over the two years[4]. It is noted that the countries in this sample did not include any countries from Western Europe or North America, and thus the generalisability of the findings to these regions are limited. The median value of prevalence was 0.46%, with the upper and lower quartiles as 0.26% and 0.77% respectively.

Selecting the independent variables. The independent variables were scraped from online sources such as the World Bank Development Indicators (2018), UNAIDS (2019), the WomanStats project [47], the Early Warning Project, CIRI Human Rights Data Project [55], [33] and the UN's Sustainable Development Goal (SDG) indicators. A total of 106 features were selected and collated[5]. The features, feature descriptions, and sources can be found in the Supplementary Materials. The data sources were verified to ensure that sufficient information on how the data was coded/collected was available, missing data was minimal for the 70 data points, and information existed for the time period that the data reflected. If the variable was a composite scale or score then the variables from which it was constructed from, and how, needed to be clear before being selected.

Data Pre-processing. Information on the open source variables indicated that often data had been collected over multiple years. When this was the case, the most recent year in the collection time frame was recorded. The year the data was collected was an important part of the data processing as for 22 countries there were two dependent variable data points - slavery prevalence in 2016 and slavery prevalence in 2018. Therefore, data was sought to cover both these time periods. Overall, there was not enough data specific to the years 2016 and 2018 so the data was grouped by time period (2016 and before, and post 2016) and the most recent data from each group selected[6]. As an example, the Physical Security of Women Scale had data only from the years 2014 and 2019, therefore the former went in group 1 to predict slavery in 2016 and the latter went into group 2 to predict slavery in 2018. Grouping the data in this manner reduced the missing data, but with the caveat that the features were assumed to be relatively stable between 2011 and 2016 for group 1, and between 2016 to 2019 for group 2.

After this, any missing data that remained was then dealt with in the following way. First, variables which had more than 50% of data missing were discarded. Then, remaining data were imputed in two steps. Where a country had either 2016 or 2018 data, the data from the year where it was available was used for both years. (Therefore in a small number of cases the same feature value was used to predict two different dependent variable values - prevalence in 2016 and 2018). After this, the subsequent missing values were minimal, and can be viewed in the Appendix, Figure A6. The final step was to use a multivariate feature imputation method with regression trees for the remaining missing values[7]. Variables were then normalised to be between between 0 and 1.

Analysis

The methodology used in this study was exploratory, inductive and data driven. This involved evaluating a number of different approaches and model classes to assess what configuration of method and model produced the most accurate predictions. Firstly, the utility of using *all* the features versus a smaller pool of features selected from the literature (the approach used by the GSI [34]) was tested. This theory driven selection process was undertaken using a literature review and by consulting with domain experts. The 106 features were reduced to 35 spanning poverty, globalisation and the country's wealth, education, politics, violence, and conflict (the full list can be found in the Appendix, Figure A6). Secondly, the value in using feature decomposition prior to modelling verses inputting the raw variables into the model was assessed. Non-negative Matrix Factorisation (NMF) was chosen as it forces the matrices to be non-negative, making the emergent components easier to interpret than when using other methods such as Principle Component Analysis (PCA). Finally, three different model classes (also selected for their interpretability) linear regression, decision tree, and random forest, with optimised meta-parameters, were also evaluated. The resultant sixteen combinations of different methods and models types can be found in Table 1. Leave one out cross validation (LOOCV) was used to evaluate the model parameters selected to ensure that the model was generalizable to the full data set. Given the lack of data and that fitting the model for an explanation, rather than pure prediction, was the goal of the study, no data was held back to create a separate test set.

have been queried [7, 33]. Therefore, this analysis only uses estimates which were derived directly from the GWP survey data (as in [33]).

[4] Given that the sampling procedure for both the 2016 and 2018 Gallup World Poll was random [48], there was no known reason to assume that the 2016 survey would have directly influenced the 2018 survey; and thus the country prevalence estimates for 2016 and 2018 were treated as independent.

[5] Features were selected based on past literature, the authors' intuition, and conversations with domain experts. Initially there were 137 variables in the feature pool, before 31 were discarded due to missing data, duplication, or measuring a phenomena too similar to the dependent variable.

[6] For ease, the 2018 SDG data was used for 2016 and 2018 due to the multiple sources and dates from which the data were collected

[7] The CART regression tree method was chosen due to features being too highly correlated to perform regression based imputation methods, such as predictive mean matching (causing singularities in the matrix). Further, as linear models such as NMF and linear regressions were being utilised in this analysis, a non-linear imputation method was preferable to avoid the variables' collinearity being inflated by the imputation process.

To harness the model for out-of-sample prediction (the prediction of slavery in countries where no survey data exists) additional independent variable data were scraped for the 171 countries (in 2018) unseen by the model. Where independent variable data was not directly available, the same imputation method was used as described previously. For more information on the missing data prior to imputation for this sample see Appendix, Table A1. The model was then trained on the 70 data points for 2016 and 2018 where survey data was used in the estimates, and out-of-sample predictions made for the countries with no survey data.

References

1. ILO. Global estimates of modern slavery: Forced labour and forced marriage, 2017.
2. WFF. Gsi 2018 methodology, prevalence. section: Data limitations, 2018. https://www.globalslaveryindex.org/2018/methodology/prevalence/.
3. Kevin Bales, Olivia Hesketh, and Bernard Silverman. Modern slavery in the uk: How many victims? *Significance*, 12(3):16–21, 2015.
4. Jacqueline Joudo Larsen and Davina P Durgana. Measuring vulnerability anelta estimating prevalence of modern slavery. *Chance*, 30(3):21–29, 2017.
5. Todd Landman. Measuring modern slavery: Law, human rights, and new forms of data. *Human Rights Quarterly*, 42(2):303–331, 2020.
6. Lax Chan, Bernard W Silverman, and Kyle Vincent. Multiple systems estimation for sparse capture data: Inferential challenges when there are nonoverlapping lists. *Journal of the American Statistical Association*, pages 1–10, 2020.
7. Bernard W Silverman. Demonstrating risks is not the same as estimating prevalence. *Paper presented at Delta 8.7 Modelling the Risk of Modern Slavery Symposium*, 2018. https://delta87.org/2018/12/demonstrating-risk-not-same-estimating-prevalence/.
8. Pablo Diego-Rosell and Jacqueline Joudo Larsen. Modelling the risk of modern slavery. 2018.
9. Kelly Gleason. Facing choices when modelling modern slavery risk. *Paper presented at Delta 8.7 Modelling the Risk of Modern Slavery Symposium*, 2018. https://delta87.org/2018/12/demonstrating-risk-not-same-estimating-prevalence/.
10. James Cockayne, Kelly A Gleason, Pablo Diego-Rossell, Shannon Stewart, Laura Gauer Bermudez, Jacqueline Joudo Larsen, and Bernard W Silverman. Modelling modern slavery risk. 2019.
11. Monti Narayan Datta, Olivia Gustafson, Chloe Lubin, Gioia Kelleher, and Rebecca Berg. Assessing the global slavery index. *The SAGE Handbook of Human Trafficking and Modern Day Slavery*, page 38, 2018.
12. Anne T Gallagher. What's wrong with the global slavery index? *Anti-Trafficking Review*, (8), 2017.
13. Andrew Guth, Robyn Anderson, Kasey Kinnard, and Hang Tran. Proper methodology and methods of collecting and analyzing slavery data: an examination of the global slavery index. 2014.
14. G Kessler. Be wary of precise figures in an underground industry such as sex trafficking., 2015.
15. Ronald Weitzer. Miscounting human trafficking and slavery. *Open Democracy*, 2014.
16. Maarten Cruyff, Jan van Dijk, and Peter GM van der Heijden. The challenge of counting victims of human trafficking: Not on the record: A multiple systems estimation of the numbers of human trafficking victims in the netherlands in 2010–2015 by year, age, gender, and type of exploitation. *Chance*, 30(3):41–49, 2017.
17. Kevin Bales, Laura T Murphy, and Bernard W Silverman. How many trafficked people are there in greater new orleans? lessons in measurement. *Journal of Human Trafficking*, pages 1–13, 2019.
18. Bernard W Silverman. Model fitting in multiple systems analysis for the quantification of modern slavery: Classical and bayesian approaches. *arXiv preprint arXiv:1902.06078*, 2019.
19. Iain M Johnstone and D Michael Titterington. Statistical challenges of high-dimensional data, 2009.
20. Brian W Head and John Alford. Wicked problems: Implications for public policy and management. *Administration & society*, 47(6):711–739, 2015.
21. Kathleen Fitzgibbon. Modern-day slavery? the scope of trafficking in persons in africa. *African Security Studies*, 12(1):81–89, 2003.
22. Bill Rau. *Combating child labour and HIV/AIDS in sub-Saharan Africa.* International Labour Office, 2002.
23. ILO. Children in prostitution–a rapid assessment. *ILO Tanzania*, 2001.

24. Kate Manzo. Exploiting west africa's children: trafficking, slavery and uneven development. *Area*, 37(4):393–401, 2005.

25. Kevin Bales. Testing a theory of modern slavery. *Free the Slaves*, 2006.

26. Ivan Manokha. Modern slavery and fair trade products: Buy one and set someone free. In *The Political Economy of New Slavery*, pages 217–234. Springer, 2004.

27. Olubukola S Adesina. Modern day slavery: poverty and child trafficking in nigeria. *African Identities*, 12(2): 165–179, 2014.

28. Monti Narayan Datta and Kevin Bales. Slavery is bad for business: analyzing the impact of slavery on national economies. *The Brown journal of world affairs*, 19(2): 205–223, 2013.

29. Hans Van de Glind and Joost Kooijmans. Modern-day child slavery 1. *Children & Society*, 22(3):150–166, 2008.

30. Max Roser and Hannah Ritchie. Hiv / aids. *Our World in Data*, 2018. https://ourworldindata.org/hiv-aids.

31. David Brown, Doreen S Boyd, Katherine Brickell, Christopher D Ives, Nithya Natarajan, and Laurie Parsons. Modern slavery, environmental degradation and climate change: Fisheries, field, forests and factories. *Environment and Planning E: Nature and Space*, page 2514848619887156, 2019.

32. David Tickler, Jessica J Meeuwig, Katharine Bryant, Fiona David, John AH Forrest, Elise Gordon, Jacqueline Joudo Larsen, Beverly Oh, Daniel Pauly, Ussif R Sumaila, et al. Modern slavery and the race to fish. *Nature communications*, 9(1):4643, 2018.

33. Todd Landman and Bernard W Silverman. Globalization and modern slavery. *Politics and Governance*, 7(4), 2019.

34. WFF. Gsi 2018 methodology, vulnerability model., 2018. `https://www.globalslaveryindex.org/2018/methodology/vulnerability/`.

35. Jiayun Dong and Cynthia Rudin. Exploring the cloud of variable importance for the set of all good models. *Nature Machine Intelligence*, 2(12):810–824, 2020.

36. Laura L Nathans, Frederick L Oswald, and Kim Nimon. Interpreting multiple linear regression: A guidebook of variable importance. *Practical Assessment, Research, and Evaluation*, 17(1):9, 2012.

37. Tal Yarkoni and Jacob Westfall. Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, 12(6): 1100–1122, 2017.

38. Been Kim, Kayur Patel, Afshin Rostamizadeh, and Julie Shah. Scalable and interpretable data representation for high-dimensional complex data. 2015.

39. Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1 (5):206–215, 2019.

40. Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019.

41. Gavin Smith, Roberto Mansilla, and James Goulding. Model class reliance for random forests. *Advances in Neural Information Processing Systems*, 33, 2020.

42. André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10): 1340–1347, 2010.

43. Leo Breiman. Random forests. *Machine learning*, 45 (1):5–32, 2001.

44. Lesia Semenova and Cynthia Rudin. A study in rashomon curves and volumes: A new perspective on generalization and model simplicity in machine learning. *arXiv preprint arXiv:1908.01755*, 2019.

45. Leo Breiman et al. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001.

46. Galit Shmueli et al. To explain or to predict? *Statistical science*, 25(3):289–310, 2010.

47. Mary Caprioli, Valerie M Hudson, Rose McDermott, Bonnie Ballif-Spanvill, Chad F Emmett, and S Matthew Stearmer. The womanstats project database: Advancing an empirical research agenda. *Journal of Peace Research*, 46(6):839–851, 2009.

48. Gallup World Poll. How does the gallup world poll work? measures the attitudes and behaviors of the world's residents, 2020. `https://www.gallup.com/178667/gallup-world-poll-work.aspx`.

49. Gregor Engelmann, Gavin Smith, and James Goulding. The unbanked and poverty: Predicting area-level socio-economic vulnerability from m-money transactions. In

2018 IEEE International Conference on Big Data (Big Data), pages 1357–1366. IEEE, 2018.

50. Doreen S Boyd, Bethany Jackson, Jessica Wardlaw, Giles M Foody, Stuart Marsh, and Kevin Bales. Slavery from space: Demonstrating the role for satellite remote sensing to inform evidence-based action related to un sdg number 8. *ISPRS journal of photogrammetry and remote sensing*, 142:380–388, 2018.

51. Bethany Jackson, Kevin Bales, Sarah Owen, Jessica Wardlaw, and Doreen S Boyd. Analysing slavery through satellite technology: How remote sensing could revolutionise data collection to help end modern slavery. *Journal of Modern Slavery*, 4(2), 2018.

52. Bethany Jackson, Doreen S Boyd, Christopher D Ives, Jessica L Decker Sparks, Giles M Foody, Stuart Marsh, and Kevin Bales. Remote sensing of fish-processing in the sundarbans reserve forest, bangladesh: an insight into the modern slavery-environment nexus in the coastal fringe. *Maritime Studies*, pages 1–16, 2020.

53. Giles M Foody, Feng Ling, Doreen S Boyd, Xiaodong Li, and Jessica Wardlaw. Earth observation and machine learning to meet sustainable development goal 8.7: Mapping sites associated with slavery from space. *Remote Sensing*, 11(3):266, 2019.

54. Katrina Nakamura, Lori Bishop, Trevor Ward, Ganapathiraju Pramod, Dominic Chakra Thomson, Patima Tungpuchayakul, and Sompong Srakaew. Seeing slavery in seafood supply chains. *Science advances*, 4(7): e1701833, 2018.

55. David L. Richards Cingranelli, David L. and K. Chad Clay. The ciri human rights dataset, 2014. `http://www.humanrightsdata.com`.

Appendix A

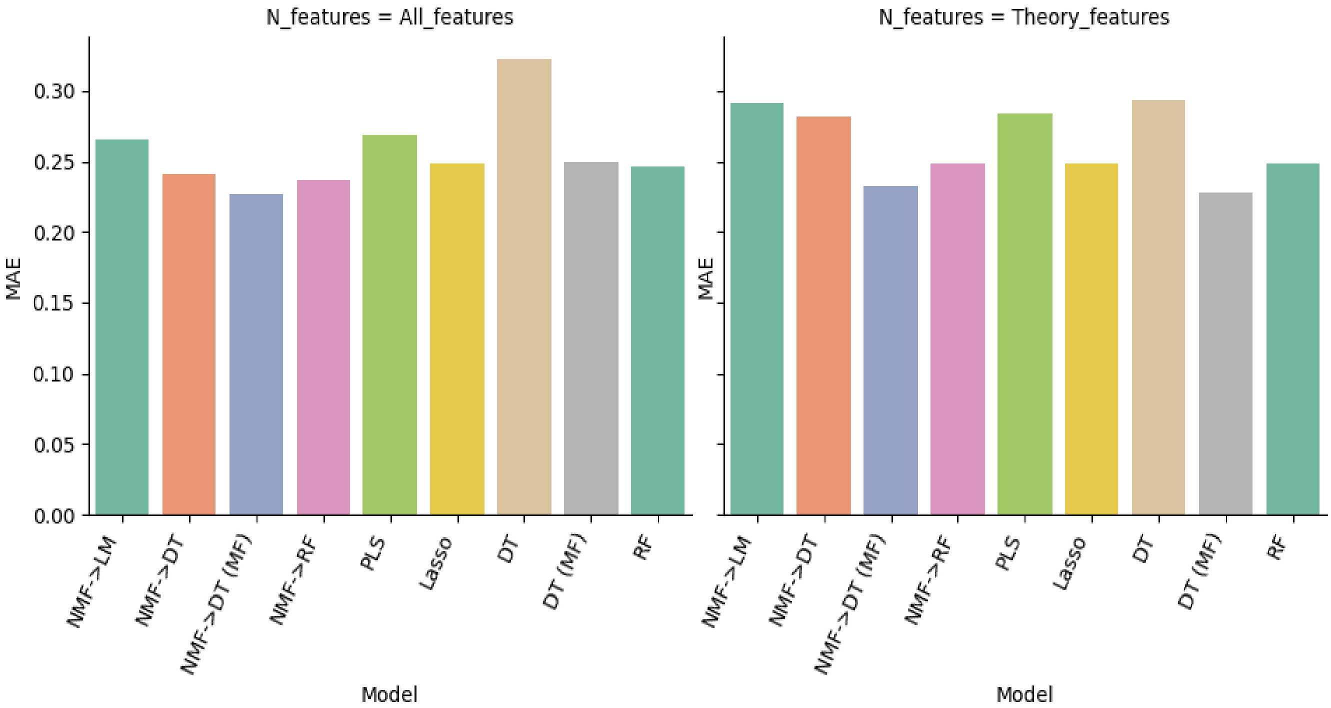
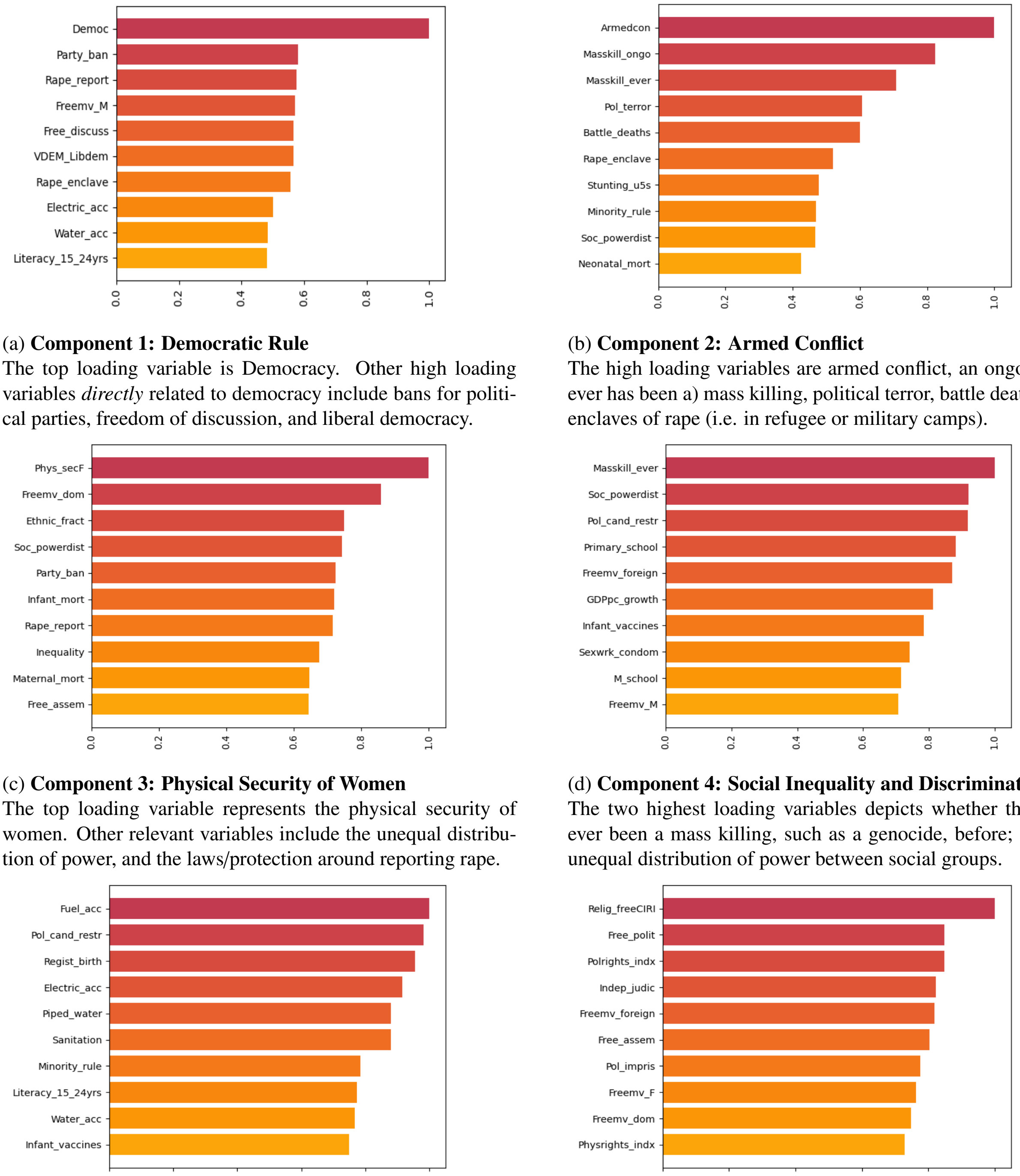


Figure A1. Graphs to visually compare the pipeline performances. NMF = Non Negative Matrix Factorisation; DT = Decision Tree; MF = Max features as a parameter available to the decision tree; RF = Random Forest; LM = Linear (regression) Model; PLS = Partial Least Squares regression.



(a) **Component 1: Democratic Rule**
The top loading variable is Democracy. Other high loading variables *directly* related to democracy include bans for political parties, freedom of discussion, and liberal democracy.

(b) **Component 2: Armed Conflict**
The high loading variables are armed conflict, an ongoing (or ever has been a) mass killing, political terror, battle deaths, and enclaves of rape (i.e. in refugee or military camps).

(c) **Component 3: Physical Security of Women**
The top loading variable represents the physical security of women. Other relevant variables include the unequal distribution of power, and the laws/protection around reporting rape.

(d) **Component 4: Social Inequality and Discrimination**
The two highest loading variables depicts whether there has ever been a mass killing, such as a genocide, before; and the unequal distribution of power between social groups.

(e) **Component 5: Access to Resources**
High loading relevant variables include access to fuel, electricity, water, and sanitation. Registered births, poor literacy, and a lack of infant vaccines also denote poor access to infrastructure/services.

(f) **Component 6: Religious and Political Freedoms**
High loading variables include religious and political freedoms, political and physical rights, independent judicial system, foreign and domestic freedom of movement, freedom to assemble, political imprisonment, and freedom of movement for women.

Figure A2. The six components from the best model. On the y axis are the original raw variables, and the bars show their loading value onto the new component (normalised to be between 0 and 1). The full variable names, a short description, and their source can be found in the Supplementary Materials. Note that here only the top 10 loading variables are shown for simplicity. The themes the components represent were decided after conversations with domain experts.

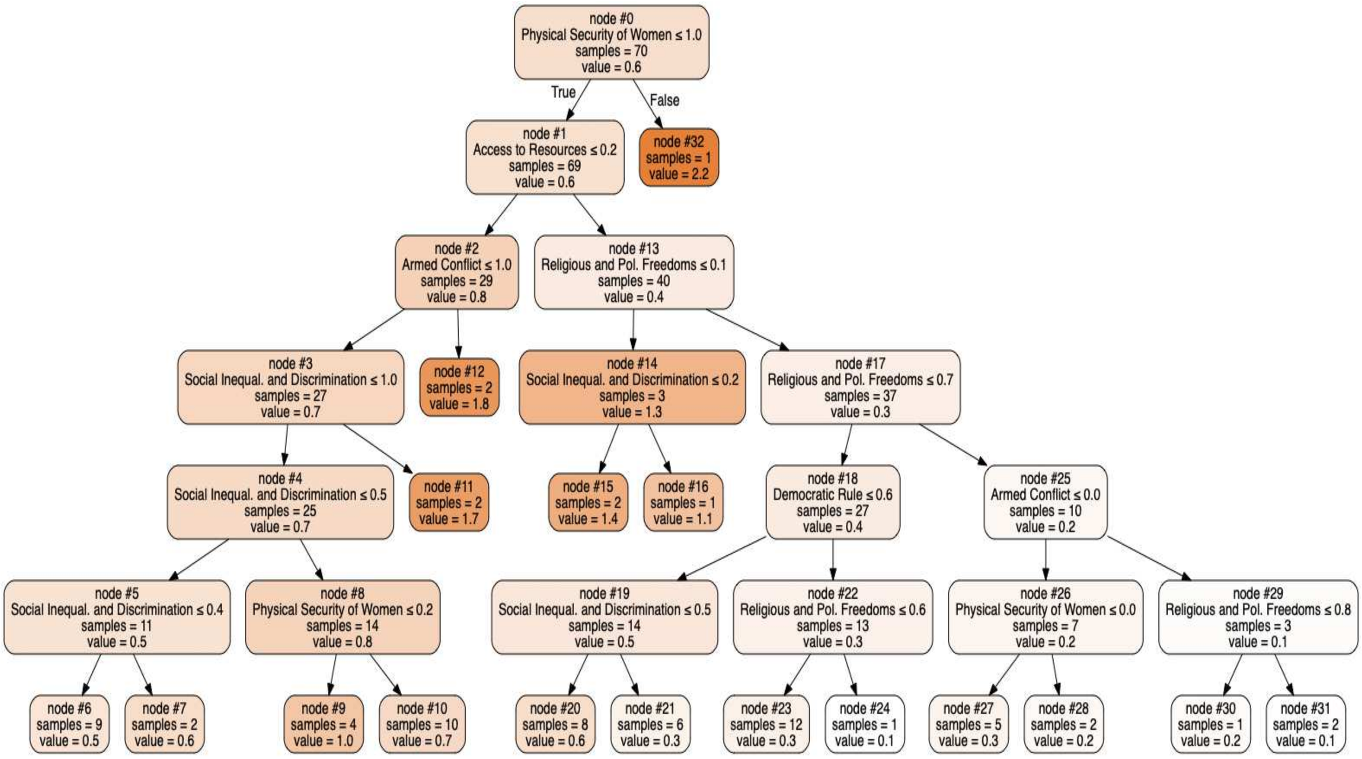
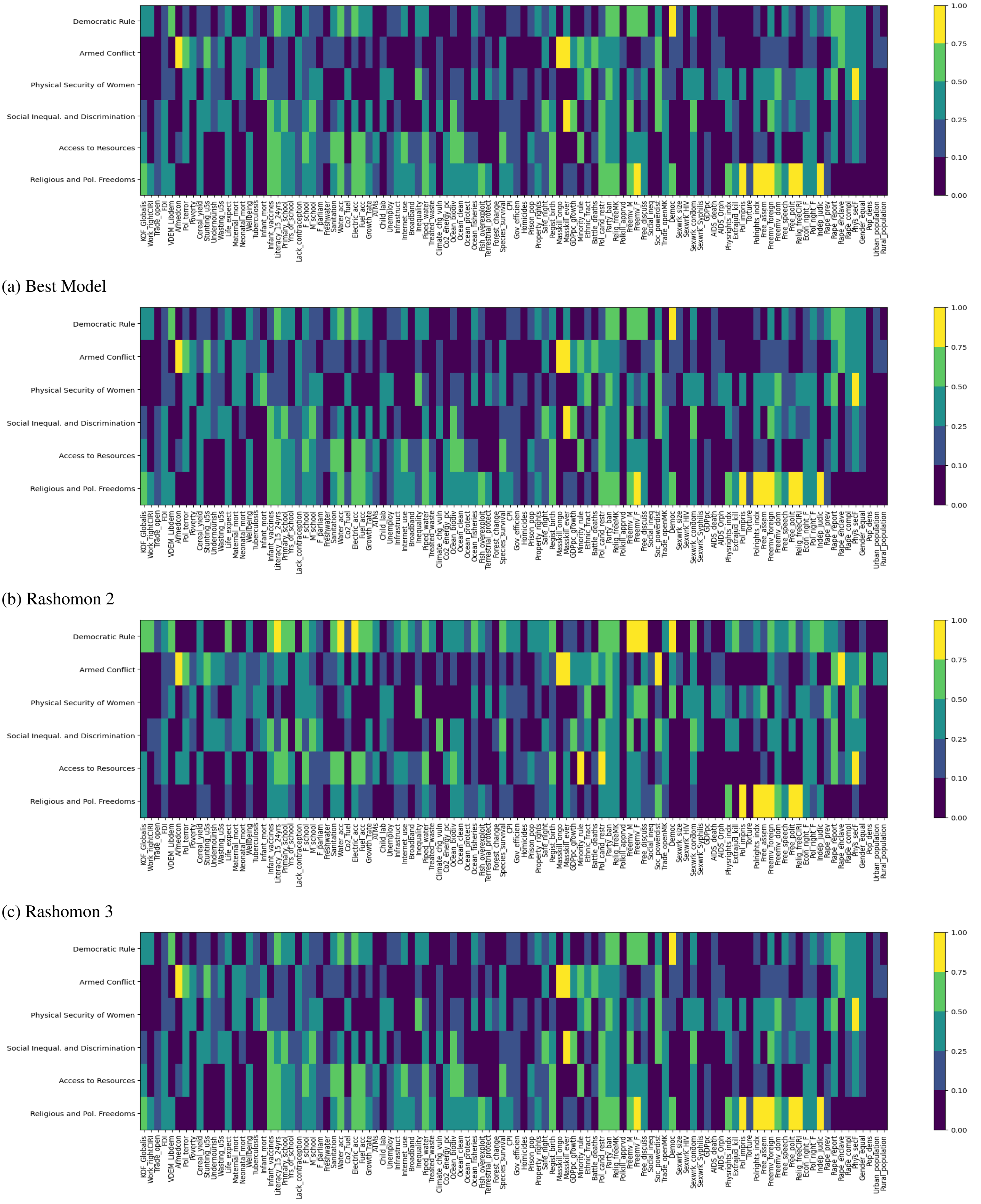


Figure A3. The best decision tree model predicting prevalence of modern slavery using the six components from the NMF. 'Samples' represents the number of country-year data points at each node (out of 70). The 'value' refers to the predicted prevalence (slavery as a percentage of the population) for that node.



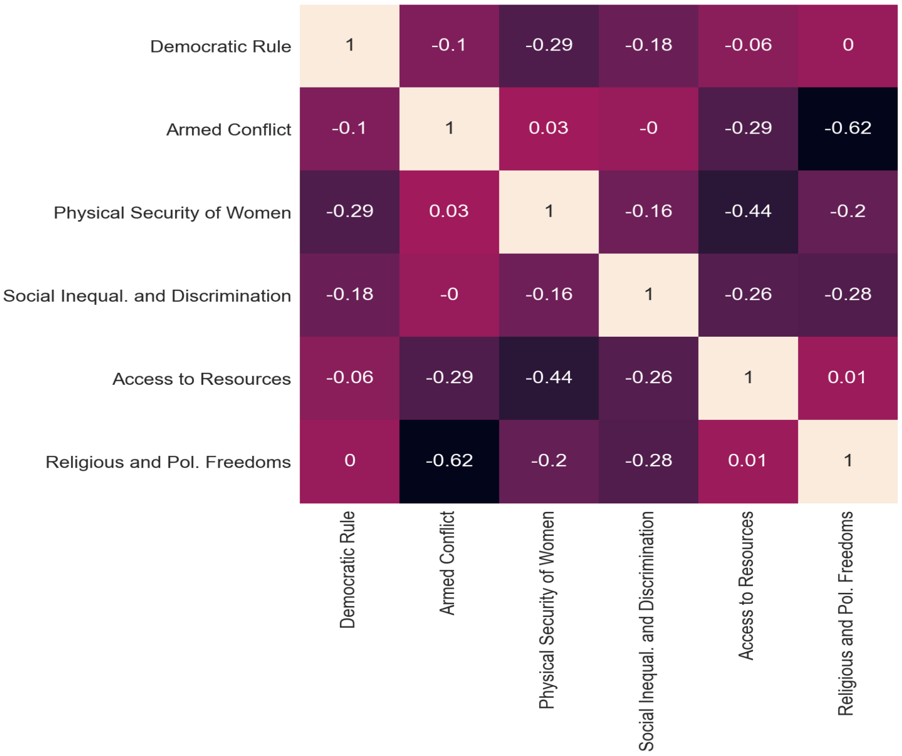
(a) Best Model

(b) Rashomon 2

(c) Rashomon 3

(d) Rashomon 4

Figure A4. The normalised variable loadings onto the components (H matrix from the NMF) for the Rashomon models with the same pipeline as the best model (all the features and the model class NMF->DT (MF)). Some differences in loadings occur due variances in the NMF paramterisation (different random seeds being used for coordinate descent, and different constants that multiply the regularisation terms (alpha)) yet the overarching themes remain stable.



	Democratic Rule	Armed Conflict	Physical Security of Women	Social Inequal. and Discrimination	Access to Resources	Religious and Pol. Freedoms
Democratic Rule	1	-0.1	-0.29	-0.18	-0.06	0
Armed Conflict	-0.1	1	0.03	-0	-0.29	-0.62
Physical Security of Women	-0.29	0.03	1	-0.16	-0.44	-0.2
Social Inequal. and Discrimination	-0.18	-0	-0.16	1	-0.26	-0.28
Access to Resources	-0.06	-0.29	-0.44	-0.26	1	0.01
Religious and Pol. Freedoms	0	-0.62	-0.2	-0.28	0.01	1

Figure A5. A correlation matrix showing the Pearson r correlations between the components. Here a high value for Physical Security of Women, depicts less security.

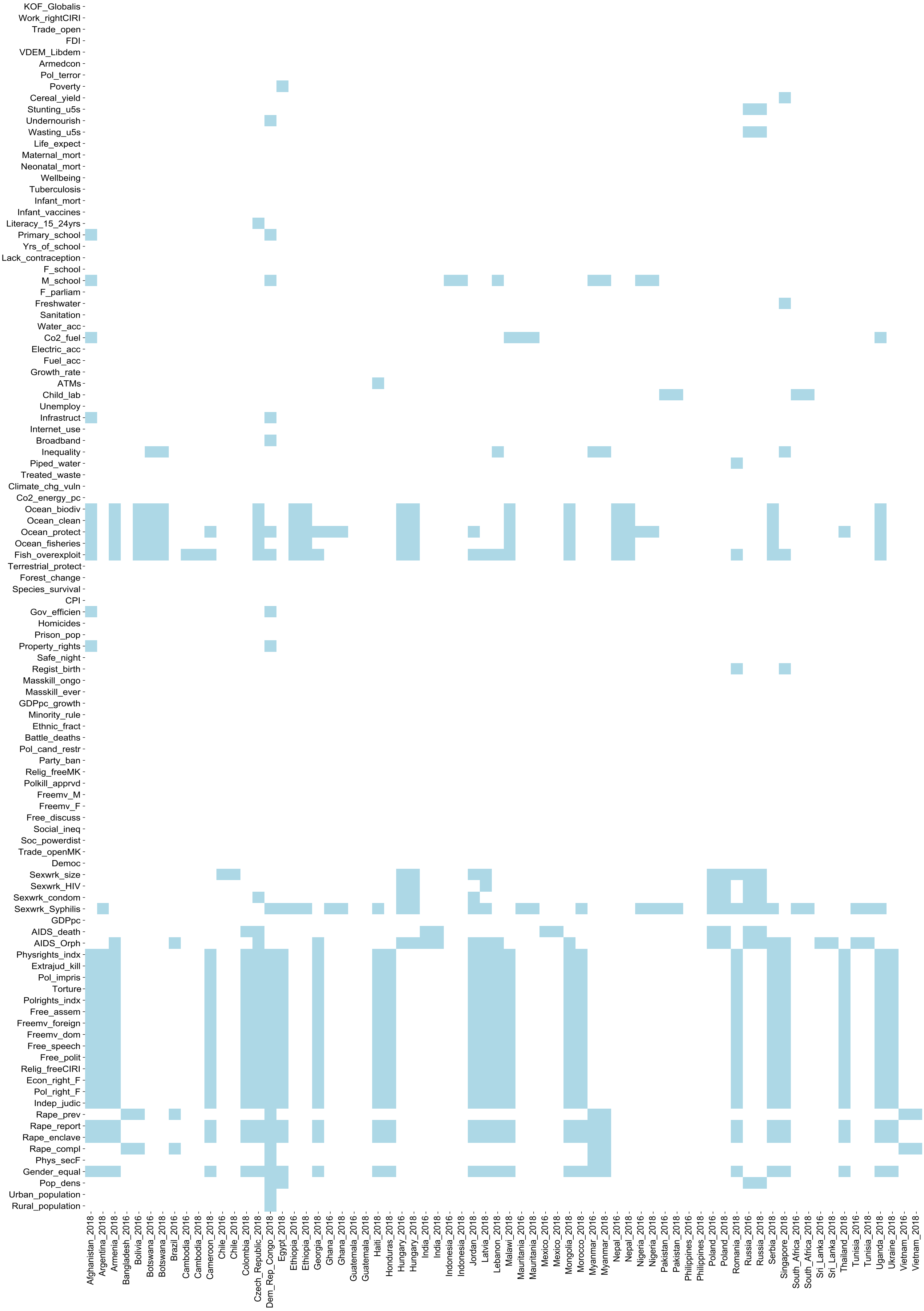


Figure A6. Missing Data for all 106 variables before data was imputed using CART.

Table A1

A comparison of 2018 prevalence estimates made by our model compared those made by the WFF's model as part of the 2018 GSI. Missing data represents the percentage of missing independent variable data for our model before imputations were made.

Country	% Missing Data	Our Best Model Estimate	GSI Estimate
Albania	5	0.1	0.69
Algeria	2	0.61	0.27
American Samoa	94	0.5	0.26
Andorra	83	0.18	0.08
Angola	4	0.69	0.72
Antigua and Barbuda	82	0.18	0.33
Aruba	92	0.5	0.3
Australia	4	0.18	0.06
Austria	11	0.26	0.17
Azerbaijan	8	0.28	0.45
Bahamas	78	0.18	0.38
Bahrain	39	0.61	0.19
Bangladesh	0	0.69	0.37
Barbados	53	0.18	0.27
Belarus	9	0.18	1.09
Belgium	7	0.18	0.2
Belize	75	0.18	0.34
Benin	3	0.5	0.55
Bermuda	94	0.69	0.4
Bhutan	74	0.69	1.13
Bolivia	10	0.5	0.21
Bosnia and Herzegovina	8	0.28	0.34
Brazil	1	0.18	0.18
British Virgin Islands	96	0.18	0.29
Brunei Darussalam	62	0.61	1.09
Bulgaria	3	0.1	0.45
Burkina Faso	10	0.5	0.45
Burundi	8	0.69	4.0
Cabo Verde	25	0.5	0.41
Canada	7	0.18	0.05
Cayman Islands	93	0.18	0.21
Central African Republic	17	0.5	2.23
Chad	8	0.5	1.2
China	7	0.28	0.28
Comoros	76	0.28	0.55
Congo Republic	7	0.69	0.8
Costa Rica	2	0.18	0.13
Cote d'Ivoire	8	0.69	0.59
Croatia	6	0.18	0.6
Cuba	39	0.61	0.38
Curacao	94	0.61	0.4
Cyprus	8	0.18	0.42
Denmark	6	0.18	0.16
Djibouti	36	0.28	0.71
Dominica	85	0.5	0.34
Dominican Republic	0	0.28	0.4
Ecuador	2	0.18	0.24

Table A1 continued from previous page

Country	% Missing Data	Our Best Model Estimate	GSI Estimate
El Salvador	0	0.61	0.25
Equatorial Guinea	40	0.28	0.64
Eritrea	45	0.28	9.3
Estonia	2	0.18	0.36
Eswatini	8	0.61	0.88
Faeroe Islands	93	0.18	0.77
Fiji	80	0.18	0.52
Finland	4	0.18	0.17
France	5	0.18	0.2
French Polynesia	96	0.18	0.4
Gabon	2	0.61	0.48
Gambia	5	0.64	0.58
Germany	3	0.18	0.2
Gibraltar	96	0.18	0.4
Greece	5	0.18	0.79
Greenland	94	0.18	0.2
Grenada	83	0.18	0.41
Guam	94	0.18	0.29
Guinea-Bissau	36	0.5	0.75
Guinea	2	0.69	0.78
Guyana	7	0.5	0.26
Hong Kong	92	0.61	0.14
Iceland	25	0.18	0.21
Iran	7	0.61	1.62
Iraq	11	0.61	0.48
Ireland	3	0.18	0.17
Isle of Man	95	0.15	0.06
Israel	6	0.08	0.39
Italy	4	0.18	0.24
Jamaica	3	0.18	0.26
Japan	5	0.26	0.03
Kazakhstan	7	0.61	0.42
Kenya	2	0.5	0.69
Kiribati	83	0.69	0.52
Kuwait	10	0.61	0.15
Kyrgyz Republic	6	0.61	0.41
Laos	12	0.69	0.94
Lesotho	6	0.5	0.42
Liberia	2	0.5	0.74
Libya	40	0.61	0.77
Liechtenstein	85	0.18	0.17
Lithuania	6	0.18	0.58
Luxembourg	27	0.18	0.15
Macao	92	0.18	0.29
Macedonia	11	0.18	0.87
Madagascar	1	0.69	0.75
Malaysia	2	0.61	0.69
Maldives	77	0.28	0.64
Mali	7	0.5	0.36
Malta	74	0.18	0.37
Marshall Islands	85	0.5	0.33

Table A1 continued from previous page

Country	% Missing Data	Our Best Model Estimate	GSI Estimate
Mauritius	6	0.5	0.1
Micronesia, Fed. Sts.	86	0.5	0.87
Moldova	11	0.18	0.55
Monaco	84	0.18	0.06
Montenegro	6	0.18	0.59
Mozambique	3	0.69	0.54
Namibia	4	0.5	0.33
Nauru	83	0.15	0.17
Netherlands	3	0.18	0.18
New Caledonia	96	0.69	0.67
New Zealand	5	0.18	0.06
Nicaragua	0	0.61	0.29
Niger	9	0.5	0.67
North Korea	52	0.28	10.46
Northern Mariana Islands	94	0.61	0.29
Norway	6	0.26	0.18
Oman	10	0.28	0.21
Palau	88	0.18	0.02
Palestine	91	0.61	0.53
Panama	0	0.18	0.21
Papua New Guinea	39	0.69	1.03
Paraguay	5	0.15	0.16
Peru	0	0.69	0.26
Portugal	2	0.18	0.25
Puerto Rico	92	0.61	0.4
Qatar	8	0.61	0.15
Rwanda	6	0.69	1.16
Saint-Martin	98	0.69	0.96
Samoa	82	0.28	1.09
San Marino	83	0.18	0.15
Sao Tome and Principe	77	0.5	0.4
Saudi Arabia	10	0.61	0.19
Senegal	0	0.5	0.29
Seychelles	78	0.18	0.18
Sierra Leone	4	0.5	0.5
Sint Maarten	95	0.61	0.21
Slovakia	10	0.18	0.29
Slovenia	6	0.18	0.22
Solomon Islands	77	0.69	1.0
Somalia	37	1.06	1.55
South Korea	12	0.26	0.19
South Sudan	41	0.5	2.05
Spain	2	0.18	0.23
St. Kitts and Nevis	83	0.18	0.42
St. Lucia	77	0.18	0.4
St. Vincent and the Grenadines	82	0.18	0.17
Sudan	34	0.61	1.2
Suriname	21	0.15	0.23
Sweden	8	0.18	0.16
Switzerland	10	0.18	0.17
Syria	44	0.61	0.73

Table A1 continued from previous page

Country	% Missing Data	Our Best Model Estimate	GSI Estimate
Tajikistan	6	0.28	0.45
Tanzania	6	0.69	0.62
Timor-Leste	42	0.5	0.77
Togo	8	0.5	0.68
Tonga	80	0.5	0.75
Trinidad and Tobago	13	0.18	0.3
Turkey	6	0.28	0.65
Turkmenistan	40	0.28	1.12
Turks and Caicos Islands	94	0.18	0.24
Tuvalu	86	0.5	0.64
United Arab Emirates	11	0.61	0.17
United Kingdom	7	0.18	0.21
United States	7	0.18	0.13
United States Virgin Islands	94	0.64	0.24
Uruguay	0	0.18	0.1
Uzbekistan	35	0.28	0.52
Vanuatu	78	0.5	0.41
Venezuela	12	0.61	0.56
Western Sahara	98	0.69	0.48
Yemen	5	0.61	0.31
Zambia	5	0.5	0.57
Zimbabwe	7	0.61	0.67

Appendix B
Best Model Parameters

The best performing model (MAE=0.227) used the following meta-parameters for the NMF and Decision Tree functions in the Scikit-learn python package: random state = 45, NMF K components = 6, NMF solver = 'cd', NMF tolerance = 0.005, NMF alpha = 2, NMF max iterations = 350, max depth of the tree = 6; tree max features = 0.3, tree minimum samples to split on = 3.

Figures

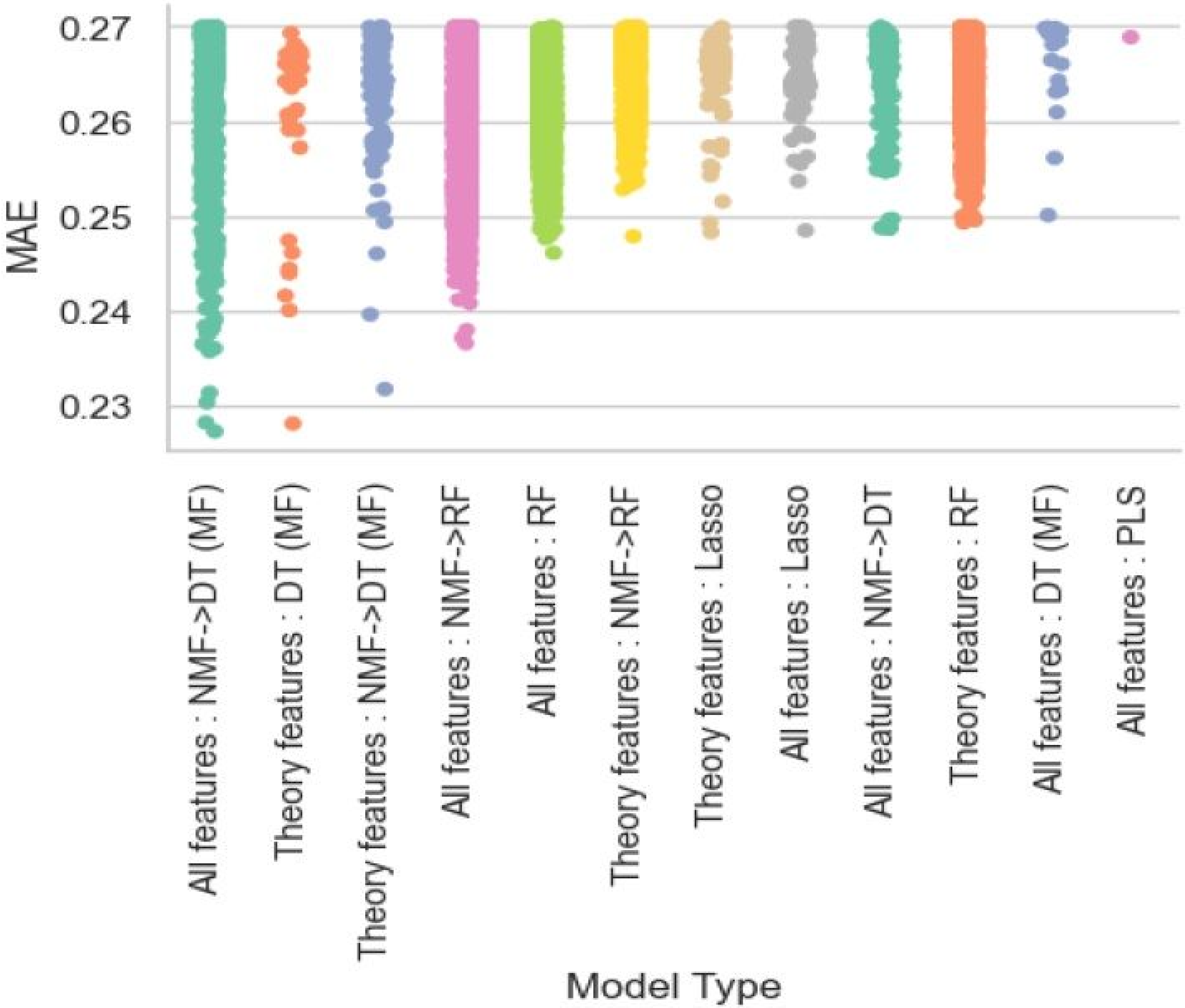


Figure 1

The LOOCV performance of all models where MAE was less than 0.27.

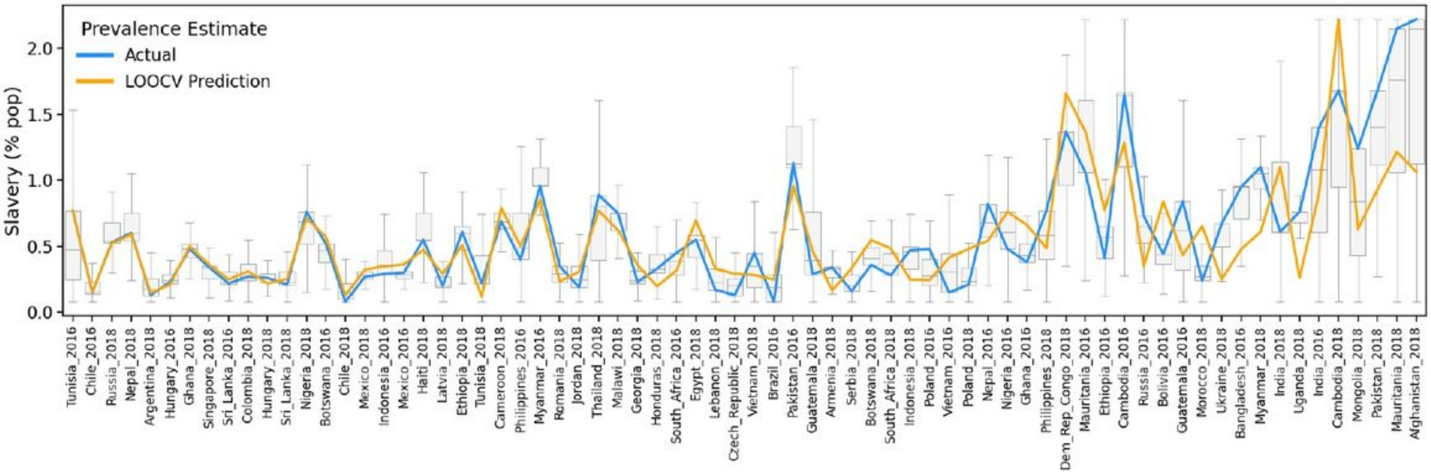


Figure 2

The predictions of slavery prevalence (individuals enslaved as a % of the population) made by the best model using leave-one-out cross validation (LOOCV), compared to the ‘actual’ prevalence as estimated using the GallupWorld Poll (GWP) survey data. The grey box plots illustrate the distribution of 10,000 bootstrapped LOOCV predictions (using the full pipeline NMF->DT) to help illustrate the uncertainty associated with our model’s predictions. The box shows the quartiles of the bootstrapped predictions while the whiskers extend to show the rest of the distribution, except for points that were determined to be “outliers” (using a function of the inter-quartile range) which are not plotted. The x axis is ordered by the MAE.

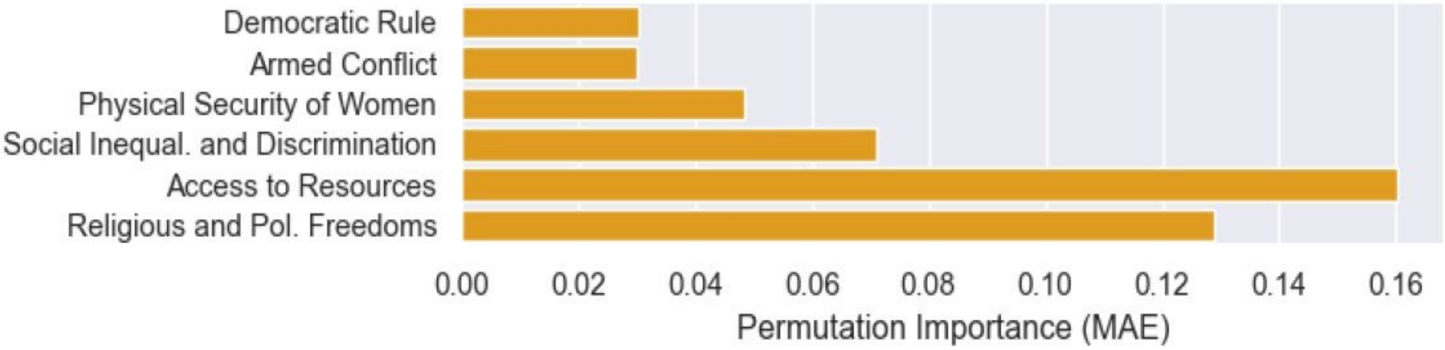


Figure 3

The permutation importance [42] of the NMF components (identified latent variables) in the best performing model.

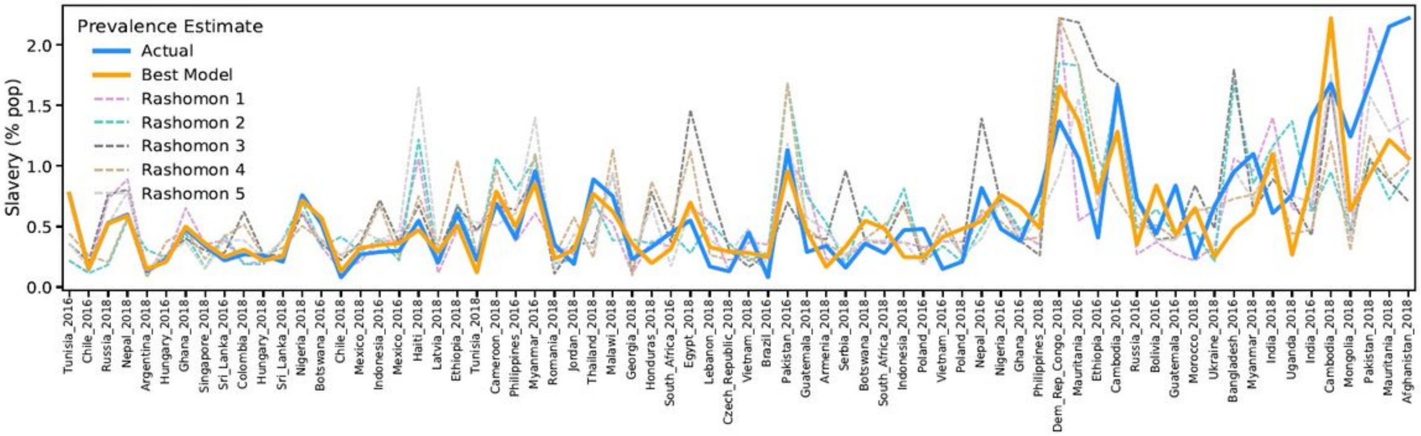


Figure 4

The predictions of slavery prevalence (individuals enslaved as a % of the population) made by the best model and the five other well performing models in the Rashomon set (see Table 2). The x axis is ordered by the MAE between our best model and the 'actual' prevalence as estimated by the GWP survey data.

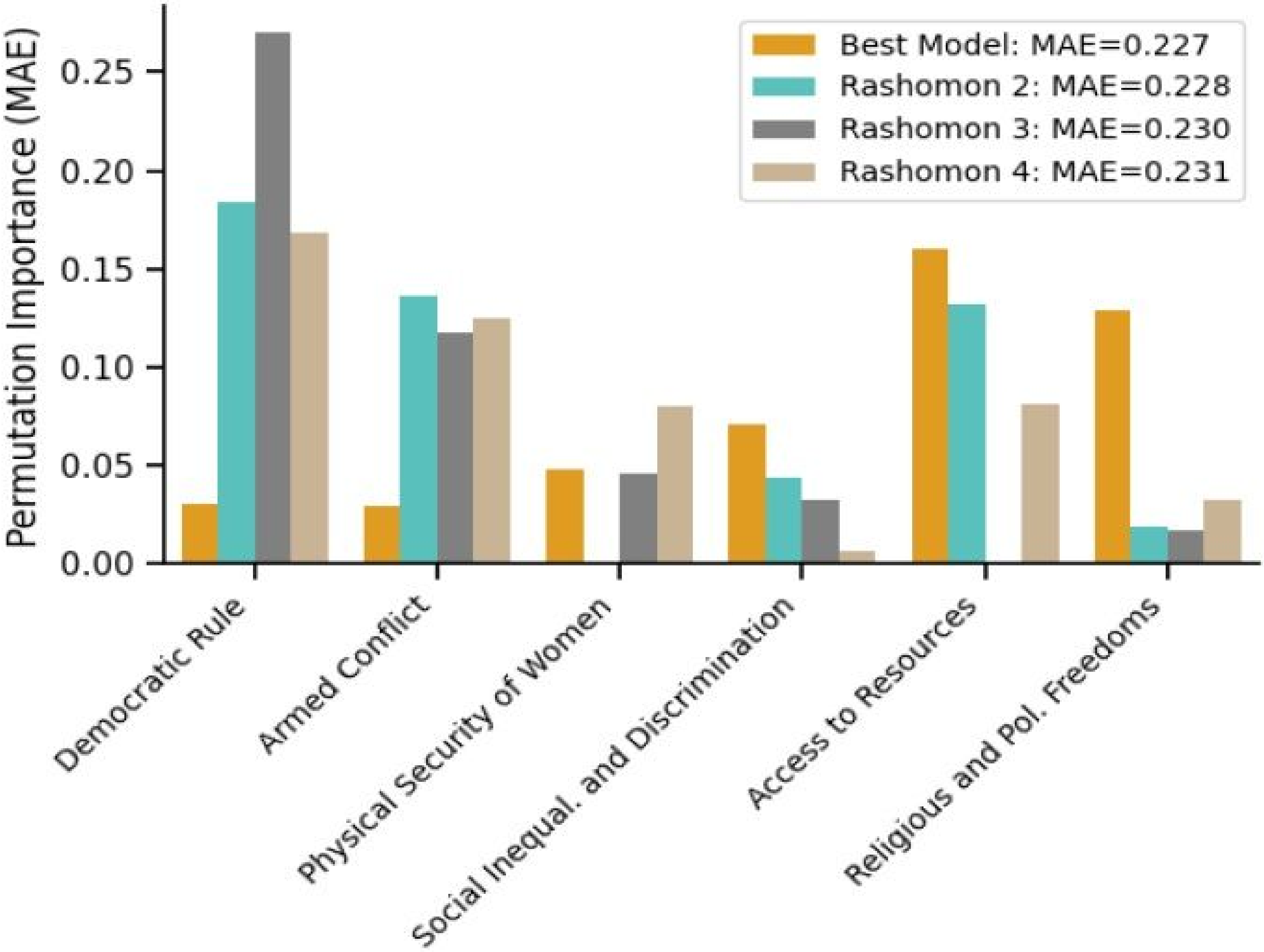


Figure 5

Comparison of variable importance ratings all models in the Rashomon set, omitting those in the same model class as the best model (NMF->DT (MF).

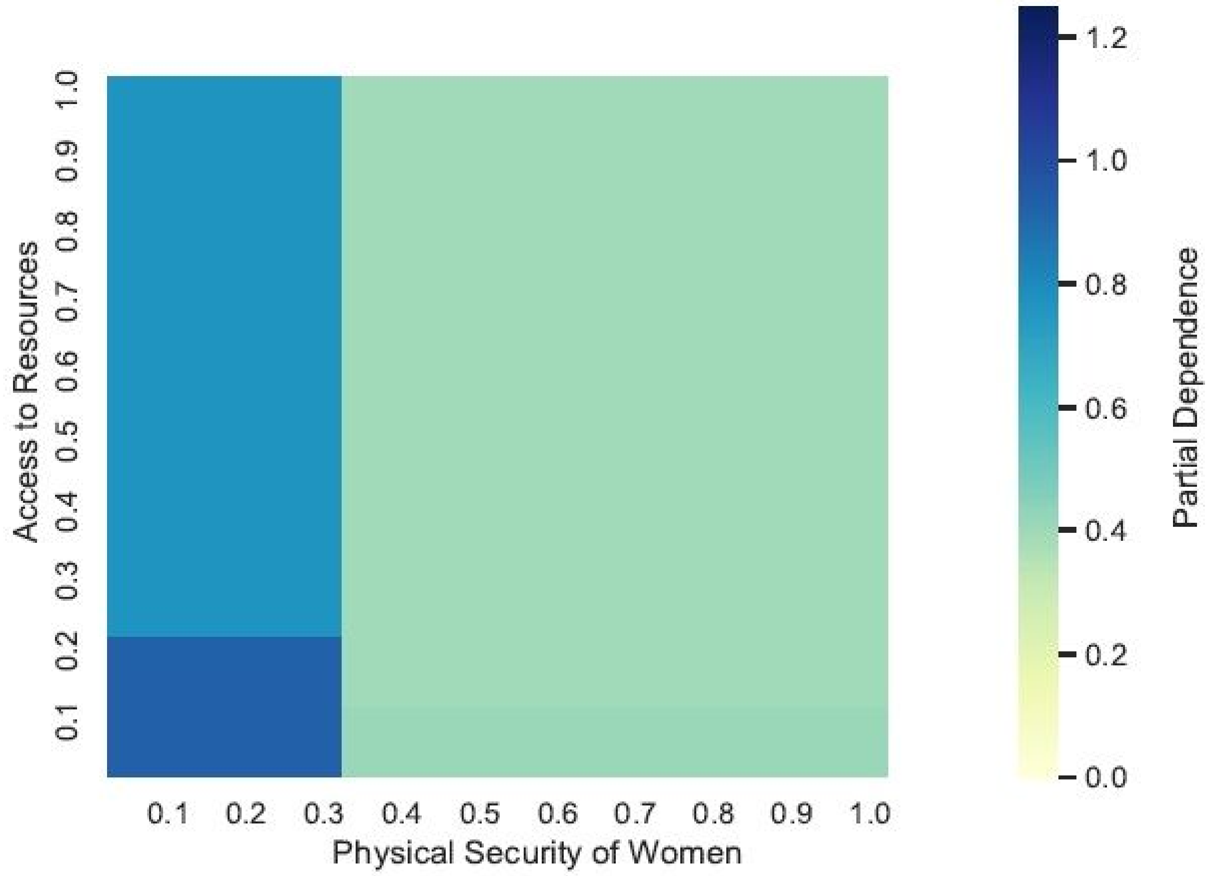


Figure 6

A heat map illustrating that (for the best performing model) the partial dependency of the prevalence prediction is especially high when both Access to Resources and Physical Security of Women are low. Here, a lower score for Physical Security of Women indicates less security.

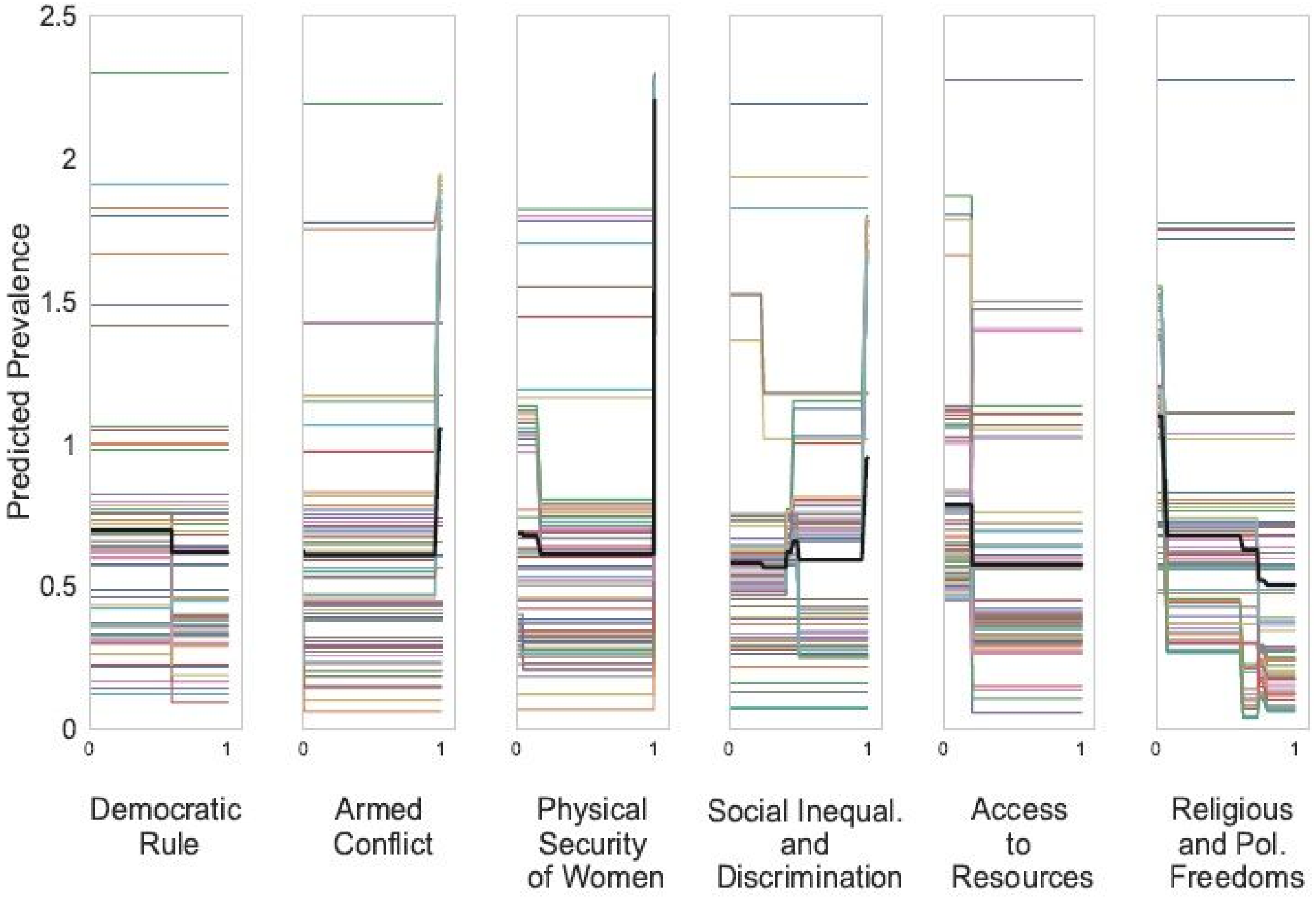


Figure 7

Individual Conditional Expectation (ICE) plots to illustrate the non-linear effects of components on the best model's predictions. Each coloured line is a country-year data point with some jitter applied. The thick black line is the average partial dependency of the component on the prevalence prediction.

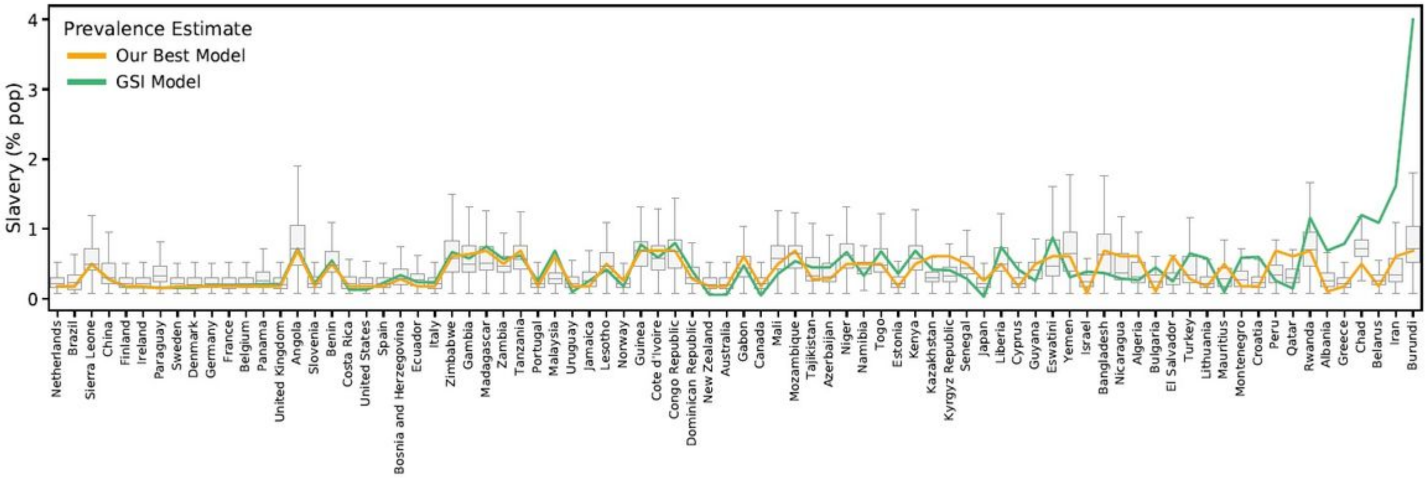


Figure 8

The predictions made for the 2018 prevalence of slavery (individuals enslaved as a % of the population) in countries where no GWP survey data exists. The estimates made by the best performing model are compared to those made by the GSI. The grey box plots illustrate the distribution of 10,000 bootstrapped LOOCV predictions (using the full pipeline NMF->DT) to help illustrate the uncertainty associated with our model's predictions. The box shows the quartiles of the bootstrapped predictions while the whiskers extend to show the rest of the distribution, except for points that were determined to be "outliers" (using a function of the inter-quartile range) which are not plotted. The x axis is ordered by the disagreement between our model's predictions and the predictions made by the GSI model. The countries displayed are those which had <10% missing independent variable data. For all 172 out-of-sample predictions and information on missing data see Appendix, Table A1.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Variabledescriptions.pdf