9-2015

Building a Better Model: Variable Selection to Predict Poverty in Pakistan and Sri Lanka

Marium Afzal
World Bank

Jonathan Hersh
Chapman University, hersh@chapman.edu

David Newhouse
World Bank

Building a Better Model: Variable Selection to Predict Poverty in Pakistan and Sri Lanka

Building a better model: Variable selection to predict poverty in Pakistan and Sri Lanka

Marium Afzal[1]

Jonathan Hersh[2]

David Newhouse[3]

September 2015

PRELIMINARY AND INCOMPLETE. PLEASE DO NOT CITE OR CIRCULATE.

Abstract

Numerous studies have developed models to predict poverty, but surprisingly few have rigorously examined different approaches to developing prediction models. This paper applies out of sample validation techniques to household data from Pakistan and Sri Lanka, to compare the accuracy of regional poverty predictions from models derived using manual selection, stepwise regression, and Lasso-based procedures. It also examines how much incorporating publically available satellite data into the model improves its accuracy. The five main findings are that: 1) Lasso tends to outperform both discretionary and stepwise models in Pakistan, where the set of potential predictors is large. 2) Lasso and stepwise models give comparable results in Sri Lanka, where the set of predictors is smaller. 3) The accuracy of the prediction model depends considerably on the poverty threshold 4) Including publically available satellite data makes poverty predictions more accurate in Sri Lanka, where predictors are scarce, but slightly less accurate in Pakistan and 5) Including the satellite data increases the benefit of using Lasso in Sri Lanka. We conclude that among the three model selection methods considered, lasso-based models are preferred for generating poverty predictions, especially when the pool of candidate variables is large. Furthermore, when the pool of candidate variables available from household surveys is smaller, incorporating publicly available satellite data can considerably improve the accuracy of regional poverty predictions.

Keywords: model selection, poverty mapping, poverty estimation, machine learning

JEL classification: I32, C50

[1] mafzal@worldbank.org, Poverty Global Practice, World Bank, 1818 H Street, Washington DC, 20433

[2] jhersh@worldbank.org, jhersh@bu.edu, Department of Economics, Boston University, 270 Bay State Road, Boston, MA 02215, and Poverty Global Practice, World Bank.

[3] dnewhouse@worldbank.org, Poverty Global Practice, World Bank, 1818 H Street, Washington DC, 20433

1. Introduction

Given the proliferation of different types of household data, survey to survey imputation, defined as predicting a variable present in one survey into another using variables common to both, is becoming increasingly popular. Survey to survey imputation allows analysts to examine the relationship between variables found in two different surveys, if they were collected at roughly the same time and represent the same population.[4] One important application is to impute consumption, which is the primary indicator of household economic welfare in most low and lower middle-income countries, into labor force or demographic and health surveys that do not collect consumption data, in order to examine the labor or health outcomes of the poor. Another common application is to generate small area estimates of poverty by predicting consumption or income into a larger target dataset, such as a census, that is representative at a more disaggregated geographic level.[5]

Despite the increasing popularity of survey to survey imputation, economists have devoted little attention to determining how best to select models from a potentially large set of common variables. In a series of papers, Leamer (1983, 1985) outlined a method for global sensitivity analysis he called extreme bound analysis to evaluate the robustness of covariates in econometric models. Except for a few prominent examples (Fernandez et. al, 2001; Levine and Renelt, 1992) this line of research has had little impact in how economists construct models. Heckman, et al (2014) tests the robustness of model selection by considering the distribution of coefficients across a variety of potential specifications, but this approach also has yet to be widely adopted. Survey to survey imputation is a natural context to consider model selection methodology in a rigorous way, since the accuracy of the prediction in the target survey depends heavily on the model used to generate it.

This paper tests three methods of model selection in the context of estimating relative poverty rates in different regions of Pakistan and Sri Lanka. The three methods are: Manual selection, where a researcher uses a mix of judgment and goodness-of-fit measures to select a model, forward stepwise regression using a p-value threshold of 0.05 as the inclusion criteria, and post-lasso regularized regression. The resulting variables are used to predict poverty, including stochastic error terms generated using a non-parametric version of the ELL estimator developed in Elbers, Lanjouw, and Lanjouw (2003). Out-of-sample cross-validation techniques are used to assess the accuracy of the prediction of the share of the population in the bottom 10, 20, 30, and

[4] Survey to survey imputation can also be used to track changes over time in some cases, but the maintained assumptions are far stronger and may not always hold (Newhouse et al, 2014).
[5] References to a number of poverty maps can be found at http://web.worldbank.org/WBSITE/EXTERNAL/TOPICS/EXTPOVERTY/EXTPA/0,,contentMDK:20239128~menuPK:462078~pagePK:148956~piPK:216618~theSitePK:430367~isCURL:Y,00.html

40 percent of the household size-adjusted consumption distribution, across 8 urban and rural areas of each province in Pakistan and 25 districts in Sri Lanka.[6]

While the problem of selecting which control variables to include in a model might seem innocuous, the results show anything but: we find the accuracy of poverty estimation greatly depends on the choice of covariates used to approximate the data generating process. The gains to different model selection methods, however, depend both on the context in which they are employed and on the relative poverty threshold used to classify households as poor. Our preferred measure of prediction accuracy is the average absolute value of the discrepancy between the predicted and actual poverty rates, averaged across regions or districts. By this measure, the gains to using the Lasso estimator are much more apparent in Pakistan, where 138 variables are available to build a model, than in Sri Lanka where only 71 variables are available. When predicting the share of the population in the bottom 30 and 40 percent of the consumption distribution in Pakistan, the Lasso model is twice as accurate as the stepwise model, and two to four times more accurate than the manually selected model. In Sri Lanka, on the other hand, stepwise was if anything slightly more accurate than lasso, and each was 30 to 60 percent more accurate than the manual model.

Besides the choice of model selection method, a related question is how much adding publicly available ancillary data, taken from satellite photography, improves the accuracy of the prediction. We therefore examine the effects of adding approximately 35 variables, such as night time lights, elevation, and the EVRI vegetation index. In Pakistan, the additional of these variables generally makes the models slightly less accurate across all prediction methods. The one exception is when using post-lasso to predict the bottom 40 percent, in which case the spatial variables increase accuracy by 58 percent. After including the satellite data in Sri Lanka, the lasso-based estimates improve substantially, on the order of 20 to 25 percent, when predicting membership in the bottom 20, 30, and 40 percent of the welfare distribution. Furthermore, after these new variables are included in the set of candidate predictors, lasso outperform stepwise by a considerable margin in Sri Lanka. Therefore, when the 35 spatial variables are included, the lasso model strictly dominates the stepwise and manually selected models in accurately predicting poverty.

The rest of the paper is laid out as follows: the remainder of section I gives a short overview of poverty estimation literature and specifically the ELL/small are estimation method for poverty estimation. Section II describes the methodology employed in detail, including defining the Post-Lasso ELL estimator, and describing how external cross-validation techniques are used to evaluate the accuracy of different methods. Section III presents the main model selection results using the set of covariates derives from the household survey. Section IV considers the addition

[6] This is the lowest level at which the poverty surveys are representative.

of approximately 35 variables generated from publicly available satellite data, to see how much they improve prediction accuracy for different methods. Section V concludes.

1.2 Literature Review

Small area estimation methods have gained traction because direct estimators cannot accurately provide conclusions about 'small areas' or subpopulation from survey data (e.g., Ghosh and Rao 1994, Haslett et al 2010). Most household surveys, as opposed to censuses, contain a large number of variables, but have a relatively small sample size. It is likely that a majority of 'small areas' will not be sampled from at all, or will contain a handful of observations. Furthermore, if the poverty rate is less than 50 percent, the precision of the poverty estimate also suffers as poverty rates decline. A number of indirect estimators have therefore been developed and applied to larger surveys, which don't contain consumption, to estimate poverty at more granular geographical levels. Elbers, Lanjouw and Lanjouw developed a methodology which is widely used especially at the World Bank (Elbers et al 2000, 2003). Briefly, the approach is to use survey data to estimate a model for consumption expenditure (or, alternatively, income) using variables present in both the survey and a larger dataset, such as a census. This predicted expenditure is then used to predict poverty or other measures of welfare (Elbers et al 2003)[7].

Molina and Rao (2010) subsequently proposed an improvement to ELL by merging it with the empirical Bayes (EB) method. In their simulations, the authors find that ELL has a slightly greater bias than EB, and a significantly larger prediction error variance, even larger than direct estimators. EB especially shows improvement over ELL in areas that are represented in the detailed survey, by reducing the random area effects. This improved variant of the ELL methodology was subsequently incorporated in the latest version of PovMap[8] (van der Weide 2014) a software program often used at the World Bank to generate small area poverty estimates. Van der Weide proposes further improvements to EB by relaxing the assumption of homoscedastic errors maintained by Molina and Rao (2010), as well as proposing modifications to GLS to improve the estimation of model parameters. A major remaining critique of ELL is that it assumes that the error terms, representing unobserved consumption, are not correlated across clusters. In situations where even minor area fixed effects or intracluster correlations are present, ELL will tend to underestimate the variance in errors. Tarozzi and Deaton (2009) show that this underestimation of error can be significant.

Apart from ELL, several alternate methodologies have also been developed to address the problem of small area estimation. Ghosh and Rao (1994) and Rao (2003) review these in detail,

[7] Demombynes et al (2003) evaluate ELL and its variants in terms of accuracy of confidence intervals, bias and correlation with true values, and the factors that affect each (2002).
[8] PovMap 2.5 is available for free download from: iresearch.worldbank.org/PovMap/

particularly empirical Bayes, hierarchical Bayes and empirical best linear unbiased predictor that have been used widely, as well as various design-based estimators.[9] Haslett (2010) reviews and conducts a comparison of ELL with two other methodologies: spatial microsimulation, which is primarily developed by geographers; and mass imputation, a statistical technique that is similar in principle to ELL.

In short, most of the substantial literature on this topic is concerned with improving the methodology used to generate estimates of the error term, conditional on a set of predictors. Most papers in this literature take as given that the researcher has identified the true model, which is the core assumption we are relaxing here. As we show below, however, the specification of the independent variables also has major implications on the accuracy of the estimates.

2. Methodology

Numerous methodological approaches exist to select models for prediction. We distinguish between two broad classes here: *manual model selection*, where covariates for the prediction model are chosen directly by the researcher; and *algorithmic model selection*, in which the researcher employs an algorithm to build a prediction model. It is far from clear ex-ante that one approach strictly dominates the other. In cases where one has strong prior information regarding the data generating process, a manual approach to model selection may be appropriate. Absent strong priors, algorithmic model selection mechanisms may better minimize model error. In practice, researchers may also use a blend of approaches, beginning with an algorithmic model selection approach, and then removing or adding covariates depending upon their strong priors for inclusions or exclusion. Although we do not test blended models below, such an approach may balance the pros and cons of each approach.

Applying the problem to real-world poverty data, we consider manual models that were developed by researchers to predict poverty rates at the sub-national level. For Pakistan, we use the model developed and published by a researcher affiliated with a Pakistan university.[10] This model was built for the purposes of predicting poverty at the district level using a direct OLS estimator, meaning there is no simulation of the error term as is the case with ELL or its variants. For Sri Lanka, we use a model developed by the Department of Census and Statistics and the World Bank for the purposes of predicting poverty at the district level using an ELL methodology framework. (Department of Census and Statistics and World Bank, forthcoming)

[9] Pfefferman (2010) updated the review with updates and variations to the methodologies mentioned in Rao's appraisal.
[10] A proper citation is available from the authors at the reader's request.

For the latter, we adapt the modeling approach by estimating one model of the entire country rather than 22 separate models for different geographic areas.[11]

For algorithmic model selection approaches, we consider two common methods for model selection: forward stepwise using a p-value as selection criteria and a Lasso estimator with Bayesian shrinkage for model selection. Since the lasso estimator has only recently become popular among economists, we describe this estimator below.

2.1 Description of Lasso Estimator

The Lasso estimator is a member of the family of regularized regression estimators first developed by Tibshirani (1996)[12]. Regularization refers to adding a component to the typical loss function, which is the residual sum of squares, that penalizes the inclusion of additional covariates. To be explicit, the lasso estimator β_{lasso} solves the optimization problem:

$$(1) \quad \beta_{lasso} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \underbrace{\sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{K} x_{ij} \beta_j \right)^2}_{Sum\ of\ squared\ residuals} + \underbrace{\lambda \sum_{j=1}^{K} |\beta_j|}_{Shrinkage\ factor} \right\}$$

Where $i = 1 \ldots . N$ indexes the number of observations and $j = 1 \ldots K$ indexes the number of parameters to search over. The left hand side of the objective function is identical to the residual sum of squares loss function from an unconstrained OLS regression. The novel component here is the right component in the optimization problem, which applies an l_1 loss function over the coefficients, which are then summed across all coefficients and multiplied by λ, the factor which determines the degree of Bayesian shrinkage for the problem. The choice of λ is left unspecified by theory. As $\lambda \to 0$ the objective function becomes the OLS objective function and $\beta_{lasso} \to \beta_{OLS}$. However, for any positive value of λ the coefficients of β_{lasso} will deviate from the OLS solution, and as $\lambda \to \infty$ β_{lasso} converges to the zero vector of dimension K, implying that all coefficient estimates will have been "shrunk" to zero. The coefficient estimates therefore depend heavily on the choice parameter of λ. In practice, this parameter is chosen through cross validation.[13] Before computation, it is standard to center variables around a mean of zero and standard deviation of one, and then present the untransformed version of the coefficients.

---

[11] We build country-level models for the purpose of simplicity in comparison. Our algorithmic modeling approach results in more covariates being selected in general than ad hoc methods, however this is not an artifact of building country-level models versus provincial level ones. When using algorithmic models on provincial level data our results generate roughly the same number of selected covariates as when building country-level models.

[12] For more on the history of the Lasso estimator see the review article Tibshirani (2011).

[13] Several reasonable options for the choice of λ exist. The canonical choice is the value which minimizes cross-validated mean squared error (MSE), λ_{min}. However, if a more parsimonious model is desired the choice of λ is often parameterized at the value of λ_{min} plus one estimated standard error of λ (Hastie et. al, 2009). This particular

The Lasso estimator provides variable selection by penalizing the model based on the sum of the absolute value of the standardized coefficients. Optimizing this modified objective function sets some variables to zero. We consider a variable selected by the Lasso estimator if they remain non-zero after optimizing the objective function. The nature of the shrinkage path for the coefficients is not always monotonically decreasing towards zero; in most cases $\beta_{OLS,j} > \beta_{Lasso,j}$ although this is not universally true for all values of λ. However, it is true for all non-pathological cases that if a coefficient is estimated at zero given some value of λ, larger values of λ, say $\tilde{\lambda}$ such that $\tilde{\lambda} > \lambda$, will results in that coefficient estimate remaining at zero. The Lasso estimator, in a sense, is weakly monotonic in λ with respect to shrinking coefficients to zero.

The Lasso estimator has been applied to economic problems in a variety of contexts. Varian (2014) gives an overview of this method and provides some examples. Bajari et. al (2015) applies it to the setting of estimating a demand function from a large set of possible covariates. Baxter and Hersh (2015) use it in the context of estimating robust covariates associated with aggregate bilateral trade flows between countries. Various extensions to the estimator have been proposed, both in economics and statistics. Belloni and Chernozhukov (2013) propose a two-step estimator ("Post-Lasso") where the first stage uses the shrinkage property of Lasso for variable selection and in the second stage, OLS coefficients are estimated over the reduced set of non-zero coefficients in the first stage. It is in the spirit of the Post-Lasso estimator of Belloni and Chernozhukov that we propose a Post-Lasso ELL algorithm. This algorithm first estimates first a Lasso model over a large set of possible coefficients, then uses the reduce set of covariates that remain non-zero after the Lasso step in the ELL framework to estimate and simulate the error term. We now move to a more formal discussion of the Post-Lasso ELL algorithm.

2.2 Post-Lasso ELL Algorithm

The Post-Lasso ELL algorithm is defined as follows:

1. Estimate a Lasso model on the training dataset, typically a household survey, containing information on household consumption and household level covariates. For choice of covariates, we initially use the largest set of reasonable coefficients for the prediction problem.[14] To choose the Bayesian shrinkage parameter, we employ cross-validation techniques and use the more parsimonious version of the optimal shrinkage parameter,

parameterization is chosen so that it results in the simplest model "whose accuracy is comparable with the best model." (Krstajic et. al, 2014).

[14] In our examples, we only consider only linear models of consumption, with the exception of a squared age and education of the head of household. This framework extends easily to non-linear functions of income. For those concerned about the hierarchical restriction of interactions, we recommend using the formulation of the Lasso estimator due to Bien et. al (2013) which will obey the hierarchical restriction of interactions when the Bayesian shrinkage parameter is applied.

that is $\lambda = lambda.1se$, or lambda plus the standard error of the λ which minimizes cross-validated MSE.

2. Letting $\hat{\beta}_{lasso}$ be the set of variables whose coefficients remain non-zero after step 1, estimate an OLS model with random effects model of the form:

$$y_{c,h} = X_{c,h}^T \hat{\beta}_{Post-lasso} + \tilde{\eta}_c + \epsilon_{c,h}$$

We follow conventions in letting $\tilde{\eta}_c$ be the random intercept cluster-level error – typically the sampling unit level – and $\epsilon_{c,h}$ is the household level error. $y_{c,h}$ is the welfare measure of interest, typically log consumption, for household h in cluster c.

3. Draw random effects $\tilde{\eta}_c$ for each cluster for R simulations. Specifically, use $\widehat{\sigma}_c$ – the estimate of the cluster level variance from step 2 – to draw $\{\tilde{\eta}_c^s\}_{s=1}^R$, or R values of η_c for each unique cluster in the test set where each $\tilde{\eta}_c \sim N(0, \hat{\sigma}_c^2)$. In our examples we set $R = 100$.
4. For each simulation, compute the predicted consumption expenditure, $\tilde{y}_{c,h}^s$ for every household on the test set using the drawn cluster simulation $\tilde{\eta}_c^s$:

$$\tilde{y}_{c,h}^s = X_{c,h}^T \ \hat{\beta}_{Post-lasso} + \tilde{\eta}_c^s$$

Where $\hat{\beta}_{Post-lasso}$ is the estimate of betas obtained from random effects[15] For each simulation, a vector of household errors can be defined as the discrepancy between actual and predicted consumption, or $\tilde{\epsilon}_{c,h}^s = y_{c,h} - \tilde{y}_{c,h}^s$ where $y_{c,h}$ is reported household consumption.

5. Sample household idiosyncratic error with replacement from each simulation and add it to $\tilde{y}_{c,h}^s$. Formally we define the sample idiosyncratic error component as $\tilde{\epsilon}_{c,h}^i = \tilde{y}_{c,h}^i - X_{c,h}^T \ \hat{\beta}_{Post-lasso} + \tilde{\eta}_c^i$. Thus our final simulated income is given by $\tilde{y}_{c,h}^s = X_{c,h}^T \ \hat{\beta}_{Post-lasso} + \tilde{\eta}_c^s + \tilde{\epsilon}_{c,h}^i$, where $i \neq s$.

6. Calculate mean and variances of $\tilde{y}_{c,h}^s$. Predicted values are defined as $\hat{y}_{c,h}^{PLELL} \stackrel{\text{def}}{=} E\left[\frac{1}{100}\sum_{s=1}^{100} \tilde{y}_{c,h}^s\right]$ and variance is given by the typical variance formula. Residuals for the estimator are defined by $\hat{y}_{c,h}^{PLELL} - y_{c,h} = \tilde{r}^{PLELL}$ where $y_{c,h}$ is the true level household consumption.
7. Repeat until all test sets have been estimated.

[15] Note, we use the coefficients from the random effects model, and not those which have had Bayesian shrinkage applied. Lasso is only used for model selection in the fashion of Post-Lasso (Belloni and Chernozhukov, 2013). Using the Lasso, i.e. shrunken, coefficients produces qualitatively similar results.

8. Let $\widehat{W}_r$ be the poverty statistic of interest, such as FGT_0, poverty headcount, or FGT_1, the poverty gap index, for a given simulation r. We simulate the expected value for the indicator over the mean of the simulations

$$\tilde{\mu} = \frac{1}{R}\sum_{r} \widehat{W}_r$$

The final statistic of interest here is $\tilde{\mu}$, that is the poverty statistic of interest averaged over the 100 simulations.

2.3 Stepwise Algorithm

We utilize a forward stepwise algorithm where the p-value of a coefficient is used as the inclusion criteria[16], parameterized at $p = 0.05$. It's possible, and in fact recommended, that this hyper-parameter p-value criterion is selected through the use of cross-validation. Although this methodology has recently been adopted in some recent World Bank predictions of poverty, it still is rarely used in practice and existing software does not easily support this approach. We therefore define our stepwise algorithm with a fixed prior p-value hyper-parameter. There are several disadvantages to the stepwise algorithm. First, the algorithm results in a non-convex objective function, which means a global minimum is not guaranteed. It further can be very computationally intensive, requiring the bulk of our computational time in our simulation exercise.[17] The non-convexity often results in discrete jumps in mean squared error (MSE) when comparing across modeling approaches – such as the selection of a hyper-parameter, for example – which can complicate the selection of hyper-parameters. Finally, the presence of highly correlated independent variables can lead to model instability.

2.4 Construction of Relative-Poverty Rates

To determine the accuracy of each modeling approach we must create a baseline from which compare each estimator's performance. For each country, we build relative poverty rates for each region, which we define as the share of sample individuals in each region whose household welfare, which is per capita consumption in Sri Lanka and per adult equivalent consumption in Pakistan, falls below a given percentile of the national distribution.[18] We select relative poverty rates at the 10%, 20%, 30% and 40% of the welfare distribution. This will further help us

[16] For an explicit definition of the forward stepwise algorithm we refer the reader to Hastie et al. (2009).

[17] The stepwise algorithms in this paper required computational time on a desktop machine of around 2-4 hours for each country. In comparison, the other methods were computed in a matter of minutes or in some cases seconds.

[18] The per adult equivalence measure in Pakistan gives a weight of 0.8 to children under the age of 18 and 1 to adults over 18. It is used for calculating national poverty statistics in Pakistan, while per capita consumption is used in Sri Lanka.

understand how the accuracy of the estimators depend on the incidence of poverty. Since we are concerned about the representativeness of the poverty estimates, we define regions according to the most disaggregated geographic level at which the consumption survey is considered to be representative. For the Pakistan Household Income and Expenditure Survey, this is the province/urban-rural level, and since there are four provinces in our sample this gives us a total of eight regions. For Sri Lanka, the HIES is considered to be representative at the District level, therefore we will construct the relative poverty rates for a total of 25 regions in our sample.

2.5 The Importance of External Cross Validation for Evaluating Out-of-Sample Performance

To assess the performance of each modeling method we use external k-fold cross-validation to fit predicted consumption from each modeling approach. Why not just fit a consumption model over the data and use the fitted model to generated predicted values? The concern is that this would produce a good in-sample fit, but doesn't guarantee a high performance out of sample. We instead apply a K-fold cross-validation approach. This involves partitioning the data into several training and test folds, fitting a model on the training set and predicting into the withheld fold, and repeating the process until all withheld folds have been used for prediction.

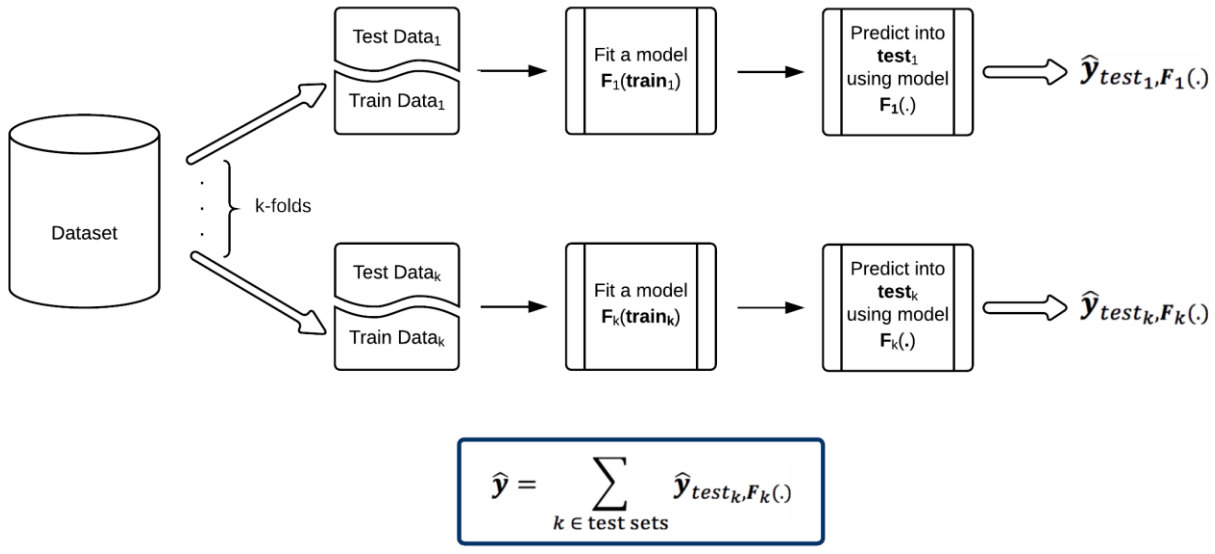


FIGURE 1: ALGORITHM FOR EXTERNAL CROSS VALIDATION

The algorithm for external exhaustive k-fold cross validation, which is shown in figure 1, is as follows: first, setting $k = 10$, we partition the data into 10 folds of equal size. Starting with fold 1, we fit a model using folds 2 through k of the data, estimating the model $f_1(X_{2\ldots k})$. Using this estimated model we predict into the withheld fold, X_1, generating predicted values $\hat{y}_{test_1,f_1(.)}$

which is a column vector of dimension $rowrank(X)/k$. We repeat until all folds have been withheld, and we have predicted values for all observations in X. In our Lasso model selection methodology, we choose the Bayesian shrinkage parameter λ through cross validation within each testing fold.

3. Model selection Results

3.1 Pakistan Data

We first consider the performance of the model selection methods in Pakistan. We utilize a sample of 16,340 households from the 2010-11 round of the Household Income Expenditure Survey (HIES), which is part of the Pakistan Social and Living Standards Measurement Survey (PSLM). In total the PSLM surveyed 76,546 households that year in Pakistan, though only a 21 percent subset of these were asked the consumption module (HIES) which constitutes our core sample. Since the PSLM is considered to be representative at the district level, small area estimation techniques can be utilized to generate estimates at the district level.

The HIES/PSLM is a rich survey, covering topics related to household education, employment, health, assets, amenities, housing quality and sanitation, and other facilities related to the Millennium Development Goals (MDGs). Table 1 presents the summary statistics for these variables and gives some sense of the richness of the dataset, which includes many variables on household level assets as well as some unconventional variables such as "time to water source" and "area economic assessment". In total, we consider 138 different variables in the set of possible variables each modeling approach can utilize. Clearly a model which utilizes all possible variables will suffer from issues of overfitting, therefore this presents a particularly good test for the modeling approaches, to see which ones are capable of identifying the optimal statistical model in terms of out of sample performance.

The survey was designed to be representative at the urban/rural provincial level, therefore we identify the region – which will become our testing area to compare model performance – at this level. The HIES only covers the four main provinces of Punjab, Sindh, Khyber Pakhtunkhwa (North-West Frontier Provinces), and Baluchistan, and as noted above, is only considered to be representative at the urban and rural level of each province. Therefore we are limited to defining 8 regions for the purposes of calculating error between predicted and actual relative poverty statistics.[19] Defining a small number of regions is a clear disadvantage using this approach; but

[19] To be explicit our 8 regions are: Punjab-rural, Punjab-urban, Sindh-rural, Sindh-urban, KP-rural, KP-urban, Baluchistan-rural, and Baluchistan-urban.

given the importance of cross-validating poverty rates using the HIES, it is not clear there is a better alternative.

We consider the performance of four separate modeling approaches: (1) OLS with manual model selection, (2) ELL (2003) with ad hoc model selection, (3) Post-Lasso ELL and (4) Stepwise ELL. [20] For the manually selected model, we utilize a model of household consumption that includes the following covariates: number of household members, household dependency ration, head of household years of education, spouse years of education, highest education level in household, a dummy if the head of household of less than 40 years of age, a dummy if the head is unemployed, a dummy if the head is employed with a consistent wage, a household asset score (sum of number of household asset), and finally provincial dummies. This model is used to generate both the manual OLS and the manual ELL estimates, with the only difference being the manual ELL model will follow ELL (2003) in estimating a cluster level effect at the primary sampling unit level, and simulating draws from this cluster level effect and from the household level residuals. The manual OLS model, on the other hand, assumes that the error term is zero for each household, and therefore compares exponentiated predicted log per capita consumption to the poverty line to determine if a household is poor or not.

3.2 Pakistan Performance Comparison

After estimation using modeling approaches (1)-(4), we derived optimal models using lasso and stepwise, the results of which are summarized in table 2. Panel A shows the results for Pakistan. The first column shows the average number of variables selected across the $k = 10$ folds. For ad hoc OLS and ad hoc ELL the number of variables selected is set manually at 20.[21] For the Post-Lasso ELL algorithm, an average of 62.2 variable were selected across folds, whereas for stepwise this results in an average of 105 variables selected. The stepwise algorithms selected more variables than the Lasso algorithm. The average R^2 between Post-Lasso ELL and stepwise models are nearly identical, at 0.68 and 0.67 respectively, whereas the manual OLS and ELL models have a lower average R^2 of 0.53 and 0.51 respectively. The next column shows the average household level consumption residual, that is $\hat{y}_i - y_i$, where y_i is the consumption variable in logs. All of these estimators appear unbiased, showing very low average residuals of between -0.0158 and 0.0299. In this table we further present the standard deviation, min and max of the household level residuals.

The performance in terms of generating region poverty rates is presented in table 4, and summarized in the top left panel of figure 2. We present three measures of region poverty rate

[20] Post-Lasso ELL refers to the estimation of a non-parametric ELL model using coefficients selected by Lasso.
[21] In the number of variables we partition factor variables into binary dummy variables, thus each distinct level in a factor is considered a separate dummy variable.

accuracy: mean region error, which is defined as $\frac{1}{N}\sum_{j=1}^{N}(\hat{y_j} - y_j)$, where j indexes each region; mean region absolute error, which is defined as $\frac{1}{N}\sum_{j=1}^{N} abs(\hat{y_j} - y_j)$, and mean region weighted absolute error, $\frac{1}{N}\sum_{j=1}^{N} w_j * abs(\hat{y_j} - y_j)$, where w_j is the weight of each region j[22] determined by its population size. All of these measures present error in terms of average percentage points across regions. Our preferred measure is mean region weighted absolute error (MWAE), which accounts for both absolute error differences and adjust for population differentials across regions.

The results are stark: in 1 out of 4 examples stepwise ELL performs the worst, and further in all but one case Post-Lasso ELL outperforms all other methods as a poverty estimator. The differences become more pronounced as the relative poverty rate decreases. In predicting the below 40% poverty rate, Post-Lasso ELL is slightly worse than stepwise, at 2.118 versus 1.47 MWAE. However, at the 10% relative poverty rate, Post-Lasso ELL greatly outperforms stepwise, with error rates of 3.391 versus 8.347. The manual model specifications perform somewhere in between, with manual OLS performing the worst in 3 out of 4 examples. Average error rates for manual ELL, the method typically used to build models, are 4.4 at the 10% relative poverty rate (RPR), 3.4 for the 20% RPR, 3.5 for the 30% RPR, and 4.3 for the 40% RPR. Both the Post-Lasso ELL model and the Stepwise ELL algorithms improve as the relative poverty rate increases, suggesting that these algorithms are better able to make use of the richness of the datasets to predict consumption at higher levels of consumption.

This example demonstrates three main points. The first simply confirms the importance of simulating error terms; using the ELL approach, not surprisingly, leads to far more accurate poverty estimates than setting the error term equal to zero. Secondly, in the context of a "data rich" poverty estimation environment, in which a multitude of variables are available, Lasso outperforms other methods of model selection and with one exception, shows uniformly lower error rates across relative poverty rates. Third, in-sample R^2 may not be an accurate measure of model performance, since models based on Lasso and stepwise produce similar R^2 values even though Lasso performs far better out of sample.

[22] Weights are used to compare error rates for regions that have different population counts. Weights are defined as $\frac{p_j}{\bar{P}}$, where p_j is region population, and $\bar{P}$ is average region population.

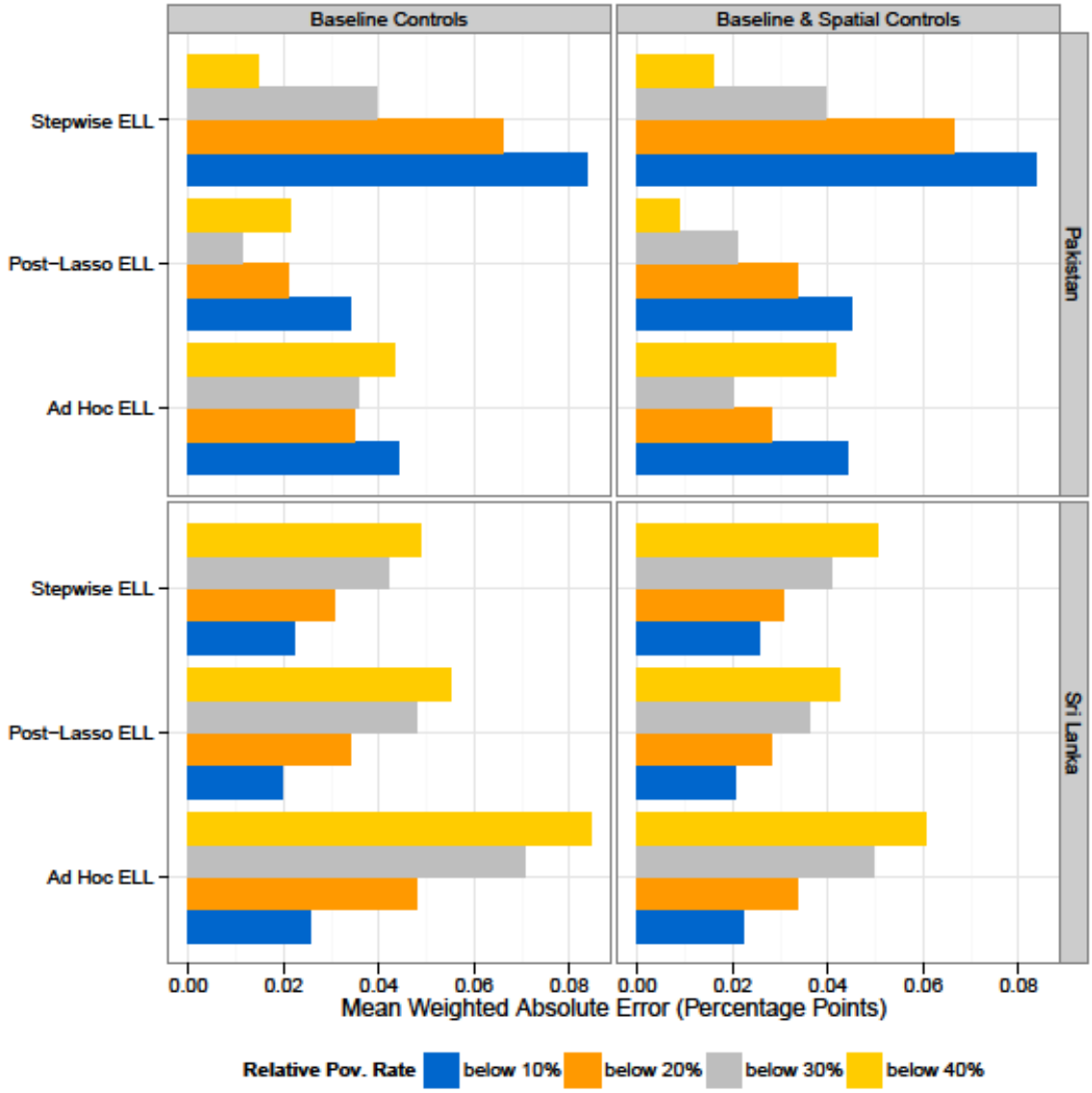


FIGURE 2: PREDICTION COMPARISON, BY MODEL SELECTION METHOD AND RELATIVE POVERTY RATE (SHORTER BARS INDICATE BETTER PERFORMANCE)

3.3 Sri Lankan Data

We turn now to the problem of estimating poverty in Sri Lanka. In contrast to the dataset for Pakistan, when building poverty rates for Sri Lanka we limit ourselves to using only the common variables available in both the Sri Lankan Census and the Household Income and Expenditure

Survey. This results many fewer variables that we were able to use above for Pakistan. We use a sample of 20,540 households in the 2012/13 Household Income and Expenditure Survey (HIES). The sampling frame of the HIES is representative at the district level, therefore we define the region to be used for the purposes of calculating poverty statistics at the district level. Since there are 25 districts, we will have 25 regions from which to compare average error rates in terms of calculating poverty.

The summary statistics for the possible variables that each algorithm can select are shown in table 4. In total, there are 71 common variables from which to select to build a model. This is only slightly over half of the total number of variables at our disposal for the models in Pakistan. We are purposefully limiting ourselves to the set of variables available in the Sri Lankan census, which is small relative to the HIES consumption survey we used in Pakistan. Noticeably absent is the richness of household asset and employment variables in the Pakistan HIES. In the Sri Lanka data detailed spousal and head information on education or employment status is not available, nor do we have any or as much detailed information roof type, water source, toilet type, cooking fuel, lighting fuel, residence type, family and area subjective economic assessment, phone type, household primary language, and time to travel from the household to key public services.

For the manual model selection we utilize a model recently used to construct poverty estimates at the DS division level developed by the World Bank and the Sri Lankan Department of Census and Statistics (2015). That model includes the following independent variables: a dummy for male household head, age of head, employment status of head, household size, household dependency ratio, highest education in household, a dummy if the household uses firewood for heating, a dummy if the household has access to electricity, a dummy if the house is owned, indicators if walls and roof are of high type, an indicator if the household indicates it has safe drinking water, and finally dummy variables if the household contains the assets: toilet, waterseal, radio, television, landline based phone or mobile phone. The total number of variables used in the manual model is 21.

3.4 Sri Lankan Performance Comparison

The performance in terms of average region level error rates in relative poverty rate is shown in the lower left panel of figure 2, and table 2 shows a summary of the performance of each algorithm at the household level. The manual model selection uses 21 variables, and both the ELL and OLS variants share an R^2 value of 0.42. The Lasso algorithm selected an average of 51.9 variables across folds, for an average R^2 of 0.55. The forward stepwise using a p-value of 0.05 selects close to this number, 51. All of the estimators appear to be unbiased, with a mean residual varying between -0.0406 for ad hoc ELL, and 0.0024 for ad hoc OLS. These models do show a larger standard deviation of residuals than the Pakistan models, and further have noticeably smaller R^2 values, indicating less of the variation in consumption is captured by the generated models. This is somewhat expected given that limiting ourselves to the variables

available in the Sri Lankan dataset – roughly half of the number of variable we could use in Pakistan -- will result in a poorer fit ceteris paribus.

Turning to the error rates at the region level shown in table 6, and summarized in the bottom left panel of figure 2, we first look the performance at the 10% relative poverty rate (RPR). Manual OLS performs by far the worst, showing a mean weighted absolute error of 7.063, and a mean region error of -9.016, meaning manual OLS under-predicts the regional poverty rate at an average of 9 percentage points. Ad hoc ELL does much better, showing an MWAE of 2.574, followed by stepwise at 2.214. Post-Lasso ELL performs slightly better, showing an error rate of 1.989. When predicting the 20% RPR, ad hoc OLS performs very poorly, under-predicting the poverty rate by an average of 12.39. Ad hoc ELL does much better, with an average MWAE of 4.78, Post-Lasso ELL's MWAE improves to 3.381 and stepwise performs slightly better with an error rate of 3.078. For the 30% and 40% RPRs, the results show stepwise performing slightly better than Post-Lasso ELL, ad hoc OLS performing poorly, and ad hoc ELL performing increasingly worse as the relative poverty rate increases.

Unlike for Pakistan, error rates decrease monotonically as the relative poverty rate increases. Stepwise performs roughly as well as Lasso in this reduced-variable context, and manual performance decreases as the poverty threshold increases. It is difficult to explain why prediction accuracy generally increases with relative poverty rates in Pakistan but decreases with relative poverty rates in Sri Lanka. It appears that the relationship between prediction accuracy and relative poverty rates depends on the context and the data.

Other results are consistent across both countries: 1) The simulation approach of ELL (2003) results in much lower error in comparison to non-simulated methods. 2) Lasso greatly outperforms both manual and stepwise model selection when the set of variables is large. 3) Stepwise model selection performs approximately as well as Lasso when the set of variables is small. 4) Even using the best model, Post-Lasso ELL, region poverty rates have an error of around 2 percentage points.

4. Modeling Performance when Adding District-Level Spatial Variables

This section turns to evaluating how the four different modelling approaches fare when additional variables are added to set of candidate predictors. We augment the household models by including publicly available satellite data to assess the impact, if any, that such data can have on the accuracy of the modeling approaches. Satellite data has several potential advantages as a complement to nationally household survey data: it is cheap – publically available data of this kind is freely available as data products online; it is also ubiquitous and coverage includes most settled areas of the world; finally, it is frequently updated, with many data products being published at the yearly level. We chose several broad categories of satellite data, chosen on the basis of availability and likelihood of correlation in the relevant country. Some of these data have been used before in similar contexts, such as the use of night lights (Henderson, Storeyguard, and

Weil, 2012), which has been shown to be highly correlated with economic activity. Others, such as percent land cover of a given type or standard deviation of elevation, have seen less use in this context but may prove useful for the purposes of model building.

Taking the same household data for Pakistan and Sri Lanka as used in the previous section, we add indicators at the district level for the following:

- Land cover: classification of land into over 20 land cover types, including water bodies, built up urban area, irrigated and rain-fed cropland, vegetation of varying types and density, etc.
- Elevation: land elevation above sea level
- Population density: estimates population distribution per sq. km., based on multiple data sources including: census counts, land cover, roads, slope, urban areas, village locations, and high-resolution satellite imagery analysis
- Vegetation index: normalized difference vegetation index (NDVI), which gives a measure for how much live green vegetation is present in an area.
- GDP: an estimate of GDP per capita at the gridded spatial level, produced by the World Bank Development Economics Research Group by combining time-series data on GDP with Landscan's gridded population map.
- Night time lights: satellite imagery captured at night is widely used in analyzing economic activity and population distribution globally.
- Radiance calibrated night time lights: This product is an improvement over standard nightlights imagery as it captures more variation within very bright zones, such as cities, or very dim zones.

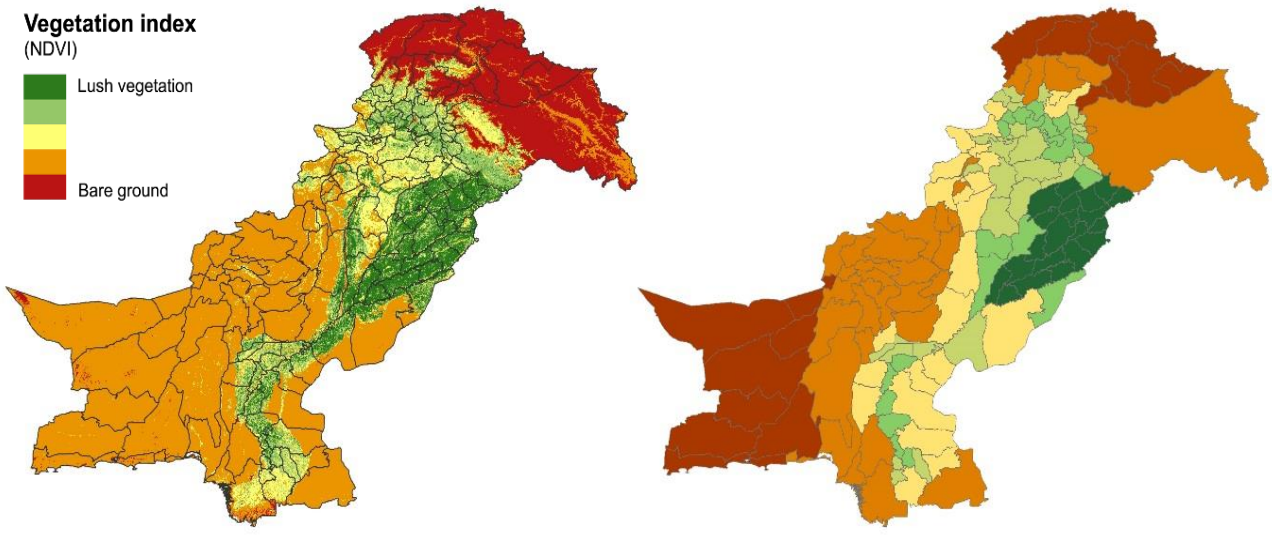


FIGURE 3: NDVI MAP FOR PAKISTAN AS A GRID (LEFT), AND AGGREGATED AT THE DISTRICT LEVEL (RIGHT)

This raw imagery is available publically online in the form of grids, or rasters, of varying spatial resolution.[23] We aggregated this data in the simplest form: using the mean and standard deviation for each district in Pakistan and the DS division level for Sri Lanka[24]. A sample spatial feature is shown in figure 4 above, which depicts the NDVI vegetation index, and shows the aggregation process. The left panel shows the raw raster image with the district boundaries of Pakistan displayed on top. We see a long band of lush vegetation starting at Hyderabad in the south, swirling around Lahore and Islamabad, and leading into Khyber Pakhtunkhwa. On the right panel we see the result of the aggregation, in this case average of NDVI by district. The averaging process gives the raw average of NDVI raster pixels contained in the district boundaries, resulting in the right panel. Many of the districts of Punjab have high degree of lush vegetation, and resultantly have a high district average in the right panel. Some intermediate districts, such as the Umer Kot in Sindh province, are only half covered in lush vegetation, which result in an intermediate average. This process is similar for all of the spatial variables.

We merge the spatial variables into the household data, at the level of the district in Pakistan and the DS division in Sri Lanka. Table 7 presents summary statistics for the spatial data used in Pakistan, and table 8 shows the summary statistics for the spatial data used in Sri Lanka. These summary statistics show mean over household level observations. The land cover variables enter the model as separate variables for each land cover type (there are 22 in total) giving the percent of total DS or district covered by this type of land. Most land-cover (31.5%) is bare land in Pakistan whereas the most common land-cover type is evergreen or semi-deciduous forest in Sri Lanka. For radiance calibrated night lights we include two time periods: 2010 and 1996, and for raw night lights we include the time periods 1992 and 2012. Notably absent is the vegetation index data for Sri Lanka, however with that exception the variables are similar across the two countries.

Many of the satellite-based variables are high collinear. The Lasso estimator is typically robust to the inclusion of highly correlated variables, whereas their inclusion can present some problems for stepwise (Dornmann, et al, 2013). Similarly, adding all of the available spatial variables to the ad hoc model resulted in unacceptably large variance inflation factor (VIF) for some covariates. We sequentially removed spatial variables with the largest VIF until all included spatial variables showed VIF scores below 10. This resulted in the exclusion of 4 land type variables for Pakistan and 7 land type variables for Sri Lanka.

Table 3 shows the model performance at the household level using the various model selection algorithms. For Pakistan, the ad hoc models estimate a model with 57 predictors, akin to adding every available coefficient that meets the VIF requirement. The stepwise algorithm selects 28 of

[23] Appendix A includes a more detailed description for how these variables were produced and where the source location for each of them is located. For the curious reader, we also present some raw raster maps for some of the data aggregated to the DS/District level in Appendix A.
[24] Description, data sources and methodology for each variable is described in greater detail in appendix A.

the available predictors, estimating a model with 133 covariates. Lasso adds an average of 13 of these available spatial covariates, excluding the majority of them. Both stepwise and Lasso show R^2 values of around 0.67 to 0.68, while the ad hoc R^2 values are both around 0.54. Turning to Sri Lanka, the ad hoc models add an additional 32 predictors to estimates a model with 53 total predictors. Stepwise selects a total of 38 predictors and Lasso selects an average of 73 across folds, an increase of 20 from the non-spatial set of controls.

Table 9 shows the region error rates for Pakistan between predicted and true relative poverty rates when we include spatial controls, which we also summarize in the upper right panel of figure 2. Comparing between the models with and without spatial data, the errors for stepwise and Lasso are similar to those in the section above that did not have access to spatial data. However, the manually selected ELL does considerably better, roughly as well as Post-Lasso ELL in two out of four cases, better than Lasso in predicting the below 20% poverty rate, and much worse than Post-Lasso ELL in predicting the below 40% relative poverty rate. Due to the richness of the data used to build the first set of models for Pakistan, including the satellite-based indicators made only marginal improvements for stepwise and Lasso. However, because the manual models included fewer predictors than either stepwise or Lasso, the satellite indicators increase their accuracy considerably. Comparing between the algorithmic approaches, stepwise performs quite poorly, especially in predicting the lowest 10% relative poverty rate, showing a MWAE of 8 percentage points. However, this performance improves as the RPR threshold increases, a similar pattern as to what was seen with the non-spatial examples.

Table 10 shows the region error rates for Sri Lanka comparing relative poverty rates based on actual and estimated per capita consumption, summarized in the lower right panel of figure 2. Here, almost without exception, the models with spatial data considerably outperform those when the algorithms did not have access to spatial data. The larger contribution of satellite-based indicators in Sri Lanka may result from the fewer variables employed in the Sri Lanka model. The higher resolution of the Sri Lankan satellite data may have also played a role, as Sri Lanka contains roughly 300 DS divisions and Pakistan data only contains 118 Districts. Including the satellite variables causes the relative error rates in the manual models to worsen by about 1-2 percentage points, while the Lasso models improves their error rates by between 1 and one half percentage points error, and the stepwise models see no discernable improvement. After adding the satellite-based variables, the Lasso-based model outperforms stepwise by a considerable margin. We conclude from this that Lasso was able to make use of the additional variables to generate more accurate predictions of poverty, while the other two methods were not.

Comparing the models with and without spatial data is summarized in Table 11. Performance varies across the poverty threshold. In Sri Lanka, the addition of spatial variables using the Post-Lasso ELL methodology improves performance by an average of 15% across poverty thresholds. In Pakistan the addition of publically available spatial variables tends to slightly lower average performance, which we attribute to the large number of predictors in the household data.

5. Conclusion

The results indicate that in the context of predicting poverty, model specification matters. For practitioners who predict poverty into ancillary surveys or censuses, we believe that model selection deserves the same type of rigorous attention that has been devoted to modelling the error term. We have shown, using Pakistan and Sri Lanka as two examples, that the Lasso estimator followed by the ELL simulation method to model errors can offer considerable improvements over the three other methods of model selection considered here: manual OLS, manual ELL, and forward stepwise. Lasso never performs substantially worse than these other methods. In cases where the set of predictor variables large, namely Pakistan and when the Sri Lankan data was augmented with publicly available satellite indicators, the gains to using a Lasso-based model selection process increased. In Sri Lanka, where there are fewer candidate predictors, performance does not appear different from existing methods, notably stepwise. Nonetheless, the results suggest that Lasso should be used more frequently as a model selection tool.

An ancillary question is to what extent adding publicly available satellite data improves the accuracy of the predicted estimates of poverty. For the purposes of this exercise, Pakistan is the atypical case, because of the availability of a larger and more representative survey with an exceptionally large number of common variables. Sri Lanka, where small area estimates of poverty would predict consumption into a census with fewer common variables, is a more canonical case. Though we must be cautious extrapolating from these two examples, our recommendation is that publically available spatial variables can improve prediction when the set of variables is scarce, such as is the case when estimating poverty into a census with limited covariates. The flip-side is a cautionary tale, in that publically available satellite data may worsen predictions if the set of covariates is already rich.

This research suggests several lines of future work. An important next step is to verify whether the main results documented here generalize to other contexts, particularly the weak dominance of the Lasso method of model selection, the monotonically improving performance of Lasso as the set of variables increases, and the improvement due to the inclusion of publicly available satellite data in Sri Lanka. The choice of household error was taken as given in our examples, but of course the optimal methodology depends on a combination of approaches to the selection of independent variables as well as error structure. Another unanswered question is why the shape of performance profile across relative poverty rates differs by country. In Pakistan, error rates decline as the poverty line rises and the models perform best when predicting membership in the bottom 40 percent of the national distribution. In contrast, in Sri Lanka error rates monotonically increase as the poverty line rises and the models perform best when distinguishing the bottom 10 percent of the national distribution. Further analysis could seek to better explain which pattern is more typical and the underlying factors behind this result.

References

Bajari, P., Nekipelov, D., Ryan, S. P., & Yang, M. (2015). Demand estimation with machine learning and model combination (No. w20955). National Bureau of Economic Research.

Baxter, M. and Hersh, J. (2015). Robust Determinants of Bilateral Trade Flows. Working paper.

Belloni, A., & Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2), 521-547.

Belloni, A., Chernozhukov, V., & Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2), 608-650.

Bien, J., Taylor, J., & Tibshirani, R. (2013). A lasso for hierarchical interactions. The Annals of Statistics, 41(3), 1111-1141.

Bonhomme, S., & Manresa, E. (2012). Grouped patterns of heterogeneity in panel data (No. wp2012_1208).

Demombynes, G., Elbers, C., Lanjouw, J. O., & Lanjouw, P. (2007). How good a map? Putting small area estimation to the test. *Putting Small Area Estimation to the Test (March 1, 2007). World Bank Policy Research Working Paper*, (4155).

Department of Census and Statistics and World Bank, forthcoming, "The Spatial Distribution of Poverty in Sri Lanka", mimeo

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., ... & Lautenbach, S. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36(1), 27-46.

Elbers, C., Lanjouw, J. O., & Lanjouw, P. (2003). Micro–Level Estimation of Poverty and Inequality. *Econometrica*, *71*(1), 355-364.

European Space Agency (ESA), Université catholique de Louvain (UCL). Globcover 2009 Land Cover Map Version 2.3. ESA, France, 2009.

Fernandez, C., Ley, E., & Steel, M. F. (2001). Model uncertainty in cross‐country growth regressions. *Journal of applied Econometrics*, 16(5), 563-576.

Gelman, A., Stevens, M., & Chan, V. (2003). Regression modeling and meta-analysis for decision making: a cost-benefit analysis of incentives in telephone surveys. *Journal of Business & Economic Statistics*, 21(2), 213-225.

Ghosh, M., & Rao, J. N. K. (1994). Small area estimation: an appraisal. *Statistical science*, 55-76.

Haslett, S., Jones, G., Noble, A., & Ballas, D. (2010). More for Less? Comparing small area estimation, spatial microsimulation, and mass imputation. *JSM*, 1584-1598.

Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Friedman, J., & Tibshirani, R. (2009). The elements of statistical learning (Vol. 2, No. 1). New York: springer.

Heckman, James J., John Eric Humphries, and Tim Kautz. "The economic and social benefits of GED certification." The Myth of Achievement Tests: The GED and the Role of Character in American Life (2014): 268-289.

Henderson, J. V., Storeygard, A., & Weil, D. N. (2009). Measuring economic growth from outer space. *American Economic Review* 102(2): 994-1028.

Krstajic, D., Buturovic, L. J., Leahy, D. E., & Thomas, S. (2014). Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of cheminformatics*, 6(1), 1-15.

Leamer, E. E. (1983). Let's take the con out of econometrics. *The American Economic Review*, 31-43.

Leamer, E. E. (1985). Sensitivity analyses would help. *The American Economic Review*, 308-313.

Levine, R., & Renelt, D. (1992). A sensitivity analysis of cross-country growth regressions. The American economic review, 942-963.

Molina, I., & Rao, J. N. K. (2010). Small area estimation of poverty indicators. *Canadian Journal of Statistics*, *38*(3), 369-385.

NASA, Japanese Ministry of Economy, Trade and Industry. ASTER Global DEM ASTGTM. USGS/Earth Resources Observation and Science (EROS) Center, Sioux Falls, South Dakota, 2011.

NASA Land Processes Distributed Active Archive Center (LP DAAC). MODIS 13Q1 Terra Vegetation Indices 16-Day L3 Global 250m SIN Grid V005. USGS Earth Resources Observation and Science (EROS) Center, Sioux Falls, South Dakota, 2014.

NDGC Earth Observation Group. Global Radiance Calibrated Nighttime Lights. NOAA NGDC, Boulder, Colorado, 2011.

Oak Ridge National Laboratory. LandScan High Resolution global Population Data Set. UT Batelle, Oak Ridge, Tennessee, 2012.

Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science*, *28*(1), 40-68.

Rao, J. N. (2005). *Small area estimation* (Vol. 331). John Wiley & Sons.

Tarozzi, A., & Deaton, A. (2009). Using census and survey data to estimate poverty and inequality for small areas. *The review of economics and statistics*,*91*(4), 773-792.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3), 273-282.

Van der Weide, R. (2014). GLS estimation and empirical bayes prediction for linear mixed models with Heteroskedasticity and sampling weights: a background study for the POVMAP project. *World Bank Policy Research Working Paper*, (7028).

Varian, H. R. (2014). Big data: New tricks for econometrics. *The Journal of Economic Perspectives*, 3-27.

World Bank Development Economics Research Group. Gross Domestic Product. Global Risk Data Platform, UNEP, Chatelaine, Geneva, 2010.

Table 1: Summary Table for Household Variables, Pakistan

Variable	N	mean	Variable	N	mean
HH highest Education	16,341	8.73	Rooms per person	16,341	0.37
Age of head	16,341	47.47	roofType==rcc/rbc	16,340	0.27
Age of spouse	2,497	40.75	roofType==wood/bamboo	16,340	0.36
Gender of head	16,341	1.07	roofType==steel/cement sheets	16,340	0.05
marstatHoH==Unmarried	16,341	0.02	roofType==other	16,340	0.32
marstatHoH==Married	16,341	0.92	wallType==burnt bricks/blocks	16,340	0.72
marstatHoH==Divorced/Separated	16,341	0.00	wallType==mud bricks/mud	16,340	0.21
marstatHoH==Widowed	16,341	0.07	wallType==wood/bamboo	16,340	0.02
Read/writes, head	16,341	0.56	wallType==stones	16,340	0.04
Read/writes, spouse	16,341	0.04	wallType==other	16,340	0.00
Head can do simple math	16,341	0.86	waterSource==piped water (inside	16,340	0.27
Spouse can do simple math	16,341	0.10	waterSource==out door tap	16,340	0.05
Head ever attended school	16,341	0.56	waterSource==hand pump	16,340	0.26
Spouse ever attended school	16,341	0.04	waterSource==motor pump	16,340	0.29
Max education, head	16,341	5.42	waterSource==closed well	16,340	0.01
Max education, spouse	16,341	0.45	waterSource==open well	16,340	0.04
Head ill or injured	16,341	0.10	waterSource== river/stream/pond/canal	16,340	0.04
Spouse ill or injured	16,341	0.01	waterSource==tanker/water barier	16,340	0.02
Number of HH members	16,341	7.74	waterSource==mineral water	16,340	0.00
Number of 65+ HH members	16,341	0.30	waterSource==other	16,340	0.02
Number of HH members 15-64	16,341	4.29	toiletType==no toilet	16,340	0.17
Number of HH members 0-5	16,341	1.23	toiletType==flush connected to severage	16,340	0.20
Dependency ratio of HH	16,341	107.13	toiletType==flush connected to tank	16,340	0.31
% of HH employed	16,341	0.28	toiletType==flush connected to open drain	16,340	0.17
No spouse present	16,341	0.13	toiletType==dry raised latrine	16,340	0.05
No spouse but children present	16,341	0.11	toiletType==pit latrine	16,340	0.08
HH owns land	16,341	0.30	toiletType==other	16,340	0.03
Amount of land owned	16,341	2.05	cookingFuel==wood	16,340	0.43
Amount of Agricultural Land	16,341	2.04	cookingFuel==gas	16,340	0.34
HH owns livestock	16,341	0.32	cookingFuel==carosine oil	16,340	0.00
HH owns sheep or goat	16,340	0.20	cookingFuel==dunk cakes	16,340	0.09
HH owns animals for transport	16,340	0.08	cookingFuel==electricity	16,340	0.00
HH owns chickens	16,340	0.13	cookingFuel==crop residue	16,340	0.13
Total value of HH assets	16,341	1897487	cookingFuel==coal/charcoal	16,340	0.00
HH owns electric iron	16,340	0.74	cookingFuel==other	16,340	0.01
HH owns electric fan	16,340	0.90	lightingFuel==electricity	16,340	0.92
HH owns sewing machine	16,340	0.58	lightingFuel==gas	16,340	0.01
HH owns radio	16,340	0.18	lightingFuel==carosine oil/diesel/petrol	16,340	0.06
HH owns chair	16,340	0.65	lightingFuel==wood	16,340	0.00

HH owns watch	16,340	0.79	lightingFuel==candle	16,340	0.00
HH owns television	16,340	0.57	lightingFuel==other	16,340	0.01
HH owns video player	16,340	0.05	phoneType==no	16,340	0.19
HH owns refrigerator	16,340	0.40	phoneType==land only	16,340	0.01
HH owns air cooler	16,340	0.07	phoneType==mobile	16,340	0.75
HH owns air conditioner	16,340	0.05	phoneType==both (land line and mobile)	16,340	0.05
HH owns computer	16,340	0.07	Time to water source	16,340	8.27
HH owns bicycle	16,340	0.30	Time to grocery	16,340	9.32
HH owns motorcycle	16,340	0.27	Time to public transit	16,340	11.40
HH owns car	16,340	0.04	Time to primary school	16,340	9.77
HH owns tractor	16,340	0.03	Time to middle school	16,340	16.17
HH owns mobile phone	16,340	0.80	Time to high school	16,340	19.06
HH owns cooking range	16,340	0.03	Time to clinic	16,340	18.72
HH owns burner	16,340	0.37	Time to family planning	16,340	20.93
HH owns washing machine	16,340	0.46	province==Punjab	16,341	0.57
famEconAssess==Much Worse	16,341	0.11	province==Sindh	16,341	0.24
famEconAssess==Slightly Worse	16,341	0.33	province==kpk	16,341	0.14
famEconAssess==Like before	16,341	0.41	province==Balochistan	16,341	0.05
famEconAssess==Little Better	16,341	0.13	urban==Rural	16,341	0.67
famEconAssess==Far better	16,341	0.02	HH in city high income area	16,341	0.01
famEconAssess==Dont Know	16,341	0.00	HH in city low income area	16,341	0.04
areaEconAssess==Much Worse	16,341	0.08	lang==Balochi	16,340	0.01
areaEconAssess==Slightly Worse	16,341	0.19	lang==Kashmiri	16,340	0.00
areaEconAssess==Like before	16,341	0.62	lang==Other	16,340	0.11
areaEconAssess==Little Better	16,341	0.06	lang==Pashtu	16,340	0.11
areaEconAssess==Far better	16,341	0.01	lang==Punjabi	16,340	0.36
areaEconAssess==Dont Know	16,341	0.04	lang==Sindhi	16,340	0.15
residenceType==owner occupied (self hired)	16,340	0.04	lang==Urdu	16,340	0.27
residenceType==owner occupied (not self	16,340	0.82			
residenceType==on rent	16,340	0.06			
residenceType==subsidized rent	16,340	0.01			
residenceType==rent free	16,340	0.06			

Table 2: Household Level Models Summary, Baseline Controls

	Spatial Controls?	Avg. # of variables	Avg. R^2	Mean Resid	Std Resid	Min Resid	Max Resid
Pakistan Models							
Ad hoc OLS	No	20.00	0.53	0.0200	0.3532	-6.9410	2.6080
Ad hoc ELL	No	20.00	0.51	-0.0158	0.3535	-4.0106	2.6305
Post-Lasso ELL	No	62.20	0.68	0.0299	0.2999	-3.5294	2.5554
Stepwise ELL	No	105.00	0.67	0.0142	0.2958	-3.0450	2.5287
Sri Lankan Models							
Ad hoc OLS	No	21.00	0.42	0.0024	0.4962	-1.8477	3.4233
Ad hoc ELL	No	21.00	0.42	-0.0406	0.4985	-1.9114	3.3672
Post-Lasso ELL	No	51.90	0.55	0.0021	0.4406	-1.8080	3.2173
Stepwise ELL	No	51.00	0.55	0.0010	0.4404	-1.8976	3.2123

Table 3: Household Level Models Compared, Spatial Controls

	Spatial Controls?	Avg # of variables	Avg. R^2	Mean Resid	Std Resid	Min Resid	Max Resid
Pakistan							
Ad Hoc OLS	Yes	57	0.53852	0.01721	0.35053	-7.136	2.62426
Ad Hoc ELL	Yes	47	0.54025	-0.0095	0.34784	-4.2278	2.73378
Post-Lasso ELL	Yes	75.1	0.68666	0.02881	0.29885	-3.773	2.56711
Stepwise ELL	Yes	133	0.67281	0.01441	0.29468	-3.1468	2.54317
Sri Lanka							
Ad Hoc OLS	Yes	53	0.41911	0.00683	0.49714	-1.9976	3.49481
Ad Hoc ELL	Yes	52	0.42394	-0.0067	0.49901	-2.0384	3.48125
Post-Lasso ELL	Yes	73	0.55847	0.00738	0.43533	-1.7865	3.22795
Stepwise ELL	Yes	38	0.41811	0.01397	0.49995	-2.0107	3.5432

Table 4: Region Error Rates Between Predicted and Relative Poverty Rate, Pakistan

	Spatial Controls?	Mean Region Absolute Error	Mean Weighted Absolute Error	Mean Region Error
Panel A: Bottom 10% of Consumption				
Ad Hoc OLS	No	5.977	6.193	-5.977
Ad Hoc ELL	No	4.408	4.408	5.636
Post-Lasso ELL	No	4.272	3.391	4.272
Stepwise ELL	No	9.074	8.347	9.074
Panel B: Bottom 20% of Consumption				
Ad Hoc OLS	No	7.358	7.943	-7.358
Ad Hoc ELL	No	3.471	3.471	3.061
Post-Lasso ELL	No	2.804	2.108	2.578
Stepwise ELL	No	7.137	6.611	7.137
Panel C: Bottom 30% of Consumption				
Ad Hoc OLS	No	6.296	7.348	-6.296
Ad Hoc ELL	No	3.551	3.551	-0.408
Post-Lasso ELL	No	1.320	1.120	0.284
Stepwise ELL	No	3.786	3.928	3.786
Panel D: Bottom 40% of Consumption				
Ad Hoc OLS	No	4.348	5.145	-3.929
Ad Hoc ELL	No	4.312	4.312	-3.762
Post-Lasso ELL	No	2.050	2.118	-2.018
Stepwise ELL	No	1.528	1.470	0.237

"Relative poverty" is poverty defined as a household's consumption below #% of national consumption. Each model attempts to estimate this constructed poverty rate at the household level, with results aggregated to region. Region refers to sampling frame of survey, which is at the urban/rural district level. Mean region error refers to the average error across regions, absolute error takes the absolute difference between constructed and estimated pseudo poverty rates. Weighted absolute error adjusts for population differences between regions when calculating mean error rates and weights accordingly.

Table 5: Summary Table for Household Variables, Sri Lanka

Variable	N	mean	Variable	N	mean
HH located in Urban area	20,540	0.17	HH owns electric fan	20,540	0.57
HH located in Rural area	20,540	0.79	HH owns telephone	20,540	0.37
Head is unemployed	20,540	0.01	HH owns mobile	20,540	0.81
Head is a government employee	20,540	0.10	HH owns computer	20,540	0.19
Head is privately employed	20,540	0.61	HH owns camera	20,540	0.11
HH is Hindu	20,540	0.12	HH owns bicycle	20,540	0.36
HH is Islam	20,540	0.09	HH owns motorbike	20,540	0.30
HH is Christian	20,540	0.08	HH owsn three wheeler	20,540	0.11
HH is of Other religion	20,540	0.00	HH owns van	20,540	0.07
Age, head	20,540	51.26	HH owns bus	20,540	0.02
Age Squared, head	20,540	2822.06	HH owns tractor	20,540	0.04
Head is is male	20,540	0.77	HH owns pesticider	20,540	0.03
Married, head	20,540	0.79	HH owns thresher	20,540	0.00
Widowed, head	20,540	0.16	HH owns waterpump	20,540	0.02
Education leve of head	20,540	8.14	HH owns boat	20,540	0.01
Education Squared of head	20,540	79.59	HH owns fishing net	20,540	0.01
Household size	20,540	3.88	Num bedrooms	20,540	2.38
Household size squared	20,540	17.61	HH experienced Natural calamity	20,540	0.91
Highest education in HH	20,540	12.58	HH owns toilet	20,540	0.90
Num males in HH 0-4	20,540	0.17	House owned	20,540	0.87
Num males in HH 5-9	20,540	0.18	Wall type brick	20,540	0.53
Num males in HH 10-14	20,540	0.17	Wall type cement	20,540	0.33
Num males in HH 65+	20,540	0.15	Wall type mud	20,540	0.04
Num males in HH 15-64	20,540	1.17	Roof type tile	20,540	0.48
Num females in HH 0-4	20,540	0.16	Roof type asbestos	20,540	0.36
Num females in HH 5-9	20,540	0.17	Roof type concrete	20,540	0.04
Num females in HH 10-14	20,540	0.16	Roof type wood	20,540	0.01
Num females in HH 15-64	20,540	1.37	Roof type sand	20,540	0.09
HH owns radio	20,540	0.71	Floor type cement	20,540	0.73
HH owns TV	20,540	0.83	Floor type tile	20,540	0.13
HH owns Video player	20,540	0.43	Safe drinking water	20,540	0.89
HH owns sewing machine	20,540	0.42	Firewood for cooking	20,540	0.78
HH owns washing machine	20,540	0.17	Gas for cooking	20,540	0.18
HH owns fridge	20,540	0.46	Electricity for cooking	20,540	0.00
HH owns cookers	20,540	0.43	Electrical grid lighting	20,540	0.09

Table 6: Region Error Rates Between Predicted and Relative Poverty Rate, Sri Lanka

Model	Spatial Controls?	Mean Region Absolute Error	Mean weighted Absolute Error	Mean Region Error
Panel A: Bottom 10% of Consumption				
Ad Hoc OLS	No	9.016	7.063	-9.016
Ad Hoc ELL	No	4.043	2.574	-0.813
Post-Lasso ELL	No	3.606	1.989	-0.115
Stepwise ELL	No	3.738	2.214	0.642
Panel B: Bottom 20% of Consumption				
Ad Hoc OLS	No	12.390	11.015	-12.390
Ad Hoc ELL	No	6.240	4.781	-4.387
Post-Lasso ELL	No	4.943	3.381	-2.290
Stepwise ELL	No	4.661	3.078	-1.222
Panel C: Bottom 30% of Consumption				
Ad Hoc OLS	No	12.640	12.390	-12.445
Ad Hoc ELL	No	9.203	7.079	-7.816
Post-Lasso ELL	No	6.580	4.793	-4.642
Stepwise ELL	No	5.854	4.201	-3.402
Panel D: Bottom 40% of Consumption				
Ad Hoc OLS	No	10.442	10.368	-9.238
Ad Hoc ELL	No	10.592	8.433	-9.723
Post-Lasso ELL	No	7.171	5.508	-5.739
Stepwise ELL	No	6.348	4.865	-4.432

"Relative poverty" is poverty defined as a household's consumption below #% of national consumption. Each model attempts to estimate this constructed poverty rate at the household level, with results aggregated to region. Region refers to sampling frame of survey, which is at the urban/rural district level. Mean region error refers to the average error across regions, absolute error takes the absolute difference between constructed and estimated pseudo poverty rates. Weighted absolute error adjusts for population differences between regions when calculating mean error rates and weights accordingly.

Table 7: Summary Statistics for Spatial Variables, Pakistan

Variable	Mean	Standard deviation
Land Cover (percentage area per district)		
Closed to open (>15%) mixed broadleaved and needleleaved forest (>5m)	0.6	1.6
Post-flooding or irrigated croplands	30.8	34.4
Mosaic Forest/Shrubland (50-70%) / Grassland (20-50%)	0.1	0.3
Mosaic Grassland (50-70%) / Forest/Shrubland (20-50%)	2.0	5.2
Closed to open (>15%) shrubland (<5m)	0.5	1.3
Rainfed croplands	8.8	14.1
Closed to open (>15%) grassland	6.3	12.0
Sparse (>15%) vegetation (woody vegetation, shrubs, grassland)	0.8	2.3
Closed (>40%) broadleaved forest regularly flooded - Fresh water	0.0	0.0
Closed (>40%) broadleaved semi-deciduous and/or evergreen forest regularly	0.0	0.2
Closed to open (>15%) vegetation (grassland, shrubland, woody vegetation) on	0.0	0.0
Artificial surfaces and associated areas (urban areas >50%)	0.8	3.1
Mosaic Cropland (50-70%) / Vegetation (grassland, shrubland, forest) (20-50%)	7.4	7.6
Bare areas	31.5	31.4
Water bodies	0.4	1.0
Permanent snow and ice	1.9	6.3
Mosaic Vegetation (grassland, shrubland, forest) (50-70%) / Cropland (20-50%)	6.8	5.8
Closed to open (>15%) broadleaved evergreen and/or semi-deciduous forest (>5m)	0.1	0.3
Closed (>40%) broadleaved deciduous forest (>5m)	0.1	0.3
Open (15-40%) broadleaved deciduous forest (>5m)	0.0	0.0
Closed (>40%) needleleaved evergreen forest (>5m)	1.2	3.7
Open (15-40%) needleleaved deciduous or evergreen forest (>5m)	0.0	0.0
Elevation (m)	926.81	1078.27
GDP	200.94	420.16
Population density (Landscan 2012)	308.25	465.73
Normalized Differential Vegetation Index	2987.95	1872.68
Radiance-calibrated nightlights (2010)	8.84	11.92

Table 8: Summary Statistics for Spatial Variables, Sri Lanka

Variable	Mean	S.d.
Elevation. mean	195.14	327.77
Elevation, std	50.47	90.87
GDP values from UNEP/DEC, mean	3592.20	12916.63
GDP values from UNEP/DEC, std	1386.76	1989.02
Land type: Artificial surfaces and associated areas (urban areas >50%)	9.1%	0.23
Land type: Bare areas	0.1%	0.00
Land type: Closed (>40%) broadleaved deciduous forest (>5m)	0.7%	0.02
Land type: Closed (>40%) broadleaved forest regularly flooded - Fresh water	0.0%	0.00
Land type: Closed (>40%) broadleaved semi-deciduous and/or evergreen forest reg	0.0%	0.00
Land type: Closed (>40%) needleleaved evergreen forest (>5m)	0.6%	0.01
Land type: Closed to open (>15%) broadleaved evergreen and/or semi-deciduous fo	58.7%	0.29
Land type: Closed to open (>15%) grassland	0.3%	0.01
Land type: Closed to open (>15%) mixed broadleaved and needleleaved forest (>5m	4.8%	0.06
Land type: Closed to open (>15%) shrubland (<5m)	15.4%	0.15
Land type: Closed to open (>15%) vegetation (grassland, shrubland, woody vegeta	0.0%	0.00
Land type: Mosaic Cropland (50-70%) / Vegetation (grassland, shrubland, forest)	1.2%	0.03
Land type: Mosaic Forest/Shrubland (50-70%) / Grassland (20-50%)	0.0%	0.00
Land type: Mosaic Grassland (50-70%) / Forest/Shrubland (20-50%)	0.0%	0.00
Land type: Mosaic Vegetation (grassland, shrubland, forest) (50-70%) / Cropland	1.5%	0.03
Land type: Open (15-40%) broadleaved deciduous forest (>5m)	0.0%	0.00
Land type: Open (15-40%) needleleaved deciduous or evergreen forest (>5m)	0.0%	0.00
Land type: Permanent snow and ice	0.0%	0.00
Land type: Post-flooding or irrigated croplands	3.1%	0.10
Land type: Rainfed croplands	2.9%	0.07
Land type: Sparse (>15%) vegetation (woody vegetation, shrubs, grassland)	0.0%	0.00
Land type: Water bodies	1.7%	0.03
Radiance calibrated night lights 1996, mean	16.59	30.93
Radiance calibrated night lights 1996, std	5.54	6.75
Radiance calibrated night lights 2010, mean	22.73	31.28
Radiance calibrated night lights 2010, std	5.85	6.72
Raw night lights 1992, std	3.83	3.35
Raw night lights 2012, mean	15.40	15.55
Raw night lights 2012, std	3.72	3.45
Raw night lights 1992, mean	10.33	14.89

Table 9: Region Error Rates Between Predicted and Relative Poverty Rate, Spatial Controls, Pakistan

Model	Spatial Controls?	Mean Region Absolute Error	Mean Weighted Absolute Error	Mean Region Error
Panel A: Bottom 10% of Consumption				
Ad Hoc OLS	Yes	5.462	5.768	-5.462
Ad Hoc ELL	Yes	4.396	4.396	5.669
Post-Lasso ELL	Yes	5.309	4.485	5.309
Stepwise ELL	Yes	9.182	8.355	9.182
Panel B: Bottom 20% of Consumption				
Ad Hoc OLS	Yes	7.216	7.838	-7.216
Ad Hoc ELL	Yes	2.821	2.821	3.225
Post-Lasso ELL	Yes	4.160	3.335	4.160
Stepwise ELL	Yes	7.233	6.626	7.233
Panel C: Bottom 30% of Consumption				
Ad Hoc OLS	Yes	7.083	7.702	-7.083
Ad Hoc ELL	Yes	1.997	1.997	-0.140
Post-Lasso ELL	Yes	2.276	2.089	2.102
Stepwise ELL	Yes	3.918	3.946	3.918
Panel D: Bottom 40% of Consumption				
Ad Hoc OLS	Yes	4.484	5.469	-3.985
Ad Hoc ELL	Yes	4.150	4.150	-3.422
Post-Lasso ELL	Yes	0.870	0.886	-0.070
Stepwise ELL	Yes	1.774	1.569	0.295

"Relative poverty" is poverty defined as a household's consumption below #% of national consumption. Each model attempts to estimate this constructed poverty rate at the household level, with results aggregated to region. Region refers to sampling frame of survey, which is at the urban/rural district level. Mean region error refers to the average error across regions, absolute error takes the absolute difference between constructed and estimated pseudo poverty rates. Weighted absolute error adjusts for population differences between regions when calculating mean error rates and weights accordingly. Spatial controls include district level average and standard deviation measures for night lights, radiance corrected night lights, NDVI (vegetation index), and % land cover of a given land type.

Table 10: Region Error Rates Between Predicted and Relative Poverty Rate, Spatial Controls, Sri Lanka

Model	Spatial Controls?	Mean Region Absolute Error	Mean weighted Absolute Error	Mean Region Error
Panel A: Bottom 10% of Consumption				
Ad Hoc OLS	Yes	8.454	6.914	-8.454
Ad Hoc ELL	Yes	3.604	2.230	0.644
Post-Lasso ELL	Yes	3.565	2.058	0.511
Stepwise ELL	Yes	3.956	2.563	1.571
Panel B: Bottom 20% of Consumption				
Ad Hoc OLS	Yes	11.240	11.167	-11.172
Ad Hoc ELL	Yes	4.630	3.373	-2.307
Post-Lasso ELL	Yes	4.415	2.812	-1.340
Stepwise ELL	Yes	4.602	3.076	-1.054
Panel C: Bottom 30% of Consumption				
Ad Hoc OLS	Yes	11.648	13.020	-11.382
Ad Hoc ELL	Yes	6.369	4.942	-5.271
Post-Lasso ELL	Yes	5.181	3.604	-3.501
Stepwise ELL	Yes	5.520	4.085	-3.864
Panel D: Bottom 40% of Consumption				
Ad Hoc OLS	Yes	9.251	10.704	-8.021
Ad Hoc ELL	Yes	7.394	6.052	-6.846
Post-Lasso ELL	Yes	5.653	4.222	-4.434
Stepwise ELL	Yes	6.528	5.027	-5.332

"Relative poverty" is poverty defined as a household's consumption below #% of national consumption. Each model attempts to estimate this constructed poverty rate at the household level, with results aggregated to region. Region refers to sampling frame of survey, which is at the urban/rural district level. Mean region error refers to the average error across regions, absolute error takes the absolute difference between constructed and estimated pseudo poverty rates. Weighted absolute error adjusts for population differences between regions when calculating mean error rates and weights accordingly. Spatial controls include district level average and standard deviation measures for night lights, radiance corrected night lights, NDVI (vegetation index), and % land cover of a given land type.

Table 11: Performance Comparison with and without Spatial Controls

	Sri Lanka			Pakistan		
	Mean Weighted Abs Error			Mean Weighted Abs Error		
	w/o Spatial	w/ Spatial	% imp	w/o Spatial	w/ Spatial	% imp
Panel A: Bottom 10% of Consumption						
Ad Hoc OLS	7.063	6.914	2.11%	6.193	5.768	6.86%
Ad Hoc ELL	2.574	2.23	13.36%	4.408	4.396	0.27%
Post-Lasso ELL	1.989	2.058	-3.47%	3.391	4.485	-32.26%
Stepwise ELL	2.214	2.563	-15.76%	8.347	8.355	-0.10%
Panel B: Bottom 20% of Consumption						
Ad Hoc OLS	11.015	11.167	-1.38%	7.943	7.838	1.32%
Ad Hoc ELL	4.781	3.373	29.45%	3.471	2.821	18.73%
Post-Lasso ELL	3.381	2.812	16.83%	2.108	3.335	-58.21%
Stepwise ELL	3.078	3.076	0.06%	6.611	6.626	-0.23%
Panel C: Bottom 30% of Consumption						
Ad Hoc OLS	12.39	13.02	-5.08%	7.348	7.702	-4.82%
Ad Hoc ELL	7.079	4.942	30.19%	3.551	1.997	43.76%
Post-Lasso ELL	4.793	3.604	24.81%	1.12	2.089	-86.52%
Stepwise ELL	4.201	4.085	2.76%	3.928	3.946	-0.46%
Panel D: Bottom 40% of Consumption						
Ad Hoc OLS	10.368	10.704	-3.24%	5.145	5.469	-6.30%
Ad Hoc ELL	8.433	6.052	28.23%	4.312	4.15	3.76%
Post-Lasso ELL	5.508	4.222	23.35%	2.118	0.886	58.17%
Stepwise ELL	4.865	5.027	-3.33%	1.47	1.569	-6.73%
Panel E: Average Across Relative Poverty Rates						
Ad Hoc OLS			-1.90%			-0.73%
Ad Hoc ELL			25.31%			16.63%
Post-Lasso ELL			15.38%			-29.70%
Stepwise ELL			-4.07%			-1.88%

Appendix A: Spatial variables

NDVI

Normalized Difference Vegetation Index is a widely used indicator to quantify greenery of a region and has applications in understanding the health of vegetation and characterizing land cover, amongst other uses. This product is generated by Earth Resources Observation and Science Center of the US Geological Survey using imagery from NASA's MODIS satellite. The calculation of NDVI is based on the variable ways in which different spectral bands of light are reflected by plants[25]. Healthy plants absorb large quantities of visible light for photosynthesis, whereas near-infrared light is barely absorbed and mostly reflected back. NDVI exploits this difference and is calculated by the following formula: NDVI = (NIR — VIS)/(NIR + VIS). This generates a value between -1 and 1, with higher values indicating more greenery.

Data access: http://reverb.echo.nasa.gov/

Year: 2014

Generating district-level aggregates: District level aggregates were created in ArcGIS using the Zonal Statistics tool of the Spatial Analyst toolbar[26]. For each district, the following statistics are generated:

- MEAN — Calculates the average of all cells in the value raster that belong to the same zone as the output cell.
- MAJORITY — Determines the value that occurs most often of all cells in the value raster that belong to the same zone as the output cell.
- MAXIMUM — Determines the largest value of all cells in the value raster that belong to the same zone as the output cell.
- MEDIAN — Determines the median value of all cells in the value raster that belong to the same zone as the output cell.
- MINIMUM — Determines the smallest value of all cells in the value raster that belong to the same zone as the output cell.
- MINORITY — Determines the value that occurs least often of all cells in the value raster that belong to the same zone as the output cell.
- RANGE — Calculates the difference between the largest and smallest value of all cells in the value raster that belong to the same zone as the output cell.
- STD — Calculates the standard deviation of all cells in the value raster that belong to the same zone as the output cell.
- SUM — Calculates the total value of all cells in the value raster that belong to the same zone as the output cell.
- VARIETY — Calculates the number of unique values for all cells in the value raster that belong to the same zone as the output cell.

[25] http://earthobservatory.nasa.gov/Features/MeasuringVegetation/measuring_vegetation_1.php
[26] http://resources.arcgis.com/en/help

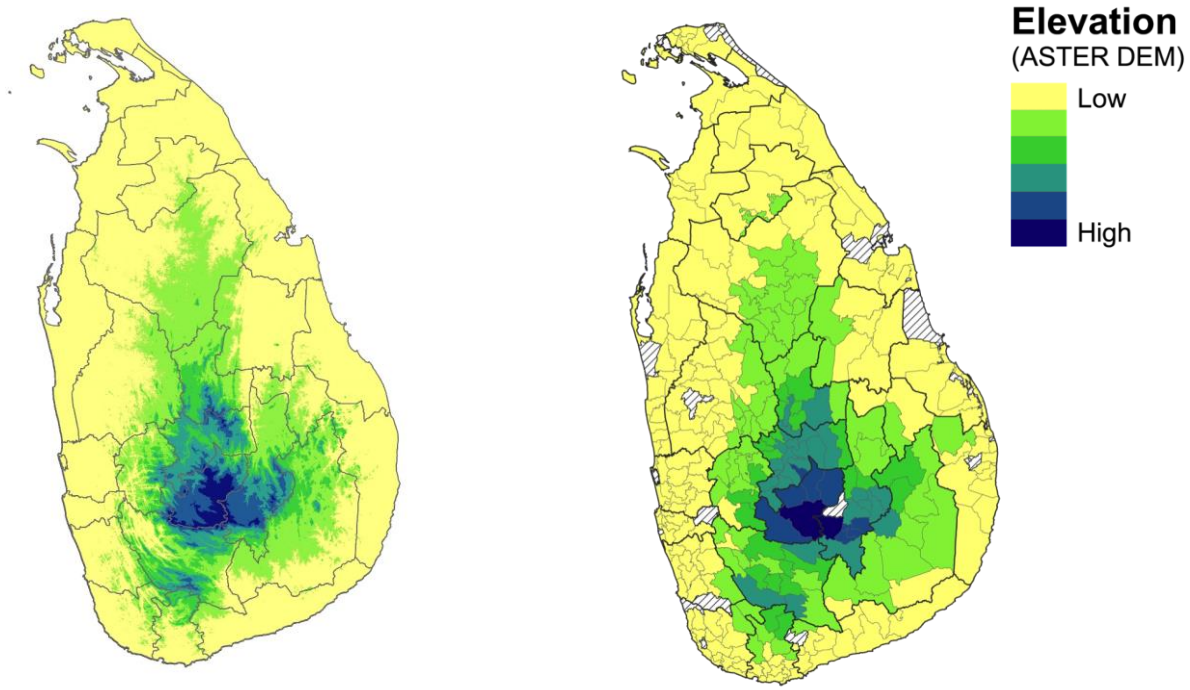
Elevation

Global Digital Elevation Model is a comprehensive elevation map produced jointly by NASA and the Japanese Ministry of Economy, Trade and Industry (METI). It is derived from imagery from the Japanese sensor ASTER (Advanced Spaceborne Thermal Emission and Reflection Radiometer) that is aboard NASA's Terra satellite. The methodology relies on correlating stereo image pairs from two angles and analyzing the variation to estimate elevation[27]. These individual DEMs are stacked with multiple DEMs covering the same scenes, and are combined to reduce bad values (e.g., occluded by clouds) and merged to create the final global DEM layer. Values of the Global DEM layers range from -500 to 9000 m, with zero representing sea level. The layer is generated at a resolution of 1 arc second, which roughly equates to 30 m at the equator. The latest Global DEM product was released in 2011.

Data access: http://gdem.ersdac.jspacesystems.or.jp/

Year: 2011

Generating district-level aggregates: District level aggregates were created in ArcGIS using the Zonal Statistics tool of the Spatial Analyst toolbar (see above).



(A) (B)

FIGURE 4: ASTER DIGITAL ELEVATION MAP FOR SRI LANKA AS (A) A GRID, AND (B) AGGREGATED AT THE DS LEVEL

[27] https://lpdaac.usgs.gov/sites/default/files/public/aster/docs/Tachikawa_etal_IGARSS_2011.pdf

Landscan

Landscan is a widely used global population distribution product generated by Oak Ridge National Laboratory. The methodology models population distribution by incorporating multiple data sources including: census counts, land cover, roads, slope, urban areas, village locations, and high-resolution satellite imagery analysis. The final product has a spatial resolution of 30 arc seconds, which is equivalent to approximately 1 km at the equator. Each pixel represents the predicted number of people per 30 arc seconds.

Data access: http://web.ornl.gov/sci/landscan/

Year: 2012

Generating district-level aggregates: District level aggregates were created in ArcGIS using the Zonal Statistics tool of the Spatial Analyst toolbar (see above).

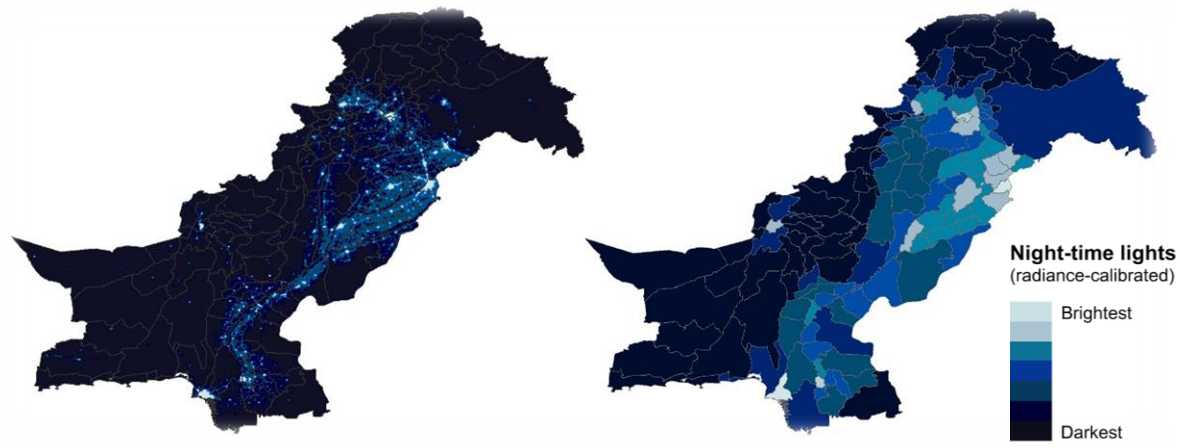
Nightlights

Global lights at night products are produced by NASA, and are widely used in analyzing economic activity and population distribution globally. The product is generated with imagery captured by the VIIRS sensors aboard NASA's Suomi NPP satellite. Nightlights In this case, we used radiance calibrated nightlights product which is an improvement over standard nightlights imagery as it captures more variation within very bright zones, such as cities, or very dim zones. It does so by capturing imagery at varying sensor sensitivity levels and merging them to create a richer dataset. The final product has a spatial resolution of roughly 750 m at the equator, with the latest version being released in 2011.

Data access: http://ngdc.noaa.gov/eog/dmsp/download_radcal.html

Year: 2011

Generating district-level aggregates: District level aggregates were created in ArcGIS using the Zonal Statistics tool of the Spatial Analyst toolbar.



(A) (B)

FIGURE 5: NIGHTLIGHTS MAP FOR PAKISTAN AS (A) A GRID, AND (B) AGGREGATED AT THE DISTRICT LEVEL

GDP

In the distributed global GDP dataset sub-national GRP and national GDP data are allocated to 30 arc second (approximately 1km) grid cells in proportion to the population residing in that cell. The method also distinguishes between rural and urban population, assuming the latter to have a higher GDP per capita. Input data are from: a global time-series dataset of GDP, with subnational gross regional product (GRP) for 74 countries, compiled by the World Bank Development Economics Research Group (DECRG). Gridded population projections for the year 2009, based on a population grid for the year 2005 provided by LandScan Global Population Database (Oak Ridge, TN: Oak Ridge National Laboratory). This dataset has been extrapolated to year 2010 by UNEP/GRID-Geneva. Unit is estimated value of production per cell, in thousand of constant 2000 USD. This product was compiled by DECRG for the Global Assessment Report on Risk Reduction (GAR)[28].

Data access: http://preview.grid.unep.ch/index.php?preview=data&events=socec&evcat=1

Year: 2010

Generating district-level aggregates: District level aggregates were created in ArcGIS using the Zonal Statistics tool of the Spatial Analyst toolbar (see above).

Global Landcover

The global landcover product by the European Space Agency and Université catholique de Louvain classifies land into over 20 land cover types, including water bodies, built up urban area, irrigated and rain-fed cropland, vegetation of varying types and density, etc. This is generated from the MODIS surface spectral reflectance to capture variation in surfaces on the ground at a spatial resolution of 300m.

Data access: http://due.esrin.esa.int/globcover/

Year: 2009

Generating district-level aggregates: District level aggregates were defined as the percentage of area of each district that was covered by each land-cover category. This was created in ArcGIS by iterating over each category and applying the Zonal Statistics tool to find district-level area per category. This was further processed to find percentage area per category within each district.

[28] Description from: http://preview.grid.unep.ch/index.php?preview=data&events=socec&evcat=1