How Accurate Is a Poverty Map Based on Remote Sensing Data? An Application to Malawi*

Roy van der Weide[†], Brian Blankespoor[‡], Chris Elbers[§], and Peter Lanjouw[¶]

March 2023

Abstract

This paper assesses the reliability of poverty maps derived from remote-sensing data. Employing data for Malawi, it first obtains small area estimates of poverty by combining household expenditure survey data with population census data. It then ignores the population census and obtains a second poverty map by combining the survey with predictors of poverty derived from remote sensing data. The two approaches reveal the same patterns in the geography of poverty. However, there are instances where the two approaches obtain markedly different estimates of poverty. Poverty maps obtained using remote sensing data may do well when the decision maker is interested in comparisons of poverty between assemblies of areas yet may be less reliable when the focus is on estimates for specific small areas.

Keywords: Poverty, Small Area Estimation, Remote Sensing Data
JEL Classification: C13, D31, I32, O15, R13, Y91

*The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent. The authors gratefully acknowledge financial support from the World Bank Knowledge for Change Program.

[†]World Bank: rvanderweide@worldbank.org
[‡]World Bank: bblankespoor@worldbank.org
[§]Vrije Universiteit Amsterdam: c.t.m.elbers@vu.nl
[¶]Vrije Universiteit Amsterdam: p.f.lanjouw@vu.nl

1 Introduction

Tracking poverty at the global, regional, and national levels helps international organizations and the development community identify lagging regions and allocate their resources accordingly (see e.g. Ravallion (2018), Ferreira et al. (2015), Bourguignon (2017)). Poverty is commonly measured by the percentage of the population whose income or consumption expenditure falls below a poverty line. The household income and expenditure data used to estimate poverty come from household surveys that are conducted every 1-5 years by the national statistical agencies. Individual countries use these data to monitor poverty at the national and subnational levels.

Elbers et al. (2007) demonstrates the value of estimating poverty at a more disaggregated level. Specifically, they estimate by how much governments could lower the costs of providing resources to the poor if they had access to a poverty map that provides estimates of poverty down to the district or municipality level. Their simulation study using data from Madagascar, Cambodia, and Ecuador suggests that the gains can be substantial. In all three countries, the reduction in poverty that is achieved with a uniform transfer to all households can often be realized with less than half the budget if the transfers could be targeted on the basis of municipality level estimates of poverty. Unfortunately, collecting data on household income and expenditure is costly and time consuming (Kilic et al. 2017; Fujii and van der Weide, 2020). As a result, survey samples tend to be relatively small, permitting only a limited degree of disaggregation, often not beyond large sub-regions.

Estimating poverty at the small area level requires alternative sources of data, ideally a population census, that are able to provide complete coverage of the population. Elbers, Lanjouw, and Lanjouw (2003), hereafter referred to as ELL, pioneered the small area estimation of poverty by combining household survey with population census data. The proposed approach imputes household income or consumption expenditure (multiple times) for each household into the population census, and then aggregates the multiple imputed values to obtain estimates of poverty at the municipality or district level. The population census arguably provides the richest possible data source for this purpose. It contains a comprehensive set of variables that are highly correlated with household income or consumption (although not, usually, income itself). Specifically, a census typically includes information on demographics, education, employment, housing characteristics, asset ownership, and community services (i.e. local amenities) for all households in the population. In the ELL procedure, a household survey is used to infer the relationship between these variables and household income, taking advantage of the fact that many of these variables are available in both the census and household survey.

Elbers et al. (2008) assess the validity of the approach using data from Brazil and find that small-area estimates based on the ELL procedure are close to direct calculations of sub-national poverty in this setting. ELL has been adopted to obtain small area estimates of poverty and inequality in over 60 countries worldwide.[1] Poverty maps are also increasingly used as a source of data in a variety of empirical applications, see e.g. Araujo et al. (2008), Baird et al. (2013), Bazzi (2017), Crost et al. (2014), Demombynes and Ozler (2005), Elbers et al. (2005), and Maloney and Caicedo (2015).

An important limitation of the ELL approach, however, is precisely its reliance on the population census. For all the strengths of a population census, a notable disadvantage is its limited availability. Population censuses are generally conducted only once every 10 years, and it often takes two years or longer to process the data. These factors combine to imply that ELL-based poverty maps are rarely up-to-date (i.e. are often already outdated at the time of release). Furthermore, even when current population census data exist and are ready for analysis, getting access to such data is often not trivial. As a result, areas that may be most in need of a timely poverty map may lack the data necessary to implement the ELL approach. An alternative approach that can produce a poverty map in the absence of population census data will thus have obvious appeal.

The objective of this study is to evaluate whether an accurate map of poverty can be obtained using predictors derived from publicly available remote sensing data instead of a population census. A requirement in such a case would be that the household expenditure survey is also geo-referenced, i.e. that the enumeration areas (or primary sampling units) can be located on a map. While this would have been a demanding requirement a decade ago, geo-referencing of household surveys is now increasingly becoming standard practice. To preserve the anonymity of sampled households, the geographic coordinates of the primary sampling units in public use datasets are not exact, i.e. some random offsets are applied. Our secondary objective is to investigate whether these random offsets undermine this alternative approach of obtaining a poverty map. The primary sampling unit, such as a village or a neighborhood in a city, will be the unit of observation (rather than the household). Correspondingly, the level of poverty in the primary sampling units will now serve as the dependent variable. Examples of possible predictors of local poverty rates include local Night-Time-Lights (NTL), road network, urban footprint, vegetation, elevation, surface temperature, rainfall.

The use of remote sensing data in the economics literature has expanded dramatically in recent years, as these data have become richer and more accessible, see e.g.

[1] See e.g. Alderman et al. (2002) and and Bedi et al. (2007). Recently, the approach has also been extended to the small area estimation of child malnutrition (Fujii, 2010). For an overview of the literature on small area estimation we refer the reader to Elbers and van der Weide (2014).

Donaldson and Storeygard (2016). Early applications to the measurement of economic activity and welfare include Henderson et al. (2012), which also established the value of these data as predictors of economic outcomes (see also Pinkovskiy and Sala-I-Martin, 2016). These early studies confirm that national income can be accurately predicted using NTL data.[2] Economists have since expanded the set of predictors extracted from remote sensing data and increased the spatial resolution at which economic welfare including poverty is estimated. Jean et al. (2016) is among the early studies to use a comprehensive set of predictors derived from remote sensing data to estimate poverty at the small area level. The studies vary in the resolution of imagery used, with higher resolution images allowing for a richer set of predictors that can be extracted from them (see e.g. Marx et al., 2019). The increased accessibility of remote sensing data combined with the continued demand for highly disaggregated poverty data has prompted a surge in the production of poverty maps in recent years. In the years following Jean et al. (2016), dozens of remote sensing-based poverty and/or household welfare maps have been produced, see e.g. Blumenstock et al. (2015), Blumenstock (2016), Bosco et al. (2017), Burke et al. (2021), Chi et al. (2022), Engstrom et al. (2022), Imran et al. (2014), Newhouse et al. (2022), Pokhriyal and Jacques (2017), Smythe and Blumenstock (2022), Steele et al. (2017), Watmough et al. (2016), Watmough et al. (2019), Ye at al. (Yeh2020), Zhao et al. (2019), and Lee and Braithwaite (2022).

This highlights the importance of assessing the accuracy of remote-sensing-based poverty maps. One logical entry-point in this regard is an assessment of how closely a poverty map obtained using remote sensing data tracks a poverty map derived from population census data. Such assessments are currently under-represented in the literature. Existing studies typically validate their small area poverty estimates by comparing them against direct-estimates obtained from the household survey or by adopting proxies for consumption poverty that can be evaluated using census data (think of asset indices as a proxy for household consumption data). Direct estimates, however, are subject to large sampling error and cover only a small subset of the total number of areas. Accordingly, they fail to confront what motivates the use of small area estimation methods to begin with. While the use of asset indices does allow for evaluation at the small area level (see e.g., Yeh et al., 2020; Chi et al., 2022), asset ownership measures a different concept of welfare than household consumption expenditure does and will generally rank households differently (see the discussion in Section 2). Engstrom et al. (2022) is a notable exception. They use the census-based small area estimates of poverty obtained using the ELL approach as the dependent variable in the model that

[2]More recent examples of empirical studies that use remote-sensing data to track local economic activity include Marx et al. (2019) and Nhu and Noy (2020), among others.

employs predictors derived from remote sensing data. While this is a valuable validation study, it does not provide a test of a poverty mapping approach that relies exclusively on household survey and remote sensing data.

The validation exercise presented here provides a comparison between poverty maps obtained with and without population census data. Employing data for Malawi, we first obtain small area estimates of poverty that combine the Malawi household expenditure survey from 2010 with the unit record population census data from 2008 using the ELL approach. This poverty map serves as our benchmark. We then ignore the population census data and obtain a second poverty map for Malawi by combining the household expenditure survey data with predictors of poverty derived from publicly available remote sensing data. It should be noted that our benchmark too represents a set of estimates (namely the ELL census-based estimates). In the absence of observing household consumption expenditures for each household, which would offer the necessary data to evaluate true small area poverty rates, one will have to rely on estimates of small area consumption poverty rates. We will argue that there are distinct advantages of considering household consumption poverty over wealth index poverty for the validation study (see Section 2.3).

An important motivation for using data from Malawi is that we have access to the exact coordinates of the survey PSUs. This enables us to produce the remote-sensing-based poverty map both using the publicly available geographic coordinates (that have been subjected to random offset to protect the anonymity of households) and exact coordinates (that are not publicly available) to investigate whether the random offsets might undermine the approach – and inspect the correspondence.

Our findings are two-fold. First, assessments of the correspondence between the remote-sensing-based poverty estimates and the census-based poverty estimates depend on how the correspondence is evaluated. The two approaches reveal the same patterns in the geography of poverty in Malawi; it is hard to tell the difference in a side-by-side comparison of the two poverty maps. The statistical correlation between the two different small area estimates of poverty is above 0.9, again confirming their close correspondence. However, there are districts for which the two approaches obtain markedly different estimates of poverty. Accordingly, poverty maps obtained using remote sensing data can be expected to do well when the decision maker is interested in estimates or comparisons of poverty between assemblies of areas. Yet, the decision maker may not be served as well by remote-sensing-based estimates when the focus is on estimates for specific small areas. Second, the random offsets embedded in the geographic coordinates of the public use data do not meaningfully alter the estimated geography of poverty. But they may, however, lead to an under-estimation of the stan-

dard errors of the small area poverty estimates (i.e. an over-estimation of precision). The reason for this is that the random offsets partially destroy the spatial correlation structure in the data and error term.

The decision to restrict attention to publicly available remote sensing data is a deliberate one. This ensures that the data are available at no cost in any given country, and so the approach can easily be replicated and readily repeated as new survey data become available. By the same token, our exercise arguably provides a conservative assessment, i.e. a lower bound of success as higher resolution remote sensing data that come at a cost will allow for more precise estimates. It should be noted that with the rapid technological advancements in this field, what is costly today will in all likelihood become low-cost or publicly-available in the near future.

The paper is organized as follows. An overview of the two small area estimation methods is presented in Section 2. In this section we also offer a methodological contribution by deriving standard error estimates for the remote-sensing based poverty map. Section 3 describes the different data used in this study. Our empirical findings are presented in Section 4. Finally, Section 5 concludes.

2 Methodologies

2.1 Population census-based small area estimation of poverty

Elbers et al. (2003; ELL) pioneered the use of unit record population census data combined with household consumption expenditure survey data for the small area estimation of poverty and inequality. ELL assume the following data generating process (DGP):

$$y_{ah} = x_{ah}^T \beta + v_a + \varepsilon_{ah},$$

where y is log per capita income or expenditure, x is a vector of predictors, and where v and ε are zero expectation error terms that are independent of each other with variances denoted by σ_v^2 and $\sigma_{\varepsilon,ah}^2$, respectively. The subscripts indicate household level h residing in small area a. The predictors x may consist of variables observed at different levels of aggregation, i.e. household level employment status and local area employment rates may both serve as good predictors of household expenditure.

The model permits heteroskedasticity in the household errors ε_{ah} to allow for instances where the predictability of expenditures varies between demographics (i.e. elder versus young), employment status (i.e. unemployed versus employed), and sectors (i.e. blue collar versus white collar). Specifically, the variance $\sigma_{\varepsilon,ah}^2$ is assumed to be a func-

tion of household and area characteristics, denoted by z_{ah}. A logistic transformation of the squared error will serve as the dependent variable (as the distribution of the untransformed squared error tends to be heavily skewed):

$$\ln\left(\frac{\varepsilon_{ah}^2}{A-\varepsilon_{ah}^2}\right) = z_{ah}^T\alpha + \epsilon_{ah},$$

where the vector z_{ah} includes the constant term. An estimate of $\sigma_{\varepsilon,ah}^2 = E\left[\varepsilon_{ah}^2|z_{ah}\right]$ is obtained by applying the inverse of the logistic transformation (see Elbers et al., 2003). Estimates of the household level variances are re-scaled such that the sample mean matches the estimate of the unconditional variance σ_ε^2, i.e. $E[\sigma_{\varepsilon,ah}^2] = \sigma_\varepsilon^2$.

For households sampled in area a, the residuals $e_{ah} = y_{ah} - x_{ah}^T\beta = v_a + \varepsilon_{ah}$ are informative of the latent area error v_a. Assuming both errors are normally distributed, it follows that conditional on the residuals, the error e_{aj} is normally distributed with mean and variance given by (see e.g. van der Weide, 2014):

$$E\left[e_{aj}|e_{a1},\ldots,e_{am_a}\right] = \gamma_a\sum_h \alpha_{ah}e_{ah} \quad (1)$$

$$var\left[e_{aj}|e_{a1},\ldots,e_{am_a}\right] = \sigma_v^2 - \gamma_a^2\left(\sigma_v^2 + \sum_h \alpha_{ah}^2\sigma_{\varepsilon,ah}^2\right) + \sigma_{\varepsilon,ah}^2, \quad (2)$$

where $\gamma_a = \sigma_v^2/\left(\sigma_v^2 + \sigma_\varepsilon^2/m_a\right)$, with m_a equal to the number of survey households sampled into the area a, and where:

$$\alpha_{ah} = \frac{\left(\frac{1}{\sigma_{\varepsilon,ah}^2}\right)}{\sum_h\left(\frac{1}{\sigma_{\varepsilon,ah}^2}\right)}.$$

In practice, the empirical residuals $e_{ah} = y_{ah} - x_{ah}^T\hat{\beta}$ that can be observed in the survey data, where $\hat{\beta}$ denotes the estimator for β, will serve as the substitute for e_{ah}.

If there are sampling weights these will be absorbed by (and modify) the weights α_{ah} (see van der Weide, 2014). Conditioning on the survey data when drawing the errors is known as Empirical Bayes (EB) prediction (see e.g. Molina and Rao, 2010). Note that EB prediction only concerns the drawing of the area errors, and only those areas that have been sampled into the survey. The gain in precision resulting from conditioning on the survey data for these areas is most notable when the variance σ_v^2 and the number of sampled households m_a are relatively large.

The small area estimates of poverty and corresponding standard errors are obtained by means of bootstrapping. Let R denote the number of simulations. The estimator for poverty then takes the form:

$$\hat{H} = \frac{1}{R}\sum_{r=1}^{R} h\left(\tilde{y}^{(r)}\right),$$

where $h(y)$ is a function that converts the vector y with (log) consumption expenditures for all households into the corresponding poverty rate, and where $\tilde{y}^{(r)}$ denotes the r-th simulated vector with elements:

$$\tilde{y}_{ah}^{(r)} = x_{ah}^{T}\tilde{\beta}^{(r)} + \tilde{v}_{a}^{(r)} + \tilde{\varepsilon}_{ah}^{(r)}.$$

where the model parameters $\tilde{\beta}^{(r)}$, $\tilde{\sigma}_{v}^{(r)}$ and $\tilde{\sigma}_{\varepsilon}^{(r)}$ and the errors $\tilde{v}_{a}^{(r)}$ and $\tilde{\varepsilon}_{ah}^{(r)}$ are drawn from their estimated distributions. We draw $\tilde{\beta}^{(r)}$, $\tilde{\sigma}_{v}^{(r)}$ and $\tilde{\sigma}_{\varepsilon}^{(r)}$ by re-estimating the model parameters using the r-th bootstrap version of the survey.[3] The error term $\tilde{v}_{a}^{(r)} + \tilde{\varepsilon}_{ah}^{(r)}$ is drawn from a normal distribution with mean and variance given by eq. (1)-(2). Finally, the point estimates and their corresponding standard errors are obtained by computing respectively the average and the standard deviation over R imputed poverty rates. Note that the standard errors of the poverty estimates are driven both by model error and idiosyncratic error. The contribution of the idiosyncratic error (the area- and the household error) will tend to zero as the size of the target population of a small area (in the population census) tends to infinity. Similarly, the contribution of the model error tends to zero when the sample size of the survey (which is used to estimate the model parameters) tends to infinity.

2.2 Remote-sensing-based small area estimation of poverty

Relying on remote sensing data to estimate small area poverty rates alters the dimensions of the model with which we work. The unit of observation will be a geographic unit, think of a small administrative unit such as a village. The dependent variable in this case will be village level poverty rather than household level consumption expenditure.[4] Furthermore, restricting the set of the covariates that can be derived from publicly available remote sensing data could conceivably increase the magnitude as well as the spatial correlation structure of the error term. Adopting a spatial error model will help fit this spatial structure (e.g. Anselin, 2001).

[3] Alternatively, the model parameters can be drawn from their estimated asymptotic distribution.

[4] Alternatively, one could consider the household as the unit of observation while all predictors vary at the village level, as is done in e.g., Newhouse et al. (2022). There is however little to be gained from adopting a regression at the household level when using only area-level explanatory variables, provided that the unit of observation used in the area-level model is at the model disaggregated level at which the predictors are observed. In other words, one does stand to lose precision if the small area level is adopted as the unit of observation, while the predictors can be observed at the village level (which is a more granular level than the small area level).

We use a spatial error model (henceforward SEM) that assumes iid normal shocks to local poverty spilling over to other locations depending on distance. It is assumed that village level poverty q can be described (in matrix notation) by:

$$q = X\theta + u,$$

where X is the matrix with geo-covariates (rows indicating villages and columns indicating covariates), u satisfies $u = \rho W u + \varepsilon$, where ε is a vector of independent local errors satisfying $\varepsilon \sim N(0, \sigma^2 I)$ and $E[\varepsilon|X] = 0$, and where W is a spatial weight matrix with elements w_{ij} measuring the spatial correspondence between villages i and j (relative to other villages). We define $w_{ij} = t_{ij}^{-\tau}$ (for $i \neq j$) as a function that is inversely related to travel time t_{ij} between i and j with $\tau > 0$. We take W to be symmetric, with $w_{ii} = 0$, and normalize it by its largest eigenvalue, i.e. $W = W^*/\lambda_{max}$, where W^* is the un-normalized spatial weight matrix and λ_{max} is the largest eigenvalue of W^* (see e.g. Bell and Bockstael (2000) and Kelejian and Prucha (2010)). Locations that are in close proximity to each other are exposed to the same latent geographical features which induces spatial correlation. These correlations are modeled as the product of W and ρ, where ρ is a scalar, expected to be positive. The spatial covariance structure of the error term $u = (I - \rho W)^{-1}\varepsilon$ is seen to satisfy:

$$\Sigma = Euu^T = \sigma^2 CC^T,$$

where $C = (I - \rho W)^{-1}$.

Let $m = X\theta$ be the mean of q conditional on X. Let the subscripts S and O indicate whether the village is "in sample" or "out of sample", respectively. For ease of exposition, let us also sort the villages such that the "in sample" villages are stacked above the "out of sample" villages, which permits for the following notation:

$$\begin{pmatrix} q_S \\ q_O \end{pmatrix} = \begin{pmatrix} m_S \\ m_O \end{pmatrix} + \begin{pmatrix} u_S \\ u_O \end{pmatrix}.$$

The corresponding partition of the variance matrix Σ takes the form:

$$\Sigma = \begin{bmatrix} \Sigma_{SS} & \Sigma_{SO} \\ \Sigma_{OS} & \Sigma_{OO} \end{bmatrix}.$$

Suppose that we know the model parameters (in practice we will have to work with estimates) such that m_S and m_O are both observed (note that X is observed for all villages). Let us also assume for now that we observe q_S, the outcome variable of

interest for a sub-set of locations. (Since the village level poverty rates are derived from a sample of households in any given village, we observe q_S with error due to sampling; we will return to this later.)

The objective then is to estimate q_O given our observations (or estimation) of m_S, m_O, Σ, as well as q_S, so as to obtain a map of poverty that covers the whole country. If the data exhibits spatial correlation (i.e. $\rho \neq 0$), then observing q_S and m_S will carry information that is relevant for the estimation of q_O above and beyond the information that is conveyed in m_O. It follows from the normality assumptions that the Best Linear Unbiased Predictor (BLUP) of q_O solves (Goldberger, 1962):

$$\begin{aligned} BP(q_O) &= m_O + \operatorname{cov}[q_O, q_S] \operatorname{var}[q_S]^{-1} (q_S - m_S) \\ &= m_O + \Sigma_{OS}\Sigma_{SS}^{-1} (q_S - m_S) . \end{aligned}$$

Henceforward, the BP estimates derived from the SEM model may be interchangeably referred to as SEM-estimates.

Let us next account for the fact that q_S is observed with sampling error. We shall denote the sampling direct estimate of q_S by $\hat{q}_S$:

$$\hat{q}_S = q_S + e_S = m_S + u_S + e_S,$$

where e_S denotes sampling error. This modifies the BLUP estimate of q_O to:

$$\begin{aligned} \tilde{q}_O &= m_O + \operatorname{cov}[q_O, \hat{q}_S] \operatorname{var}[\hat{q}_S]^{-1} (\hat{q}_S - m_S) \\ &= m_O + \Sigma_{OS} (\Sigma_{SS} + D_S)^{-1} (\hat{q}_S - m_S) , \end{aligned}$$

where D_S is a diagonal matrix with the sampling variances (i.e. variance of e_S) as diagonal elements. Note that when the magnitude of sampling variance is large (i.e. when $\hat{q}_S$ denotes a noisy observation of q_S), more weight is given to the synthetic estimator m_O.

We can similarly use the model to obtain a more precise estimate of q_S. It can be verified that the BLUP estimator of q_S in this case satisfies:

$$\begin{aligned} \tilde{q}_S &= m_S + \operatorname{cov}[q_S, \hat{q}_S] \operatorname{var}[\hat{q}_S]^{-1} (\hat{q}_S - m_S) \\ &= m_S + \Sigma_{SS} (\Sigma_{SS} + D_S)^{-1} (\hat{q}_S - m_S) . \end{aligned}$$

Note that $\tilde{q}_S$ is of the same form as the estimator put forward by Pratesi and Salvati (2008), who consider a different choice of SEM model. As Pratesi and Salvati (2008) restrict their analysis to in-sample prediction, they do not offer an estimator for q_O.

Precision In practice a prediction of q_O is also subject to error induced by the sampling variation of estimators for model parameters. Denote by a 'check' $\check{}$ estimators and predictions derived from them. The covariance matrix H of prediction errors of $\check{q}_O$ can be decomposed into three terms:

$$H = E\left(q_O - \check{q}_O\right)\left(q_O - \check{q}_O\right)^T = \ V_1 + V_2 + V_3,$$

where:

- V_1 reflects the sampling variation of $\check{m}_O = X_O\check{\theta}$, resulting from the fact that θ is estimated from the survey data (indicated by $\check{\theta}$) and the sampling distribution of $\check{\theta}$ carries over to $X_O\check{\theta}$
- V_2 represents sampling variation induced by $\check{\theta}$, as well as the estimators of the variance parameters, on $\Sigma_{OS}\left(\Sigma_{SS} + D_S\right)^{-1}\left(\hat{q}_S - m_S\right)$
- V_3 reflects the variance component that is due to the residuals $u_O = q_O - m_O$, to the extent that they are independent of (and therefore cannot be predicted by) $\hat{q}_S - m_S$

Estimators for V_1, V_2 and V_3 are proposed in Lemma 1 and in Appendix 1.

Aggregation In general, precision will be comparatively low at the level of geographical units, while predicting poverty over larger areas increases precision as idiosyncratic errors are averaged out. Aggregation over groups of units ('areas') can be represented by a linear mapping B. The estimator for the area-level poverty rate is seen to equal $B\check{q}_O$, while its variance-covariance matrix solves:

$$E\left[B\left(q_O - \check{q}_O\right)\left(q_O - \check{q}_O\right)^T B^T\right] = BHB^T.$$

The standard errors of the area-level poverty estimates reported in Table 5 (in the column headed 'SEM') are obtained by evaluating the square root of the diagonal elements of BHB^T.

Lemma 1 *Let $q, u, W, X, \Sigma, \varepsilon$; $\omega = (\rho, \theta, \sigma^2, \tau)$ and index sets S, O be as defined and discussed above; assume further that observation errors e_S are independent of ε and normally distributed with zero mean and given variance matrix $D_S = Ee_Se_S^T$. Let $\check{\omega} = (\check{\rho}, \check{\Sigma}^2, \check{\theta}, \check{\tau})$ be maximum-likelihood estimators of the corresponding parameters with variance matrix $V(\check{\omega})$, and let $\check{q}_O = X_O\check{\theta} + \check{\Sigma}_{OS}[\check{\Sigma}_{SS} + D_S]^{-1}(\hat{q}_S - X_S\check{\theta})$ denote predictions of q_O. Denote by S_{OS} the matrix:*

$$S_{OS} = \Sigma_{OS}[\Sigma_{SS} + D_S]^{-1}$$

Then: (i)

$$q_O - \check{q}_0 = X_O(\theta - \check{\theta}) + \left(S_{OS}(u_S + e_S) - \check{S}_{OS}(\hat{q}_S - X_S\check{\theta})\right) + r_O, \tag{3}$$

with:

$$r_O = u_O - E(u_O|u_S + e_S).$$

And: (ii) The three terms on the right side are asymptotically independent and their covariance matrices can be consistently estimated by:

$$\begin{aligned}
V_1 &= E_\omega\left[X_O(\theta - \check{\theta})(\theta - \check{\theta})^T X_O^T\right] = X_O V(\check{\theta}) X_O^T \\
V_2 &= E_\omega\left[\left(S_{OS}\left(u_S + e_S\right) - \check{S}_{OS}\left(\hat{q}_S - X_S\check{\theta}\right)\right)\left(S_{OS}\left(u_S + e_S\right) - \check{S}_{OS}\left(\hat{q}_S - X_S\check{\theta}\right)\right)^T\right] \\
&\approx E_\omega\left[\left(S_{OS} - \check{S}_{OS}\right)\hat{q}_S\hat{q}_S^T(S_{OS} - \check{S}_{OS})^T\right] + \check{S}_{OS}X_S V(\check{\theta}) X_S^T \check{S}_{SO} \\
V_3 &= E\left[r_O r_O^T\right] \approx \check{\Sigma}_{OO} - \check{\Sigma}_{OS}[\check{\Sigma}_S + D_S]^{-1}\check{\Sigma}_{SO}.
\end{aligned}$$

Proof Here we provide an outline of the proof. The detailed proof is given in Appendix 1. The first assertion follows from:

$$q_O = x_O\theta + u_O = x_0\theta + E(u_0|q_S - X_S\theta) + r_0 = x_0\theta + E(u_0|u_S + e_S) + r_0.$$

Asymptotic independence of the first and second term on the right hand side of eq. (3) follows from the independence of $\check{\theta}$ and $\hat{q}_S - X_S\check{\theta}$ and the fact that the information matrix of the model for $\hat{q}_S$ is block diagonal in θ and (ρ, σ^2, τ) respectively. As for independence between the third and the first two terms, note that the estimator $\check{\omega}$ is a function of $u_S + e_S$ and hence is uncorrelated with $r_O = u_O - E(u_O|u_S + e_S)$, which together with the normality assumptions implies independence.

To emphasize that the expectations of the first two terms are taken with respect to the sampling distribution of ω we have used the symbol E_ω. The expressions for the asymptotic covariance of the first and third terms of eq. (3) are obvious. Details on the covariance matrices of the three terms and estimators for them are discussed in Appendix 1. □

2.3 Consumption poverty versus asset index poverty

The study of poverty and inequality in developing countries is confronted by multiple challenges. Not least are those relating to the absence of pertinent data. Household surveys typically underpin the analysis of distributional outcomes, with household income or consumption most commonly serving as the key underlying indicator of economic

welfare. It remains however, that many household surveys fail to collect data on such variables, and this then limits the extent to which the surveys can inform analysis of distributional outcomes. A well-known source of household survey data are the Demographic and Health Surveys that have been fielded in nearly 100 countries. These surveys collect hugely valuable data on health outcomes and a variety of other indicators, but do not collect detailed information on incomes or expenditures. Filmer and Pritchett (2001) propose circumventing this constraint by proxying wealth from asset ownership variables based on weights derived from principal components analysis. They find that their wealth indicator is reasonably well correlated with per-capita expenditures in the India context. However, the correlation varies notably across countries. Filmer and Scott (2012) calculate the correlation coefficient between per capita consumption and the Filmer-Pritchett asset index in an additional 11 developing countries and observe an average correlation of only 0.5.

Inconsistent rankings based on asset indices versus those based on common measures of income or consumption may arise for multiple reasons. For example, most asset indices employed in practice do not reflect the net wealth position of households as, typically, they do not include the liabilities households may hold. Indebtedness can be an important dimension of poverty in its own right and can also be a source of impoverishment. Second, asset indices such as those constructed using DHS data often do not include a key asset in the developing country context, namely land. In rural settings landholdings are frequently the most important source of wealth, but collecting data on the size and, crucially, value of landholdings is typically very difficult. Drèze et al. (1998) show for the case of an intensively studied village in India, that alternative approaches to assessing the value of landholdings can dramatically alter the measured wealth status, and ranking, of rural agricultural households. While conventional estimates of income or consumption do not directly reflect the debt-status of households, nor the value of agricultural landholdings, these household characteristics would be expected to shape consumption patterns or incomes in ways that are different, and likely less well captured by asset indices. A counter argument, in favor of asset indices, is that these may not be as prey to measurement error as current income or consumption and may thus be a better proxy for long-term welfare.

Hentschel et al. (2000) describe the experience in Ecuador of constructing a poverty map using a "basic needs" index, built up from an ad-hoc weighted count of selected characteristics and assets available at the household level in the population census. They show that this index generates a very different ranking of households than would a comprehensive consumption measure. Competing "basic needs" maps within the same country, produced by different agencies or institutions, easily become a source

of disagreement or confusion. Hentschel et al (2000) propose the construction of a consumption-based poverty map based on the combination of household survey data with the population census. The underlying method they employ was subsequently refined in Elbers et al. (2002, 2003). Hentschel et al (2000) argue that an appealing feature of such consumption-based poverty maps is that they allow estimates of poverty produced at the detailed sub-national level to be assessed and interpreted alongside the national-level estimates that are routinely calculated and reported from household survey data, and that typically underpin official poverty statistics.

3 Data

Figure 1 (left panel) shows the map of Malawi, our study area. Malawi is a landlocked country in the South-East of Sub-Saharan Africa, situated between Zambia, Mozambique, and Tanzania, with a large lake alongside its Eastern border. Its territory has an elongated shape that is divided into a Northern, Central, and Southern region, with over 80 percent of the population residing in the Central and Southern regions (see Table 1). The country is still highly rural with just over 15 percent of the population (in 2010) residing in urban areas. The three largest cities span all three regions: Lilongwe (the capital) in the Central region with a population of 674,448 (2008), Blantyre in the Southern region with a population of 661,256 (2008) , and Mzuzu in the Northern region with a population of 133,968 (2008) (Malawi NSO 2008).

Both approaches considered in this paper rely on the 2010-11 Malawi Third Integrated Household Survey (IHS3), which is part of the Living Standards Measurement Study - Integrated Surveys on Agriculture (LSMS-ISA) project (Malawi National Statistical Office (NSO) and World Bank, 2012). The survey contains high-quality data on household consumption expenditure that is used to measure poverty, defined as the percentage of individuals whose total annual household consumption per capita fall below the national poverty line. Household expenditure and the poverty line are expressed in Malawi kwacha using February/March 2010 prices.[5] The poverty line, the sum of the food and non-food poverty lines, equals kwacha 37,001.68.[6]

[5]For details on the price deflator used in the Malawi IHS3, we refer the reader to Malawi National Statistical Office and World Bank (2012). For a more general discussion on spatial (and temporal) price deflation and real consumption measurement, see e.g. Gibson et al. (2017) and van Veelen and van der Weide (2008).

[6]The food poverty line in the IHS3 is Kwacha 22,956, the monetary cost of 2,400 calories per person per day, where the price of a calorie is estimated from the population in the 5th and 6th deciles of the aggregate distribution for consumption. The non-food poverty line (non-basic food needs) in the IHS3 is kwacha 14,045, and is estimated as the average non-food consumption of the population whose food consumption is close to the food poverty line.

The ELL approach combines the household survey with the Malawi 2008 Population and Housing Census (henceforward referred to as the census). The census provides individual and household level information on household composition, education, employment, dwelling characteristics and asset ownership. These data are used to obtain multiple imputed values for household consumption expenditures for all households in the census, which are then aggregated to obtain estimates of poverty at the small area level. Alternatively, the SEM approach model combines the household survey with remote sensing data, such as night time lights, urban footprints, major roads, population density, vegetation, surface temperature, rainfall. It uses these data to obtain multiple imputed values for village level (i.e. enumeration area) poverty rates which too can be aggregated to the small area level to facilitate a comparison between the two approaches. The following sub-sections provide further details on the respective databases used.

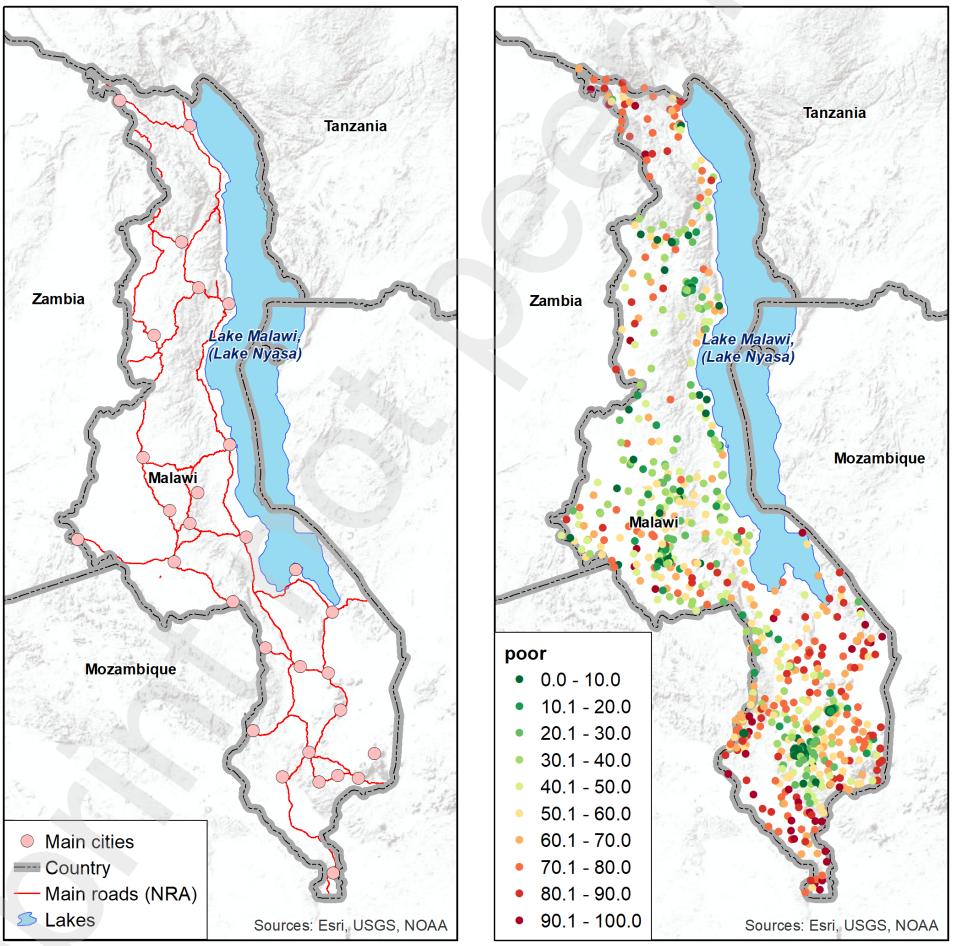

Figure 1: Study area (left panel) and LSMS survey points color-coded by poverty rate (right panel: count of individuals below the national poverty line divided by total individuals sampled per household cluster)

In 2010	% population	% urban
Malawi	100	15.2
North	13.1	14.3
Central	42.6	15.4
South	44.3	15.2

Table 1: Population in 2010

3.1 Household expenditure survey

The Third Integrated Household Survey (IHS3) was conducted between March 2010 and March 2011. Its target universe consists of the households and individuals in all districts, except for the Likoma district. A total of 12,271 households were interviewed (with 668 replacements), see Table 2. The sample is designed to be representative at the nation, region, and district level, as well as for urban and rural areas at the national and regional level. The sampling frame is based on listing information and cartography from the 2008 Malawi Population and Housing Census (the same Census used in this poverty mapping project). The IHS3 uses a stratified and two stage sampling design with: (1) 767 primary sampling units (PSUs); and (2) 16 households randomly selected from each PSU. In the actual population, at the time of the census, an EA (which serves as the PSU) on average counted a total of 235 households. A minimum of 24 EAs were chosen in each of the 31 districts included in the IHS3 (out of the 32 districts in Malawi).

The PSUs are geo-referenced such that they can be placed on a map and merged with remote sensing data. The right-panel of Figure 1 shows the observed poverty rates for the 767 PSUs (i.e. villages), obtained by evaluating for each PSU the share of sampled individuals with a household expenditure per capita that is below the poverty line. While there is a larger concentration of sampled clusters in and around major urban centers, the survey provides a complete coverage of the country including rural and more remote areas. Poverty rates are seen to vary considerably across space. On average however, poverty levels tend to be higher in more remote areas that located further away from the urban centers. Note that this does not imply that poverty is less of a concern in urban centers. In fact, the majority of the poor plausibly reside in urban centers due to the higher concentration of population there, i.e. despite the lower prevalence of poverty.

	2008 census	IHS3
# regions	3	3
# districts	32	31
# Wards/TAs	353	279
# EAs	12,575	767
# Households	2,859,720	12,271

Table 2: Descriptive statistics of 2008 census and IHS3

3.2 Population census

The enumeration for the 2008 Population and Housing Census for Malawi was carried out between June 8 and 28, 2008. It covers regular households, institutions, and the homeless. For the purpose of this exercise, institutions and the homeless are omitted. The total population was found to be 13,077,160. Table 2 reports the number of households in the census and number of sampled households in the survey at different levels of aggregation. The most disaggregated level at which poverty will be estimated is the Traditional Authority (TA) level, which we may also refer to as the small area level.

The administrative boundary data are taken from the National Statistical Office (NSO) of Malawi in order to match the boundaries with the census data at the enumeration area level. Malawi has 12,666 enumeration area boundaries from the 2008 census. The IHS3 household survey has 12,271 observations from 768 enumeration areas.[7]

3.3 Public remote sensing data

The analysis uses publicly available global remote sensing data from numerous sources and summarizes predictors derived from them at the administrative level (i.e. TA level) for Malawi. For most of the variables, we take advantage of the geographically comprehensive raster data structure, which uses pixels to form a grid that covers all land areas. From the raster structure we extract geo-referenced information at the administrative level. Using Geographic Information Systems (GIS) and Google Earth Engine, we overlay the administrative boundaries with other geo-referenced gridded data, (e.g. temperature, precipitation, elevation, etc.). When an administrative area consists of multiple pixels (which is typically the case), the pixel observations are collapsed (i.e.

[7]See the background report for more details on the design, sample and survey results at: http://microdata.worldbank.org/index.php/catalog/2939/download/41224

aggregated) to obtain summary statistics (e.g. mean and sum) at the administrative area level.[8] This process of extraction provides a nearly exhaustive correspondence. For a select few cases, we make two minor adjustments:

- Administrative areas near boundaries with water or small islands do not always overlay with the raster data. In this case, we first extract data by center point if the results from the polygon extract are missing. Second, we make a small adjustment by filling in the missing cells from the original raster with the average value of data from contiguous neighbors.

- When the polygon representing an administrative area is much smaller than the raster pixel size, we disaggregate the raster data. If the center point of the polygon is located in an area of "no data" (e.g. a lagoon or a lake), then we move the center point to the nearest land pixel. We also calculate the area at the pixel size to ensure consistency between the sum of pixel values and the corresponding land area.[9]

For the remaining data, we use vector data to calculate distances. Table 10 in Appendix 4 lists the variables considered in this paper that have been derived from publicly available remote sensing datasets. These variables are assembled into various themes related to initial conditions, agriculture, climate, environment, socio-economic conditions, population and remoteness. Specific details on the source and construction of the variables can be found in Appendix 2.

4 Empirical application to Malawi

4.1 Model selection and estimation of model parameters

4.1.1 ELL approach: Household level data with household expenditure as dependent variable

We rely on variables that are available in both the survey and the census. The first step of our model selection procedure groups candidate variables by type: (i) demographics, (ii) education, (iii) employment, (iv) dwelling characteristics, (v) asset ownership,

[8] We use two geometries (center points and polygons) to represent the administrative area. We extract raster data by polygon using the extract function in R. In Google Earth Engine, we use the Reducer function over a region at the administrative level.

[9] An area calculation with the vector data provides more accurate measurements, however it does not always yield consistent area with the pixel calculations, which is necessary to ensure variables with the share are between 0 and 1.

and (vi) village level variables. This latter group includes variables that are created by computing mean values of selected variables from the population census at the enumeration-area (EA) levels (think of local employment rates, share of the population with given levels of education and the penetration of phone use) and variables derived from the publicly available remote data (think of night-time-lights, distance to nearest city and river, local climate, and agricultural land quality).

Next, we regress the log of per capita expenditure (combined with region and urban dummies) on each group separately using Ordinary Least Squares (OLS). Independent variables that are not significant at the 5% level are sequentially dropped from the regressions, starting with those that are least significant. Once all groups have been considered we select the group that provides the best fit, combine this with the group that provides the second-best fit, and continue to trim the variables that cease to be significant. This process is continued until every single group has been given a chance to contribute one or more variables to the model. We allow for interactions with the urban-rural dummy to capture selected area heterogeneities (see Tarozzi and Deaton, 2009; Tarozzi, 2011). Finally, LASSO model selection is implemented to verify that all key variables are accounted for.[10] The same procedure is applied to obtain the heteroskedasticity model describing the variance of the errors. The resulting regression models are included in Appendix 3.

Among the different categories of explanatory variables, dwelling characteristics and household asset ownership rank as the strongest predictors of household consumption expenditure per capita, followed by household head education and employment variables. An interesting observation is that local area night-time-lights, agricultural land quality, and climate data (derived from public remote sensing data that serve as predictors in the SEM approach) are found to be significant when included on their own – but once we control for variables available in the population census, these variables lose their significance and are subsequently dropped from the model.

Once the models have been selected, the model parameters are estimated by going through the following steps: (a) estimate the models using OLS and extract the residuals; (b) estimate the unconditional variance parameters σ_v^2 and σ_ε^2, using Henderson's method III (see Henderson, 1953; and Searle et al., 1992); (c) estimate the conditional variance $\sigma_{\varepsilon,ah}^2$; and finally (d) construct the covariance matrix $\Omega = E\left[vv^T + \varepsilon\varepsilon^T|x\right] = \sigma_v^2 I_n + \text{diag}(\sigma_{\varepsilon,ah}^2)$, and compute the feasible GLS estimator for β.

[10] In this paper we restrict ourselves to using standard regression techniques, avoiding novel approaches developed by the Machine Learning community. An investigation into the value-added of ML in a context such as ours, where relatively few predictor variables are available, is beyond the scope of this study. In addition, using alternative estimation techniques would confound this paper's comparison of using different sources of data for poverty prediction.

While the regression coefficients purely capture correlations and not causal effects, it is instructive to verify which predictors stand out and inspect the signs of their effects. As one would expect, higher (lower) quality materials used for roof, floor, and walls as well as a larger (smaller) number of rooms are positively (negatively) correlated with (log) household expenditure. Household assets, particularly, owning a car, fridge, cooker, and tv, all report positive regression coefficients. Furthermore, we observe that per capita household expenditures tend to be higher (lower) for smaller (larger) households, and for households with a more (less) educated household head.

4.1.2 Village level data with village poverty rate as dependent variable

A similar procedure is adopted to build the model for village level poverty using variables derived from public remote sensing data as predictors which are grouped into: (i) population density and land size, (ii) night time lights, (iii) road connectivity and distance to cities, borders and rivers, (iv) climate and environment, and (v) agricultural suitability variables. Among these categories of variables, night time lights, market access (i.e. road connectivity), and local climate and environment report the highest correlations with village poverty with similar in-sample goodness-of-fit levels. The final model has 15 explanatory variables.

We estimate the model by maximum likelihood. The distance decay parameter τ is set to $\tau = 0.54$, which maximizes the in-sample goodness-of-fit (measured by the correlation between predicted and observed poverty rates). It turns out that, except for ρ, the parameters (as well as their conditional covariance matrix) are not very sensitive to the choice of τ. The estimation results are reported in Table 3, where poverty is measured as a percentage. Details on working out the asympototic variance-covariance matrix of the parameter estimators are given in Appendix 1.

While also here the regression coefficients capture correlations and not causal effects, it will nevertheless be instructive to inspect the signs of the correlations. Lower poverty rates are associated with higher levels of population density, higher values of night time lights and higher shares of cultivated land that can be irrigated through rainfall. In contrast, higher poverty rates are associated with increased remoteness (i.e. larger distances from major cities), higher levels of ruggedness, and higher rates of forest loss.

4.2 Comparing poverty maps

Before we inspect the estimates of poverty at the small area level, let us first evaluate the estimates at the national and the regional levels. Since the IHS3 survey is designed to be representative at these levels, the survey direct estimates provide a natural bench-

	Poverty (NSO)
(Intercept)	-0.45205 **
	(0.21987)
log population (2012)	-0.03781 ***
	(0.00734)
log ruggedness	0.02051 **
	(0.01005)
log slope	-0.06148 ***
	(0.01745)
log forest loss	0.07558 ***
	(0.01169)
econ. rents from agri. land	0.0085 ***
	(0.00328)
rain-fed crop land (share)	-0.02675
	(0.01844)
distance to major river	-0.00131 ***
	(0.00016)
log travel time to city	0.02608 **
	(0.01045)
log Night Time Lights (2010)	-0.02033 **
	(0.00968)
log Night Time Lights (1996)	-0.03566 ***
	(0.01279)
mean temperature	-0.00406 ***
	(0.00077)
temperature seasonality	-0.00021 ***
	(4e-05)
potential evapotranspiration	0.00147 ***
	(0.00016)
MSE	0.02357
	(0.24747)
ρ	0.75524 ***
	(0.00165)
Observations	765
adj. R^2	0.558

Table 3: Parameter estimates of SEM with $\tau = 0.54$
t-statistics in parentheses: * p<0.10, ** p<0.05, *** p<0.01

mark to compare our model-based estimates against. The ELL and BP estimates are obtained by aggregating the small area estimates to the national and regional level using population weights. The results are shown in Table 4. The model-based estimates from both approaches are seen to line up with the survey-direct estimates. ELL and BP

both identify the North and South as the regions with the highest rates of poverty.[11] While the two approaches disagree slightly on the ordering of these two regions, the observed difference in the poverty estimates lies well within the 95 percent confidence interval.

More striking is the difference in statistical precision between the two approaches. ELL achieves slightly greater statistical precision than the survey-direct estimates (by virtue of combining the survey with unit record population census data), while the standard errors corresponding to the BP estimates are notably larger (between two and three times the ELL standard errors). This difference in precision is in large part due to divergence in number of observations used to estimate the model parameters: 12,271 households in the ELL approach versus 767 PSUs in the SEM approach (used to obtain the BP estimates). The variance of the poverty estimates are largely driven by the variance of the model parameter estimates, which in turn scales with the number of observations used in the respective regression model. As the model adopted by ELL is at the household level, it has 16 times more observations to work with compared to the SEM approach.

	survey-direct	ELL	BP
Malawi	0.507 (0.009)	0.498 (0.006)	0.492 (0.023)
North	0.543 (0.021)	0.562 (0.018)	0.540 (0.022)
Central	0.445 (0.015)	0.438 (0.009)	0.434 (0.024)
South	0.555 (0.014)	0.537 (0.008)	0.533 (0.026)

Table 4: Estimates of poverty from the survey, ELL and BP

The small area estimates of poverty for the two approaches are presented in Figure 2, which shows the two poverty maps side by side. A visual inspection of the two maps shows a striking agreement on the spatial distribution of poverty. Both maps suggest low levels of poverty in the urban areas of Mzuzu, Kasungu, Lilongwe, Zomba, and Blantyre, as well as areas of high poverty in the far north and south of the country. The models differ slightly in areas of intermediate poverty, but the spatial patterns shown in the two maps are generally remarkably similar.

[11] BP estimates slightly underestimate poverty in areas of the Northern region with comparatively low rates of poverty.

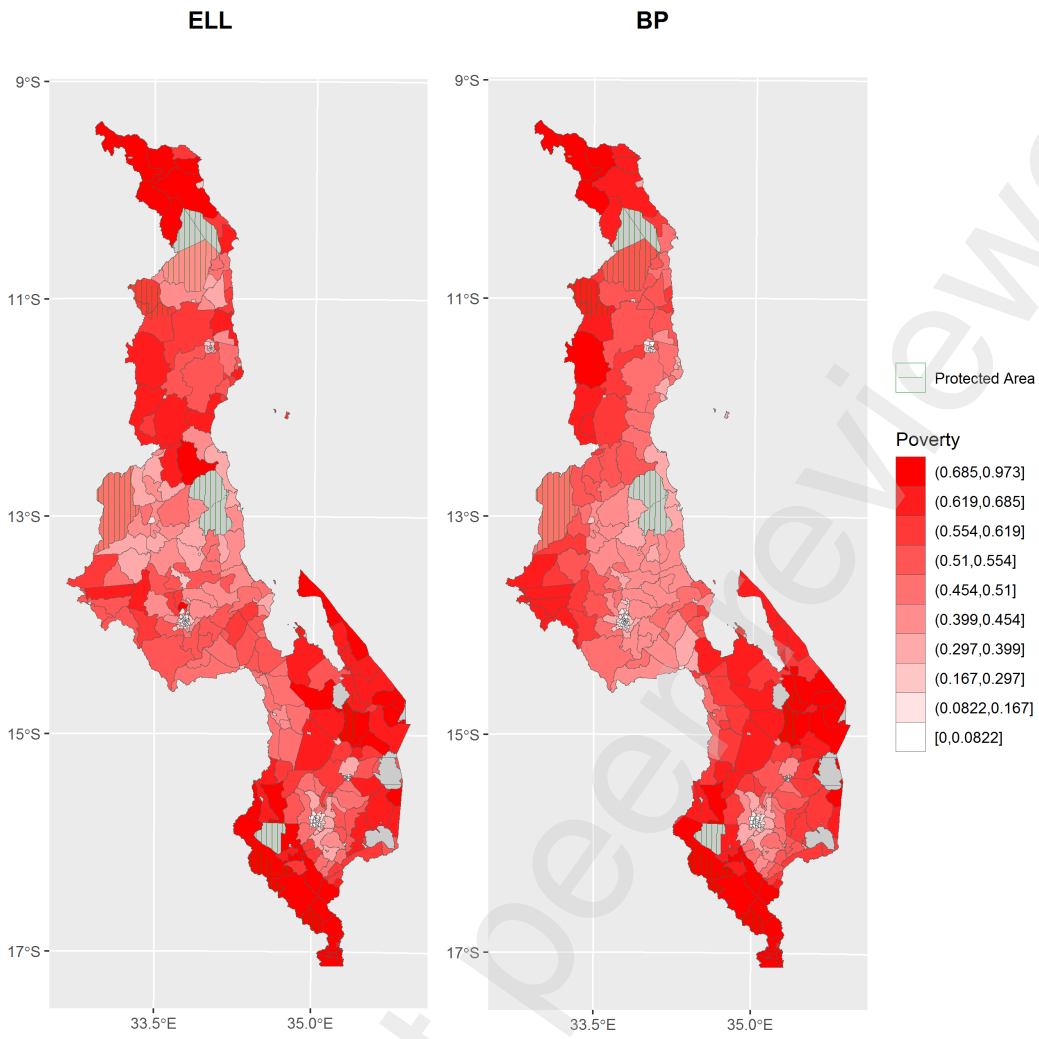

Figure 2: ELL (left) and BP (right) estimates of poverty at the TA level. National parks and game reserves are masked as "protected areas"

Figure 3 maps the difference between the BP and ELL estimates. A well-defined geographical pattern would hint to the possibility that there is a predictable structure in the data that is picked up by one model but not by the other. Noteworthy patterns include an area in the north-central part of the country between Mzuzu and Kasungu coincident with a low density of survey points. In this area the BP estimates are lower than the ELL estimates for three contiguous TAs. Conversely the areas where BP estimates point to higher poverty rates tend to be TAs in an east-west oriented band located south of Lake Malawi.

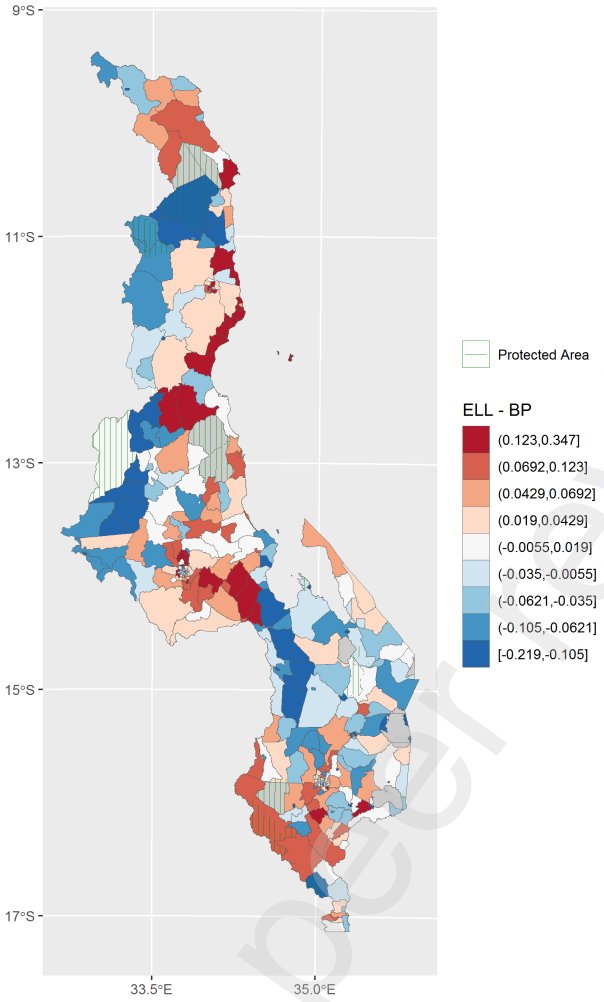

Figure 3: Spatial distribution of differences BP and ELL estimates of poverty at the TA-level

4.3 Comparing individual small area estimates

While the difference between BP and ELL estimates of poverty is small for the large majority of TAs, there are a number of TAs for which the discrepancy is more substantial. Figure 4 plots a histogram of the differences in poverty estimates (the BP estimate minus the ELL estimate). It follows that for about half of the TAs the estimates lie within +/- 5% points of each other (for 75% of TAs, the estimated poverty rates lie within 10% points of each other). The histogram also confirms that the difference between estimates are symmetrically distributed around zero by approximation; there is no meaningful bias in the BP estimates relative to ELL, and overestimates of poverty occur with the same frequency as underestimates of poverty.

Figure 5 evaluates the correspondence between the two approaches by plotting the BP estimates (vertical axis) against the ELL estimates of poverty (horizontal axis),

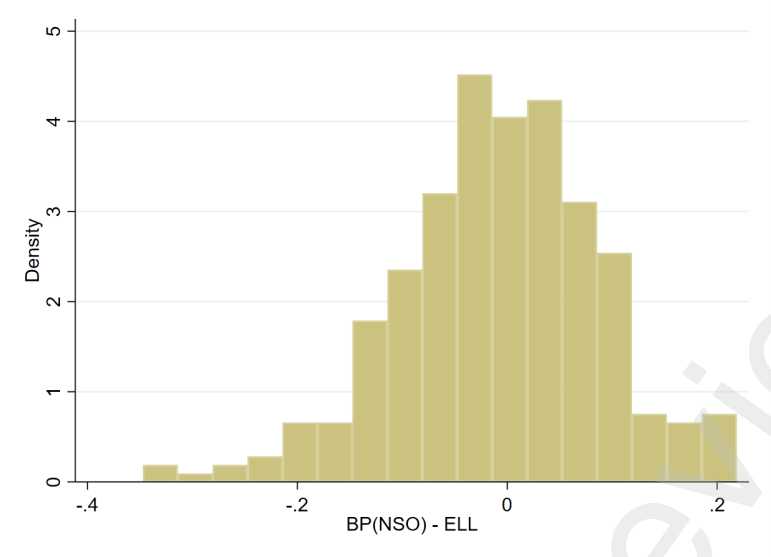

Figure 4: Histogram of difference between BP and ELL estimates of poverty at the TA level

where the size of the dot/bubble measures the population size of the TA. The green line shows the diagonal while the red line plots the fitted regression line. The two lines almost perfectly coincide indicating a near zero bias. The R^2 is 0.83 which confirms the close correspondence between the BP and ELL estimates. The figure also confirms however that despite the close correspondence for the large majority of TAs, there are a select number of TAs, including a number of populated TAs, where the two estimates of poverty differ substantially. This suggests that while the SEM approach (underlying the BP estimates) accurately captures the geography of poverty and conceivably provides accurate estimates of poverty for groups of TAs, estimates may be off for any particular TA (i.e. small area). This means that the value of the SEM approach will depend on the needs of the application.

4.4 Robustness of BP estimates to random offsetting of GPS locations

In order to preserve the confidentiality of sample households and communities, the IHS3 does not provide exact geographic coordinates of the Primary sampling Units (PSUs) in its public use dataset. Such an approach is standard practice; it is rare for exact GPS locations of PSUs to be made available in the public domain. In our assessment of the SEM approach above, however, the precise EA coordinates provided by the Malawi NSO were employed. The question thus arises whether performance of the SEM method is seriously undermined when one is compelled to work with the public use coordinates.

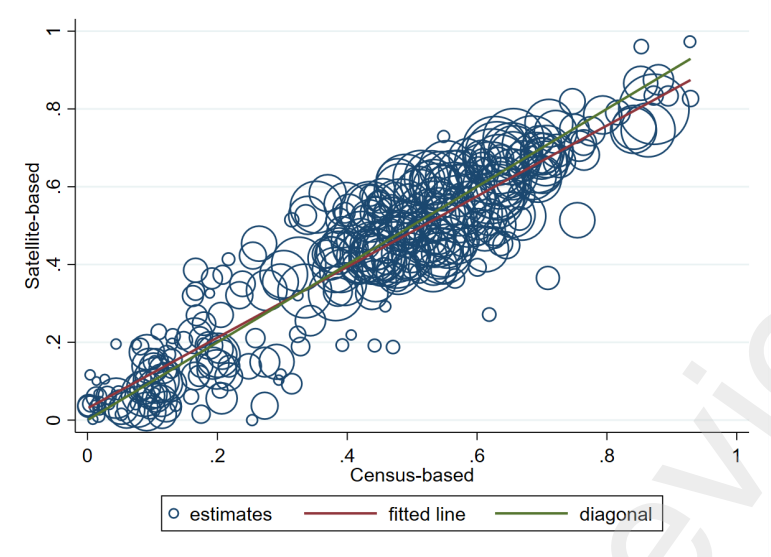

Figure 5: Scatter plot of BP against ELL estimates of poverty at the TA level

The coordinates in the IHS3 made available for public use have been modified using the MeasureDHS methodology (Perez-Heydrich et al. 2013). This method applies a random offset to the exact GPS locations within a specified range that varies between 0-2 km in urban areas, and 0-5 km in rural areas. Two caveats apply. First, the specified range for rural PSUs that are particularly remote (about one percent of rural PSUs) is extended to 0-10 km. Second, the condition is imposed that location of the random offset must be within the sampled district. The result is a set of modified EA coordinates with known limits of accuracy for public use (Malawi NSO and World Bank 2012).

Table 5 compares the regression models obtained with the NSO and the public use coordinates, respectively. Using public coordinates reduces the in-sample goodness-of-fit of the model, as is to be expected. The reduction in adjusted R^2 is reasonably modest, however, from 0.558 to 0.514. The two models are qualitatively similar; all regression coefficients remain significant and are of the same order of magnitude. Some regression coefficients are more sensitive to the choice of public versus exact NSO coordinates than others. Examples of covariates that are found to be particularly stable include: bio1, bio4m pet, slope, ruggedness, and distance to river. Not coincidentally these are covariates that exhibit a high degree of spatial correlation, and as such are less sensitive to random spatial perturbations.

Table 6 reports the statistical correlation coefficients between the two different BP estimates (obtained using public versus exact NSO coordinates) and the ELL estimates of poverty (the benchmark) at the TA level. It can be seen that: (a) While BP estimates obtained with public coordinates exhibit a lower correlation with the benchmark

	Poverty (NSO)	Poverty (Public)
(Intercept)	-0.45205 **	-0.58123 **
	(0.21987)	(0.22611)
log population (2012)	-0.03781 ***	-0.02969 ***
	(0.00734)	(0.00848)
log ruggedness	0.02051 **	0.02977 ***
	(0.01005)	(0.01076)
log slope	-0.06148 ***	-0.05727 ***
	(0.01745)	(0.01836)
log forest loss	0.07558 ***	0.04426 ***
	(0.01169)	(0.01122)
econ. rents from agri. land	0.0085 ***	0.01517 ***
	(0.00328)	(0.00357)
rain-fed crop land (share)	-0.02675	-0.02488
	(0.01844)	(0.02009)
distance to major river	-0.00131 ***	-0.00131 ***
	(0.00016)	(0.00017)
log travel time to city	0.02608 **	0.02195 *
	(0.01045)	(0.01145)
log Night Time Lights (2010)	-0.02033 **	-0.01452
	(0.00968)	(0.01016)
log Night Time Lights (1996)	-0.03566 ***	-0.03638 ***
	(0.01279)	(0.01248)
mean temperature	-0.00406 ***	-0.00401 ***
	(0.00077)	(0.00079)
temperature seasonality	-0.00021 ***	-0.00025 ***
	(4e-05)	(4e-05)
potential evapotranspiration	0.00147 ***	0.00153 ***
	(0.00016)	(0.00017)
MSE	0.02357	0.02622
	(0.24747)	(0.3463)
ρ	0.75524 ***	0.6478 ***
	(0.00165)	(0.00181)
Observations	765	738
adj. R^2	0.558	0.514

Table 5: Parameter estimates of SEM with $\tau = 0.54$
t-statistics in parentheses: * p<0.10, ** p<0.05, *** p<0.01

ELL estimates compared to the estimates obtained with exact NSO coordinates, the difference is small, and (b) as expected, the correlation between the two BP estimates is high. Figures 8 and 9 confirm the close correspondence by comparing the respective poverty maps.

	ELL	BP (NSO)	BP (Public)
ELL	1.000		
BP (NSO)	0.910	1.000	
BP (Public)	0.883	0.987	1.000

Table 6: Correlation matrix of estimates of poverty from ELL, BP (NSO), and BP (Public) at the TA level.

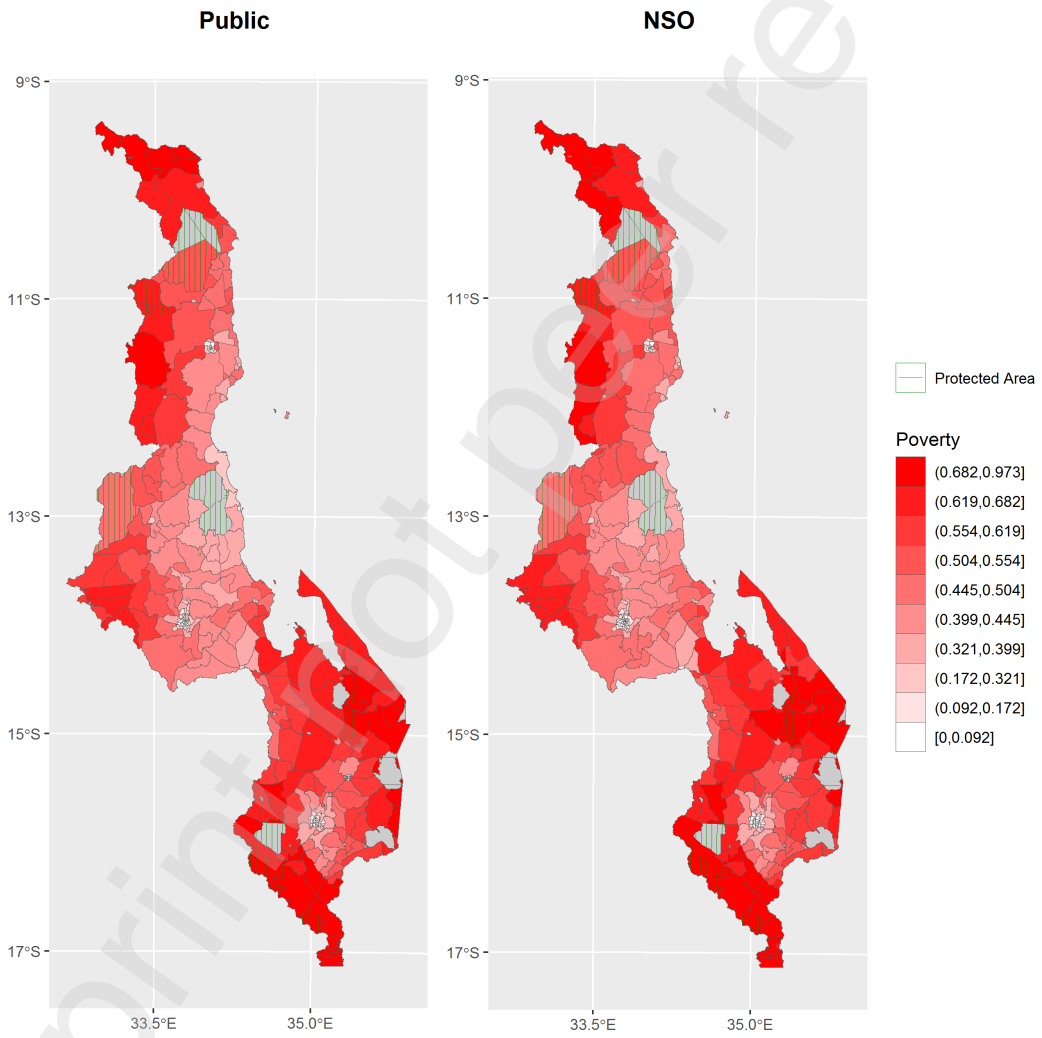

Figure 6: BP (Public) (left) and BP (NSO) (right) estimates of poverty at the TA level. National parks and game reserves are masked as "protected areas"

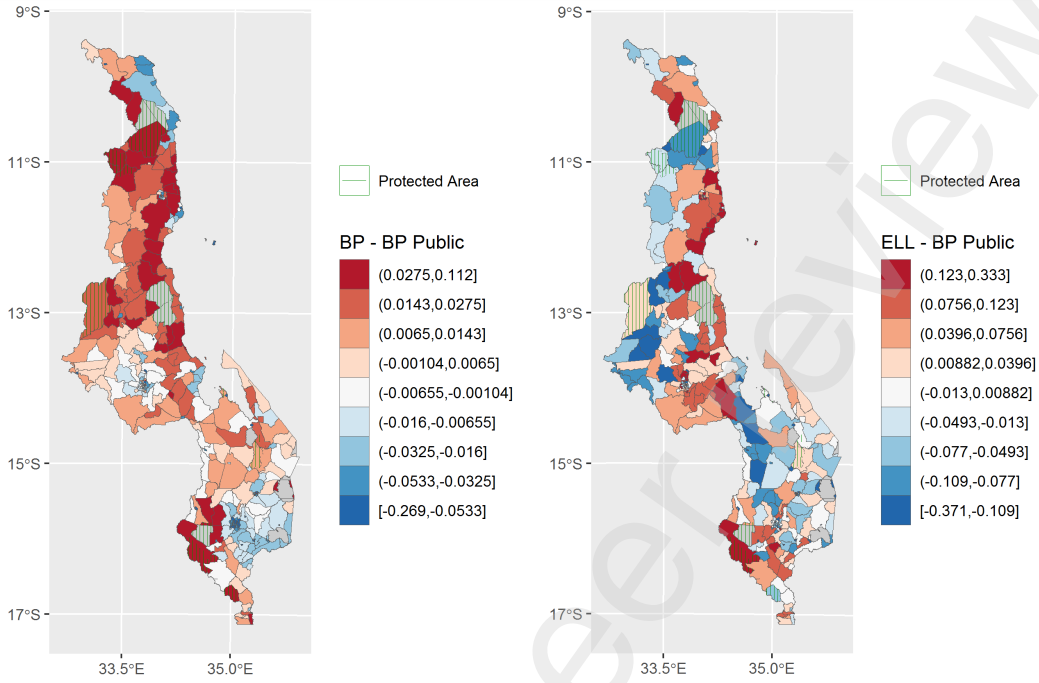

Figure 7: Spatial distribution of differences between BP (NSO) and BP (Public) (left) and between ELL and BP (public) estimates of poverty (right) at the TA-level. National parks and game reserves are masked as "protected areas".

We observe larger differences when comparing standard errors (SEs) of the small area poverty estimates obtained with public versus private NSO coordinates, see Table 7 which reports selected summary statistics of the respective SEs at the different administrative levels. Lower SEs are obtained when the public coordinates are used to estimate the spatial regression model, notably for poverty estimates at the more aggregate levels. At first glance it seems counter-intuitive that the use of randomly scrambled coordinates (when compared to using exact coordinates) yields more precise estimates. This observation is consistent, however, with the fact that scrambling destroys part of the spatial correlation structure that is present in the data (and underlying errors). An under-estimation of the spatial correlation structure in errors will lead to an over-

estimation of statistical precision, meaning that the estimated SEs obtained using the private NSO coordinates will arguably provide more accurate estimates of precision.

	BP (NSO)	BP (Public)	OLS (Public)
EA	0.155	0.165	0.186
TA	0.039	0.0.038	0.036
Region	0.0245	0.0199	0.011
National	0.0230	0.0178	0.0075

Table 7: Mean standard errors (SEs) of poverty estimates from BP (NSO), BP (Public), and OLS for different levels of aggregation.

Provided that the spatial correlation structure is only partially destroyed by scrambling the coordinates, this explanation would predict that SEs obtained with public coordinates would be higher than SEs obtained using OLS which would fully ignore any spatial correlation structure in the errors, yet lower than SEs obtained with the private NSO coordinates (where the spatial correlation structure is fully accounted for). Fully ignoring the spatial correlation structure, as the OLS model does, would lead to a comparatively larger downward bias in SEs at more aggregate levels (as errors are incorrectly assumed to average out) – relative to SEs obtained with the spatial regression models. Both predictions are born out, as can be seen in Table 7 (which also includes estimates of SEs obtained with OLS). Standard statistics measuring the degree of spatial correlation found in the residuals are reported in Table 5 (see estimates for ρ), which confirm that the errors from the regression model with public coordinates exhibit a weaker spatial correlation structure.

5 Concluding remarks

This study investigates whether an accurate poverty map could be obtained by combining remote sensing data that are available in the public domain with household consumption survey data. This approach to poverty mapping takes advantage of the fact that geo-referencing survey locations (i.e. Primary Sampling Units; PSUs) have become standard in recent years and that remote-sensing data from which predictors of poverty can be derived have become widely available. It should be noted, however, that

the publicly available geo-coordinates of survey locations are generally subjected to a random offset (i.e. are scrambled) in order to protect the anonymity of survey respondents. A second objective of this study therefore is to investigate whether these random offsets undermine the ability of this approach to produce accurate poverty maps.

Assessing the success of this alternative approach to producing a poverty map will depend on how success is measured. When the ability to capture the geography of poverty is used as the metric of success, our analysis suggests that estimating small area poverty using remote sensing data offers a promising alternative to building a poverty map using population census data. In our application to Malawi the two approaches are found to produce maps that show nearly identical geographies of poverty. When the objective is to identify poverty in a specific district or set of districts of the country, however, we find that poverty estimates obtained using remote sensing data can deviate substantially from the benchmark estimates obtained using population census data. In summary, the value of poverty maps built using remote sensing data alone will depend on the needs of the decision makers.

On the empirical question of whether randomly off-setting survey coordinates (to protect anonymity of survey respondents) may undermine the accuracy of a poverty map derived from remote-sensing data, our findings echo the comparison with census-based estimates. The random scrambling of the geographic coordinates does not meaningfully alter the estimated geography of poverty, yet may have meaningful impact on estimates for individual districts and may furthermore lead to an over-estimation of statistical precision. The latter observation arguably stems from the fact that the random offsets partially destroys the spatial correlation structure in the data and error term.

Our validation study confirms that for many practical purposes accurate poverty maps can be obtained in countries where population census data are not available. Given that remote sensing data are globally available and are typically updated on an ongoing basis, this also opens the door to updating poverty maps on a higher frequency – whenever new household survey data become available. On a practical note, since the remote-sensing-based models are estimated at the survey PSU level (rather than household level), with the area-level poverty rate as dependent variable (rather than household income or consumption), experimenting with alternative poverty measures or alternative poverty lines would require returning repeatedly to the basic modeling stages of the analysis. This is not the case in the census-based approach, where the model is estimated at the household level. In this sense the remote-sensing-based approach can become rather burdensome.

It remains to be verified whether the remote-sensing-based approach is equally successful in heavily urbanized settings. Small, though not necessarily low-population,

administrative units found in cities may be more difficult to model, particularly when wealthy and poor neighborhoods are in close proximity to each other. Differences across such neighborhoods are conceivably harder to capture using the satellite images that are currently available in the public domain. On a positive note, the technology underlying satellite imagery data is rapidly evolving, and with it the artificial intelligence algorithms that derive usable data from the images (see e.g. Jean et al., 2016; Burke et al., 2021). In conclusion, validating the approach in different locations will be important. Now that we established that the approach works well in a country like Malawi, it remains to be verified whether it is equally successful in countries with different levels of development, different levels of urbanization and different degrees of population density.

Appendix 1: Variance of the prediction error

Introduction

Consider a simple data generating mechanism:

$$y = f(x, \theta) + \varepsilon,$$

with the assumptions that ε and x are independent, realizations of the DGP are independent, and $E\varepsilon = 0, E\varepsilon^2 = \sigma_\varepsilon^2$. We want to estimate the precision of predicting y conditional on x, using parameter values $\check{\theta}$ and $\check{\sigma}^2$ for θ and σ_ε^2, consistently estimated from an appropriate sample of (y, x) observations. If (y_0, x_0) is an out-of-sample observation and y_0 is predicted by $\check{y}_0 = f(x_0, \check{\theta})$ the prediction error is:

$$y_0 - \check{y}_0 = f(x_0, \theta) - f(x_0, \check{\theta}) + \varepsilon_0.$$

The prediction error consists of two parts, which we will term *model error* $f(x_0, \theta) - f(x_0, \check{\theta})$ and *idiosyncratic error* ε_0. Under the model assumptions (and assuming a typical sampling scheme for the sample of (y, x) observations on which the estimator for θ is based) the two error components are independent and a consistent estimator for the mean square prediction error is:

$$\text{MSE}(\check{y}_0) = V(f(x_0, \check{\theta})) + \check{\sigma}^2 I,$$

where:

$$V(f(x_0, \check{\theta})) = E(f(x_0, \theta) - f(x_0, \check{\theta}))^2$$

and where the latter expectation is based on the (asymptotic) distribution of the $\check{\theta}$ estimator (typically using the 'delta method' if f is nonlinear).

Prediction errors in the SEM model

Denote by q_S and q_O the vectors of poverty rates in survey locations and non-survey locations respectively. Predictions of q_O are based on out-of-sample application of the SEM model:

$$q = X\theta + u,$$

where we have imperfect observations $\hat{q}_S$ of q_S, due to sampling variation:

$$\hat{q}_S = X_S\theta + u_S + e_S.$$

With this data generating process the best linear unbiased prediction of q is:

$$E(q|X, \hat{q}_S) = X\theta + E(u|u_S + e_S)$$

for both in-sample and out-of-sample locations, and $u_S = q_S - X_S\theta$. Let the idiosyncratic part of prediction error given X and $\hat{q}_S$ be denoted by r:

$$r = q - E(q|X, \hat{q}_S),$$

with variance matrix Err^T.

An empirical implementation involves sampling variation of estimators for θ and hence estimates of the residual $u_S + e_S = \hat{q}_S - X_S\theta$. Let $\check{\theta}$ denote the maximum-likelihood estimator for θ. Then:

$$\check{q} = X\check{\theta} + E(u|u_S + e_S = \hat{q}_S - X_S\check{\theta}).$$

With normally distributed residuals (u_S, e_S, u_O), $E(u|u_S + e_S)$ is linear in $u_S + e_S$. Because $\hat{q}_S - X_S\check{\theta}$ and $\check{\theta}$ are asymptotically independent for consistent and asymptotically normal estimators of θ, the sampling variance of $\check{q}$, given X and $\hat{q}_S$ can be estimated by:

$$V(\check{q}|X, \hat{q}_S) = XV(\check{\theta})X^T + \check{Z},$$

where $\check{Z}$ denotes the sampling variance of $\check{u} = E(u|u_S + e_S = \hat{q}_S - X_S\check{\theta})$. Combining the sources of idiosyncratic and model error, the total covariance matrix H of prediction error $q - \check{q}$ is estimated by:

$$H = XV(\check{\theta})X^T + \check{Z} + Err^T.$$

Below, these three components will be discussed in turn.

Estimation

Estimating $V(\check{\theta}, \rho, \sigma^2, \tau)$

The SEM model is estimated by maximum likelihood (ML), assuming normally distributed error terms throughout. Under these assumptions, and given the variance parameters ρ and σ_u^2 as well as exogenous $V(e_S)$, ML estimation of θ amounts to GLS. Moreover, asymptotically the full covariance matrix of the full parameter estimator is block diagonal in the components relating to ρ, σ_u^2 and θ respectively. In practice we concentrate the likelihood function on (ρ, σ_u^2), estimating the variance parameters and

their covariance matrix first. The estimator for θ and its covariance matrix then follows from GLS.

Estimating $\check{Z} = V(\check{u}_O)$

Given the residual process $u = \sigma(I - \rho W)^{-1}\varepsilon$, we have:

$$\Sigma = Euu^T = \sigma^2 CC^T,$$

where $C = (I - \rho W)^{-1}$. Recall that for $i \neq j$, $W_{ij} = T_{ij}^{-\tau}$ with T a (scaled) symmetric matrix of distances between locations i and j, and $W_{ii} = 0$. Partition the variance matrix Σ according to in-sample locations (S) and out-of-sample locations (O) so that we can write:

$$\Sigma = \begin{pmatrix} \Sigma_{SS} & \Sigma_{SO} \\ \Sigma_{OS} & \Sigma_{OO} \end{pmatrix}.$$

Also, let $\Sigma_{.S}$ denote the first 'column' of this partition:

$$\Sigma_{.S} = \begin{pmatrix} \Sigma_{SS} \\ \Sigma_{OS} \end{pmatrix}.$$

Assuming normally distributed ε and survey error e_S, it follows that:

$$E(u|u_S + e_S) = \Sigma_{.S}[\Sigma_{SS} + D_S]^{-1}(u_S + e_S),$$

where D_S is the variance matrix of e_S, established separately, based on the survey design, and taken as a given matrix of parameters.[12] The empirical counterpart is:

$$\check{u} = \check{\Sigma}_{.S}[\check{\Sigma}_{SS} + D_S]^{-1}(\hat{q}_S - X_S\check{\theta}),$$

and is subject to model error – the topic of this sub-section.

We approximate the MSE of $\check{u}$ using the Delta method. Let operator $d()$ indicate taking (infinitesimal) deviations from the true parameter values. Then:

$$d(\check{u}) = d(\check{\Sigma}_{.S}[\check{\Sigma}_{SS} + D_S]^{-1})(\hat{q}_S - X_S\check{\theta}) + \check{\Sigma}_{.S}[\check{\Sigma}_{SS} + D_S]^{-1}d(\hat{q}_S - X_S\check{\theta}).$$

To simplify notation, write $A = \Sigma_{.S}[\Sigma_{SS} + D_S]^{-1}$ and $b = \hat{q}_S - X_S\theta$, with obvious derived meaning of $\check{A}, \check{b}$. Then, denoting actual deviations from true parameter values

[12] Under the model assumptions D_S stems from variation in X_S and σ_ε^2. Alternatively, it could therefore be argued that D_S is increasing in σ^2. We ignore this complication here.

by Δ, we get:

$$\check{Z} \approx E\Delta(\check{A})bb^T\Delta(\check{A})^T + EA\Delta(\check{b})\Delta(\check{b})^T A^T + E\Delta(\check{A})b\Delta(\check{b})^T A^T + EA\Delta(\check{b})b^T\Delta(\check{A})^T. \quad (4)$$

In the expression above, symbols without accent indicate 'true' values, i.e, constants with respect to the expectation operator. In actual computation they are replaced by estimated values. The two last terms of the equation vanish asymptotically, because $\Delta(\check{A})$ and $\Delta(\check{b})$ are asymptotically independent with zero expectation.[13]

The second term of equation (4) evaluates to:

$$EA\Delta(\check{b})\Delta(\check{b})^T A^T = AX_S E\Delta(\check{\theta})\Delta(\check{\theta})^T X_S^T A^T \approx \check{A}X_S V(\check{\theta})X_S^T \check{A}^T.$$

For the first term of equation (4) we need to work out $d(\check{A})$. Note that:

$$A = \frac{1}{\sigma^2}\Sigma_{\cdot S}[\frac{1}{\sigma^2}\Sigma_{SS} + \frac{1}{\sigma^2}D_S]^{-1}$$

and that Σ/σ^2 does not depend on σ^2. In what follows Σ^* and D_S^* will refer to Σ/σ^2 and D_S/σ^2. Also, when there is no danger of confusion, we will omit accents from symbols. It follows that:

$$d(A) = d(\Sigma_{\cdot S}^*[\Sigma_{SS}^* + D_S^*]^{-1}) = d(\Sigma_{\cdot S}^*)[\Sigma_{SS}^* + D_S^*]^{-1} + \Sigma_{\cdot S}^* d([\Sigma_{SS}^* + D_S^*]^{-1}),$$

where: $d(\Sigma_{\cdot S}^*)$ is the $\cdot S$ sub-matrix of:

$$d\Sigma^* = d\left((I - \rho W)^{-2}\right).$$

Likewise,

$$d([\Sigma_{SS}^* + D_S^*]^{-1}) = -[\Sigma_{SS}^* + D_S^*]^{-1}[d(\Sigma_{SS}^*) + dD_S^*][\Sigma_{SS}^* + D_S^*]^{-1},$$

with $d(\Sigma_{SS}^*)$ the SS sub-matrix of $d(\Sigma^*)$ given above. To evaluate these expressions, it

[13] Independence follows from (i) the normality assumptions; (ii) linearity of $\Delta(\check{A})$ in $(\Delta(\check{\rho}), \Delta(\check{\sigma}^2))$ and linearity of $\Delta(\check{b})$ in $\Delta(\check{\theta})$; (iii) the information matrix $I(\theta, \rho, \sigma^2, \tau)$ of the data generating process is block diagonal, $I(\theta, \rho, \sigma^2, \tau) = \text{diag}(I(\theta); I(\rho, \sigma^2, \tau))$. See e.g., Mardia, K. V.; Marshall, R. J. (1984). "Maximum likelihood estimation of models for residual covariance in spatial regression". Biometrika. 71 (1): 135–46.

can be verified that:

$$\begin{aligned}\frac{\partial}{\partial\sigma^2}D_S^* &= -\sigma^{-4}D_S = -\sigma^{-2}D_S^* \\ \frac{\partial}{\partial\rho}\Sigma^* &= (CWC)C^T + C(CWC)^T,\end{aligned}$$

where $C = (I - \rho W)^{-1}$.

Then we have:

$$\begin{aligned}\Delta\Sigma^* &= \left(\frac{\partial}{\partial\rho}\Sigma^*\right)\Delta\rho \\ \Delta D_S^* &= -\sigma^{-2}D_S^*\,\Delta\sigma^2,\end{aligned}$$

so that:

$$\begin{aligned}(\Delta A)b &= \left(\frac{\partial}{\partial\rho}\Sigma_{\cdot S}^*\right)[\Sigma_{SS}^* + D_S^*]^{-1}b\;\Delta\rho \\ &\quad -\Sigma_{\cdot S}^*[\Sigma_{SS}^* + D_S^*]^{-1}[\left(\frac{\partial}{\partial\rho}\Sigma_{SS}^*\right)\Delta\rho - \sigma^{-2}D_S^*\;\Delta\sigma^2][\Sigma_{SS}^* + D_S^*]^{-1}b \\ &= \left(\frac{\partial}{\partial\rho}\Sigma_{\cdot S}^* - \Sigma_{\cdot S}^*[\Sigma_{SS}^* + D_S^*]^{-1}\frac{\partial}{\partial\rho}\Sigma_{SS}^*\right)[\Sigma_{SS}^* + D_S^*]^{-1}b\Delta\rho \\ &\quad +\left(\Sigma_{\cdot S}^*[\Sigma_{SS}^* + D_S^*]^{-1}\sigma^{-2}D_S^*\right)[\Sigma_{SS}^* + D_S^*]^{-1}b\Delta\sigma^2.\end{aligned}$$

Finally, when working out $E\left(\Delta A\right)bb^T\Delta A^T$, the (co)variance between $\Delta\rho$ and $\Delta\sigma^2$ is taken from the corresponding components of the covariance matrix of the estimators for the model's variance parameters.

Estimating $E[r_o r_O^T]$

Finally, using $r = u - E(u|X, \hat{q}_S)$, the idiosyncratic variance Err^T is consistently estimated by:

$$V(r) = \check{\Sigma} - \check{\Sigma}_{\cdot S}[\check{\Sigma}_{SS} + D_S]^{-1}\check{\Sigma}_{S\cdot}.$$

Appendix 2: Source of Remote Sensing Variables

We describe below the data sources and provide an overview of the specific variables that are constructed for the remote-sensing based analysis in this paper.

Initial conditions:

The capital is distinct from other areas due to its centrality of political decisions and public disbursement. We measure the bird-flight (Euclidean) distance from the center

point of each administrative area to the coordinates of the capital (from the Global Insights (v 6.1) data), which we call *dist2cap*. Likewise, we calculate distance from the center point of each administrative area to the border (*dist2border*) using the boundary file provided by the NSO.

In consideration of the initial conditions of administrative areas for regional trade, we construct distance to major rivers (*dist2major_river*) and distance to coast (*dist2coast*). Also here we use the center point of the administrative areas as the points of origin and destination. Major rivers and coastline data are from Natural Earth[14].

In addition, we consider the terrain with regards to the elevation, slope and ruggedness. The elevation data are from the Shuttle Radar Topography Mission (SRTM) Digital Elevation dataset (version 4) and is measured in meters at 3 arc-seconds and 30 arc-seconds (approximately 1km^2) (Jarvis et al. 2008). We summarize mean and max: *elev* and *elev_max30*. We derive elevation in populated places using a 1% threshold level, which is approximately 10 people per km2 from the Landscan 2012 dataset (described in the population count and density section). We take advantage of the slope function on Google Earth to process the 3 arc-seconds elevation data into a slope variable: *slope*.[15] Figure A2.1 (left panel) illustrates the variation of slope, where the Great Rift Valley is seen to run from North to South. The Ruggedness Index we use is developed by Nunn and Puga (2012) which measures the level of ruggedness in hundreds of meters of elevation difference for each grid cell. We summarize the pixel mean (*rugged*) and maximum (*rugged_max*) of this ruggedness index evaluated over the entire administrative areas as well as over populated areas (*rugged_popmask* and *rugged_max_popmask*). Populated areas are defined as areas with any population in the Landscan 2012 population model (described in population count and density section below). Figure A2.1 (right panel) shows the spatial variation of the log of *rugged*.

Agriculture:

Agriculture and arable land can indicate higher economic growthpotential, or conversely, imply a lack of urbanization and infrastructure. Agricultural activities provide evidence of economic activity, which in turn may indicate a presence of water infrastructure. In order to measure agricultural land-use we use the Global Hybrid dataset (0611-2012 V2) produced by Fritz et al. (2015), which estimates the percentage share of land used for agriculture within a one square kilometer pixel. Expert assessment of five existing global land cover products along with national and subnational crop statistics as inputs provides the likelihood of a pixel indicating agricultural land use. By mul-

[14]Large scale (1:10 million) vector data are available for download from: http://www.naturalearthdata.com/downloads/

[15]Similarly, Verdin et al. (2007) derive slope measures at 30-arc-seconds derived from the 3-arc-seconds SRTM data.

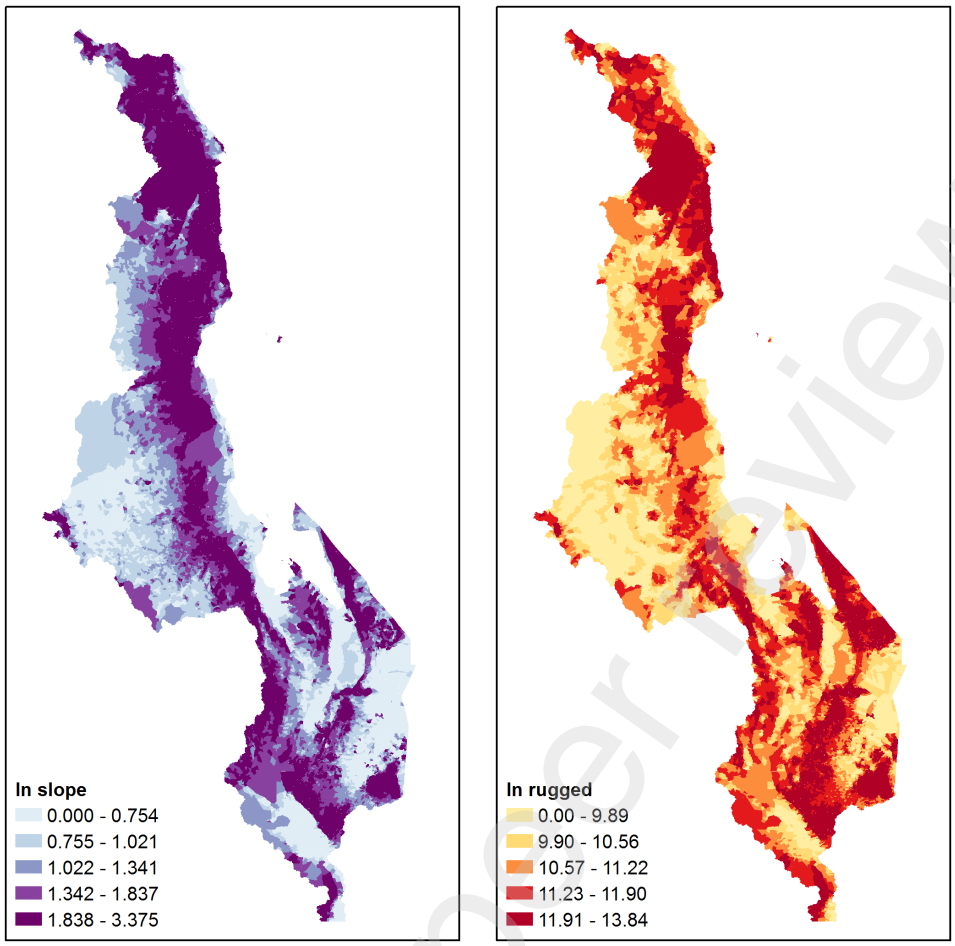

Figure 8: Map of ln slope (left panel) and map of ln ruggedness (right panel)

tiplying this percentage of agricultural likelihood by the total pixel area, we obtain a measure of total agricultural land use area in a pixel. We then aggregate these to obtain agricultural land-use data at the administrative level. The total share of agricultural land-use within the total administrative land area is labeled *sh_ag_land*. Furthermore, we consider irrigated and rainfed crop area data from GFSAD1000 V1.0, which are derived from a variety of sources including: multi-sensor remote sensing data, secondary data and field-plot data (Teluguntla et al. 2015). We process these data in Google Earth Engine and divide by the area to obtain: *rain_crop_sh* and *irrigated_crop_sh* (see Figure A2.2).

Measures of greenness consider the intensity or density of vegetation within a district over time, which identifies areas of vegetation including agricultural land and forest cover. Greenness may be affected by land fertility, irrigation, precipitation and climate. The most common measure of greenness is the Normalized Differentiated Vegetation Index (NDVI), which is derived from remote sensing data. NDVI values range between

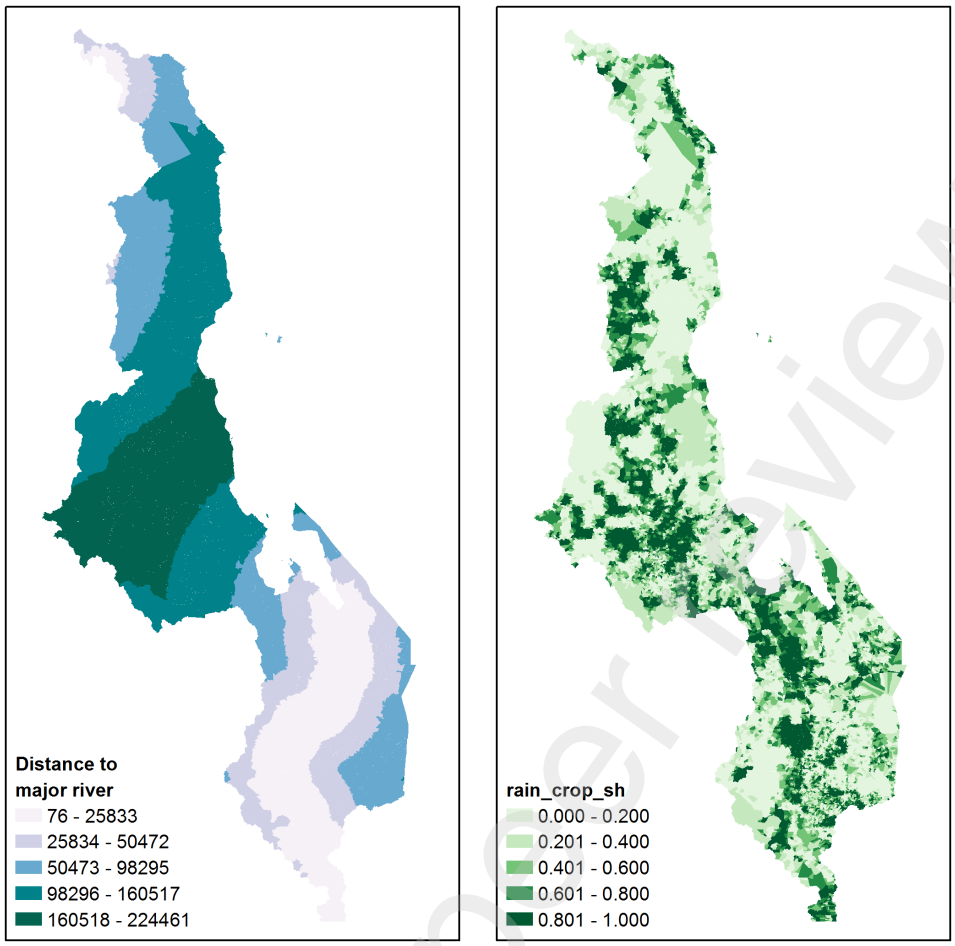

Figure 9: Map of distance to major river (left panel) and map of share of rainfed crop (right panel)

-1 (indicating complete absence of vegetation) and 1 (indicating the greatest intensity of vegetation). We examine two sources of NDVI data. The first source is hosted by Google Earth Engine at an annual level, which is the result of a composite of L1T orthorectified scenes from Landsat 7 using the computed top-of-atmosphere (TOA) reflectance. We derive five measures that summarize the data for 2008 at the administrative area level: minimum (*ndvi_min*), mean (*ndvi_mean*), median (*ndvi_median*), standard deviation (*ndvi_sd*), and maximum (*ndvi_max*). Figure A2.3 (left panel) illustrates the greenness across Malawi with high values in the North near Lake Malawi and in the South in the Shire River catchment. The second source of NDVI data is provided by the U.S. National Aeronautics and Space Administration (NASA) Global Inventory Modeling and Mapping Studies (GIMMS v.3) at a bi-monthly frequency. It measures greenness over 8 square kilometers pixels (Zhu et al. 2013).[16] Similarly, we derive five

[16]Jim Tucker (NASA) kindly provided us the data.

measures that summarize the data for 2010 at the administrative area level: minimum (*ndvi_min*), mean (*ndvi_mean*), median (*ndvi_median*), standard deviation (*ndvi_sd*), and maximum (*ndvi_max*). We also consider a similar greenness measure EVI. We use the EVI MODIS BRDF-corrected imagery (MCD43B4) monthly product available on Google Earth Engine as a result of a gap-filled method by Weiss et al. (2014). We calculate the monthly mean for 2008 and summarize it at the administrative level.

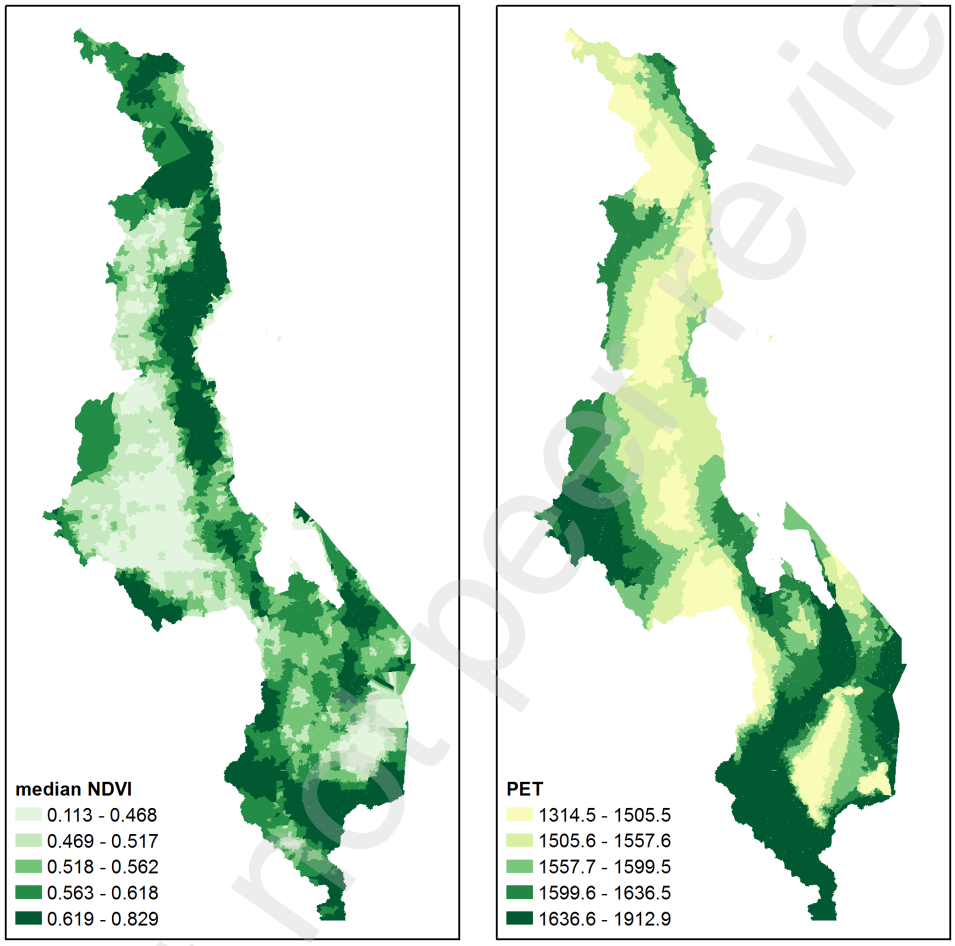

Figure 10: Map of median NDVI (left panel) and map of PET (right panel)

The Food and Agricultural Organization (FAO) and IIASA extract agricultural data from the Global Agricultural Ecological Zones (GAEZ v3.0) (Fischer et al. 2012). The database includes estimates of yield for various crops (e.g. cotton), inputs, water sources (rain-fed or irrigated). It also includes aggregated information on Net Primary Production, total crop production value per hectare, and total crop production value. The FAO also produces the Food Insecurity, Poverty and Environment Global GIS database (FGGD). The FGGD 2005 data includes a suitability index for rain-fed crops

measured at a spatial resolution of 5 arc-minutes (10 km) (Van Velthuizen et al. 2007).[17]

We also include the production of plant biomass using the measurement of Net Primary Productivity (NPP), which is the result of all the carbon used in photosynthesis minus the loss of carbon from the maintenance respiration. Google Earth Engine provides the cloud-corrected MODIS product (MOD17A3) data at 1km spatial resolution summarized at the annual level from the 45 8-day Net Photosynthesis (PSN) products. Evaluating the mean level at the administrative level yields the variable *npp*.

Naidoo and Iwamura (2007) derive a global map of gross economic rents from agricultural lands (in 2000 US dollars per hectare per year) by integrating spatial information on crop productivity, livestock density, and producer prices from FAOSTAT (http://faostat.fao.org; accessed June 2005). We extract these economic rents by pixel and summarize the mean over the administrative unit (*agr_opp_cst*).

Climate:

Climate provides the initial conditions to agricultural production and human settlement. Increasing change in climate will arguably affect conditions to support agricultural production. In a low production scenario, Hertel et al. (2010) find that the prices for major staples rise 10–60% by 2030, which would have heterogenous impacts on the poor depending on local changes in agricultural production and the sources of income for households in the area. In the case of drought and urban floods, Winsemius et al. (2015) find evidence that the poor already are overexposed to natural hazards of floods and droughts and are vulnerable to changes in climate.

BioClim is a global gridded database at 30 arc-second (1km2) spatial resolution of bioclimatic variables that are derived from monthly temperature and rainfall values from cleaned weather station data and elevation. The bio-physical model interpolates these meteorological variables and provides derivative variables of seasonality over the base period of 1960 to 1991 (Hijmans et al. 2005)[18]. We extract these 1km2 pixels for annual mean temperature (*bio1* in degrees Celsius * 10), temperature seasonality (*bio4* in standard deviation * 100), annual precipitation (*bio12* in mm * 10), precipitation seasonality (*bio15* in coefficient of variation * 10), precipitation of wettest quarter (*bio16* in mm * 10) and precipitation in driest quarter (*bio17* in mm). We aggregate the data to obtain mean values for each administrative area. In Figure A2.4 (left panel), we see the effect of elevation in the higher temperature of the Great Rift Valley compared to higher elevation areas. Figure A2.4 (right panel) shows a higher level of

[17]We modified negative values that represent pixels classified as urban, closed forest or irrigated to equal zero for the mean calculations by administrative area.

[18]For the interpolation of noisy multi-variate meteorological data, the authors use thin plate smoothing splines from the ANUSPLIN program with latitude, longitude, and elevation as independent variables.

temperature seasonality in the South compared to the North. We also use measures from a daytime and nighttime product that is derived from MODIS land surface temperature data (MOD11A2) and a gap-filling algorithm by Weiss et al. (2014): *oxmap_ngtlst* and *oxmap_ngtlst*. The difference between the nighttime and daytime land surface temperature is stored in the variable *delta_lst08*.

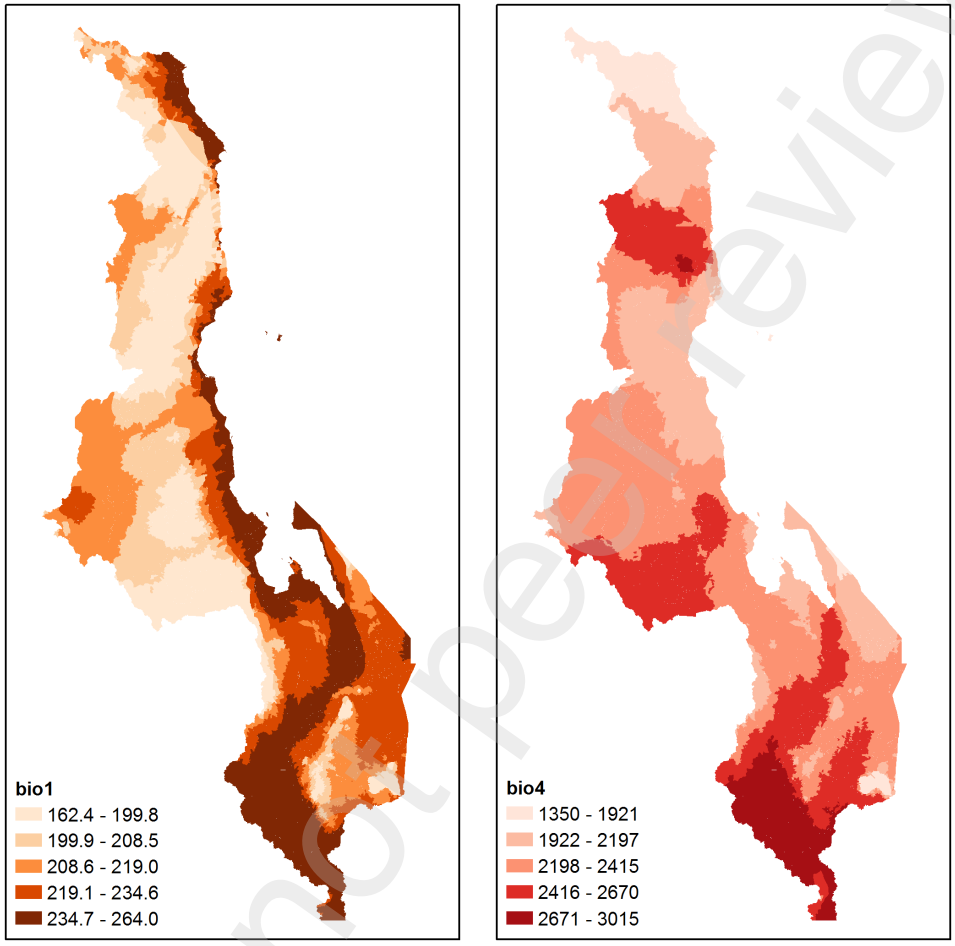

Figure 11: Map of annual mean temperature (bio1) (left panel) and map of temperature seasonality (bio4) (right panel)

The Extraterrestrial Solar Radiation (ET_solrad) monthly data (Zomer et al. 2008) provide a measure of evaporation.[19] We summarize mean values (at the administrative area level) for the months of January, April, July and October. Likewise, from the same website, we make use of annual average Potential Evapotranspiration data that measures the amount of evaporation that would occur given sufficient water resources

[19]ETRA (mm/month) data are available from the International Food Policy Research Institute (IFPRI) website http://www.cgiar-csi.org/data/global-aridity-and-pet-database (accessed 2017-03-06) and see Global Geospatial Potential EvapoTranspiration & Aridity Index: Methodology and Dataset Description for additional details.

over the period 1950 to 2000. This variable is labeled *pet* and plotted in Figure A2.3 (right panel).

In addition, we extract an aridity index and potential evapo-transpiration from models created by Trabucco and Zomer (2009). The models use data from WorldClim (Hijmans et al. 2005) as input parameters. The global mean Aridity index is Mean Annual Precipitation divided by Mean Annual Potential Evapo-Transpiration for the period from 1950 to 2000. High values of this index represent humid conditions, while low values represent arid conditions. A generalized climate classification scheme by UNEP (1997) suggests: hyper arid and arid are values less than 0.2, semi-arid has values 0.2-0.5 and dry sub-humid and humid have values above 0.5.[20]

Environment:

The environment provides context to human settlements and surrounding economic activities. Land cover provides important information on the extent and type of human activity. We use the MODIS land cover products to select the crop and urban areas from the Annual International Geosphere-Biosphere Programme (IGBP) classification. These data area are the result of both supervised classifications of MODIS Terra and Aqua reflectance data and subsequent post-processing refinements for specific classes from prior knowledge and ancillary information. We select the crop and urban classes: *crop* and *urb*. The forest data are from Hansen et al. (2013) version 1.6 that provide both the stock of the forest from 2000, which is defined as "Tree canopy cover for year 2000, defined as canopy closure for all vegetation taller than 5m in height", and forest loss, which is defined as "a stand-replacement disturbance (a change from a forest to non-forest state)": *forloss* and *forest*.

To measure pollution, we use the fine particulate matter (PM 2.5) of air pollution data estimated from a model that excludes dust and sea-salt particles (van Donkelaar et al., 2015). The model uses geographically weighted regressions along with monitoring and satellite data, including: aerosol composition and land use information. We obtain the mean (*pm25*), sum (*pm25_sum)* and standard deviation (*pm25_sd*) for each administrative unit.

Socio-economic: Night time lights

Night time lights (NTL) data, the visible light emitted from earth at night captured by satellites, are found to correlate strongly with population density and economic activity (e.g. Henderson et al., 2012). The NTL high gain data come from the Defense Meteorological Satellite Program DMSP-OLS, which are made publicly available by NOAA[21]. Specifically, we use the radiance calibrated data that is available for the years

[20]See Table 2 in CGIAR-CSI *Global Aridity Index (Global-Aridity)* and *Global Potential Evapo-Transpiration (Global-PET)* Climate Database (2009)

[21]NOAA DMPS-OLS has another night time light data product that provides a radiance correction

1996, 2005 and 2010. The variables used in our study sum all of the NTL that is emitted within an administrative level. Although the 1996 and 2010 night time lights are from different sensors, we observe an increase in the spatial extent of light areas (Figure A2.5).

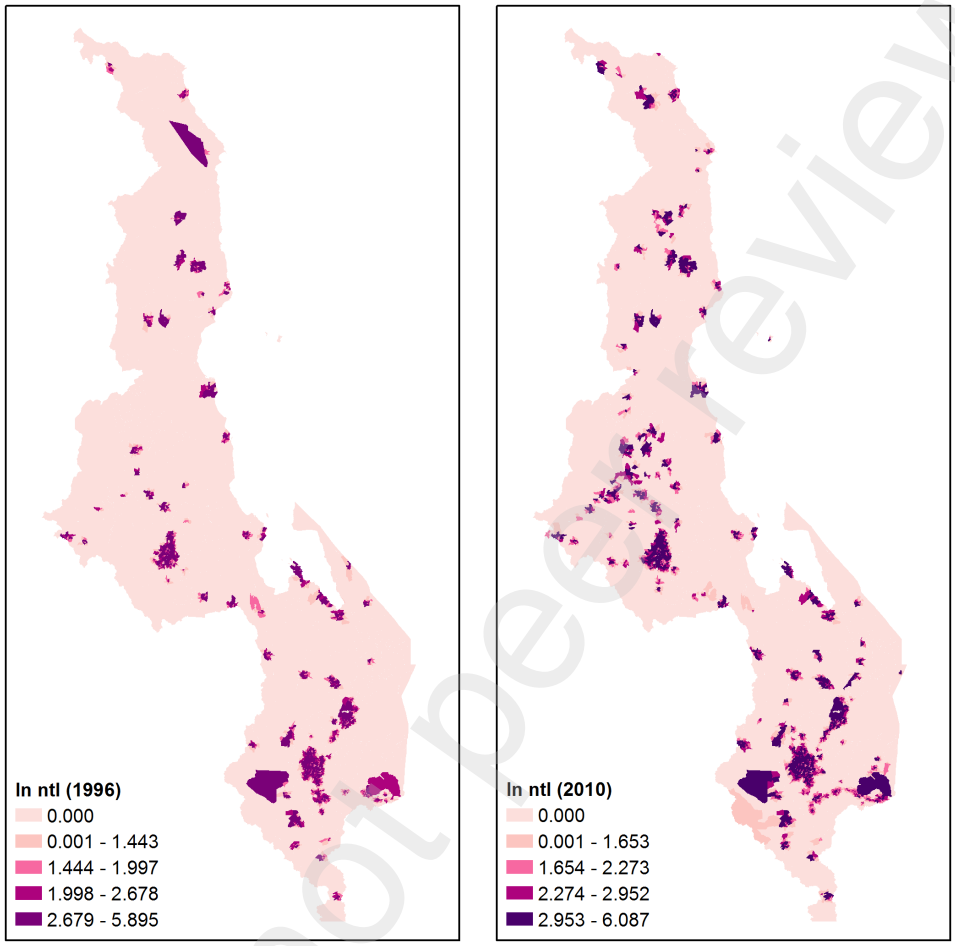

Figure 12: Map of ln night time lights (1996) (left panel) and map of ln night time lights (2010) (right panel)

Socio-economic: Population count and density

Because of the relative efficiency in the construction of infrastructure to support service delivery in higher population dense urban areas compared to rural areas, urban areas play an important role in providing service delivery, which in turn influences poverty. Different measures of urbanity have been proposed in the literature (e.g Dijkstra and Poelman, 2014; Roberts et al., 2017). We take advantage of geospatial sub-

and corrects for the top-coding problem, which is present in the high gain night time lights data. Currently, the low gain data are not available at a consistent annual basis necessary for this time-series analysis. Image and data processing by NOAA's National Geophysical Data Center. DMSP data collected by US Air Force Weather Agency.

national gridded global data to measure travel time to major cities, the agglomeration index, population count, and built-up area.

We construct population density and totals from two main sources of global model population distribution data: Landscan (e.g. Bright et al., 2013) and GHS-Pop (Freire et al., 2016).[22] Landscan is a model that uses dasymetric and spatial modeling methods to estimate population. The model takes advantage of multiple sources of spatial microdata including land cover, road, slopes and urban areas data, and population census data. Figure A2.6 illustrates the spatial distribution of the total population across the enumeration area.

GHSpop data are derived from a raster-based dasymetric model estimating global population. Using newly processed LandSat data to provide multiple periods in time (1975, 1990, 2000 and 2014), this model leverages newly available built-up data (GHS-BUILT) to define the locations where the GHSpop allocates population. The estimates of population are from the highest number of spatial units available to account for the total population known across subnational areas. We construct the following variables of total population within the administrative unit: *GHS_poptot_1975*, *GHS_poptot_1990*, *GHS_poptot_2000*, and *GHS_poptot_2014*. GHS-BUILT at 300m spatial resolution is the source of built-up area within each administrative unit from which we derive the variables: *GHSL_built_km2_1975*, *GHSL_built_km2_1990*, *GHSL_built_km2_2000*, and *GHSL_built_km2_2014*.

The share of urban population within an administrative unit is constructed as follows. We first define urban population as the number of people within a given threshold of density from the 2014 GHS population grid (*ghspop14*) and the Landscan 2012 population grid (*lpop12*). Next we use the World Urban Prospects estimates of the urban share at the country level provided by the World Bank World Development Indicators to inform a population density threshold that sums to the estimated urban share at the country level[23]. Finally, we construct the share of urban population that meet these density criteria within the administrative unit as the urban share of total population (*urban_lpop12*).

Socio-economic: Travel time and accessibility

Transport infrastructure enables improved economic growth by lowering costs of moving goods and people. Nelson (2008) derives the travel time from all land area to major cities, which are defined by a city size of 50,000 people or greater (circa 2000). They use global roads data with a travel time model based on road functional class. Remote sensing data are the input to the terrain classification. We extract these data at

[22]For a recent review of gridded population datasets see Leyk et al. (2019).

[23]Another method of country consistent urban rates from World Urbanization Prospects is the iUrban method by Aubrecht et al. (2016).

30 arc-seconds (approximately 1km^2) and label the variable *tt50k2000_mean*. Following the methodology by Uchida and Nelson (2008), Berg et al. (2017) construct a travel time model using circa 2010 data of global roads and land cover (*tt50k2010_mean*). Figure A2.6 plots the mean travel time to major cities which highlights transportation infrastructure network between cities such as Blantyre.

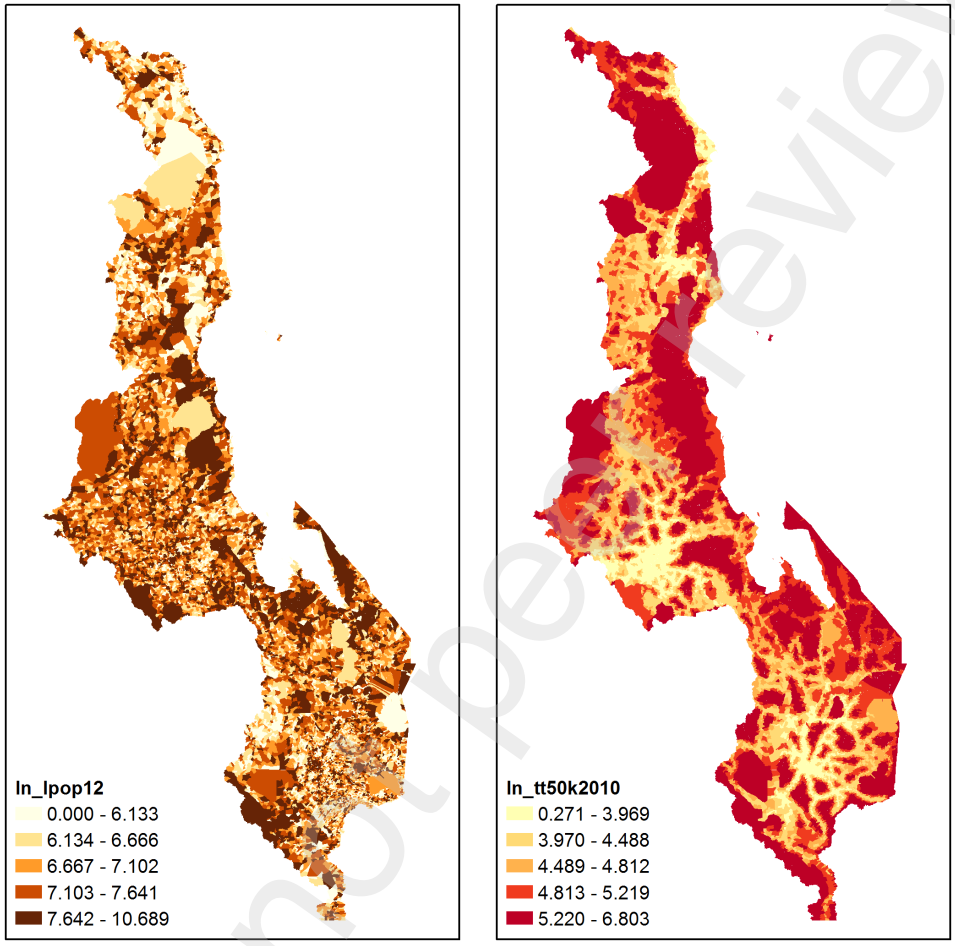

Figure 13: Map of ln population (2012), source: Landscan (left panel) and map of ln travel time to major cities (ca. 2010) (right panel)

Uchida and Nelson (2008) developed an Agglomeration Index (AI) which measures the area of land that satisfies the following criteria: (1) travel time of 60 minutes or less to a city of 50,000 people or greater, and (2) population density of at least 150 people per square kilometer. (They use the model of travel time from Uchida and Nelson (2008) and the population model from Landscan 2004.) We sum the data at the administrative area level and label it *ai50k2000*. Roberts et al. (2017) provide an updated AI circa 2010 using the newer travel time model and the population model from Landscan 2012 (*ai50k2010)*. We construct the variable of the total population in the AI for each period

and, subsequently, the share of population in an AI out of the total population of the administrative area, respectively *ai50k2000_lpop04* and *ai50k2010_lpop12*.

Appendix 3: ELL (log) consumption model

	Coefficient	Std. Err.
Intercept	10.5881	(0.1237)
EA_M_HHD_HAS_BICYCLE	-0.1779	(0.0595)
EA_M_HHD_HAS_CAR	1.2011	(0.2212)
EA_M_HHD_HAS_PHONE	0.2678	(0.0836)
EA_M_HHD_HAS_TV	-0.4604	(0.1471)
EA_M_HOH_RESID_5YR	0.1522	(0.0797)
EA_M_HOH_TR_SENA	-0.3658	(0.0513)
EA_M_HOH_TR_TUMBUKA	0.1242	(0.0555)
EA_R_CHLDBRN_LSTYR_FEM_30_45	0.2403	(0.0688)
HHD_AGE_00_05_ALL_0	0.1749	(0.0199)
HHD_AGE_00_05_ALL_1	0.0722	(0.0155)
HHD_AGE_06_15_ALL_00	0.0715	(0.0207)
HHD_AGE_06_15_ALL_01	0.0649	(0.0167)
HHD_AGE_65_ALL_0	0.0352	(0.0214)
HHD_COOK_WOOD_OTHER_1	-0.1877	(0.0276)
HHD_EDU_SECONDARY_0	-0.0814	(0.0261)
HHD_FLOOR_SOFT_1	-0.1754	(0.0203)
HHD_HAS_BICYCLE_1	0.098	(0.0136)
HHD_HAS_COOKER_1	0.1738	(0.0467)
HHD_HAS_FRIDGE_1	0.2897	(0.0444)
HHD_HAS_ITN_1	0.1469	(0.0126)
HHD_HAS_RADIO_1	0.1934	(0.0134)
HHD_HAS_TV_1	0.2631	(0.027)
HHD_OCCUPIED_06_15_0	0.034	(0.0154)
HHD_ROOF_GRASS_1	-0.1072	(0.0181)
HHD_ROOMS_01	-0.2334	(0.0236)
HHD_ROOMS_02	-0.1608	(0.019)
HHD_ROOMS_03	-0.0815	(0.0182)
HHD_SIZE_01	1.2754	(0.0375)
HHD_SIZE_02	0.7781	(0.0336)

	Coefficient	Std. Err.
HHD_SIZE_03	0.5297	(0.0273)
HHD_SIZE_04	0.3551	(0.0223)
HHD_SIZE_05	0.21	(0.0196)
HHD_SIZE_06	0.1046	(0.0187)
HHD_TOIL_NONE_1	-0.2377	(0.032)
HHD_TOIL_PIT_1	-0.1515	(0.0262)
HHD_WALL_MUDBRCK_1	-0.0376	(0.0146)
HOH_AGE_LN	-0.0597	(0.0236)
HOH_EDU_PRIMARY_1	0.1332	(0.0208)
HOH_EDU_SECONDARY_1	0.1766	(0.0346)
HOH_EDU_SOMEPRIM_1	0.1047	(0.0157)
HOH_EDU_SOMESEC_1	0.2069	(0.023)
HOH_EDU_TERTIARY_1	0.2779	(0.0455)
HOH_OCCUPIED_1	0.117	(0.0194)
HOH_OCC_PRIVATE_1	-0.0605	(0.0209)
HOH_OCC_SELFAGRO_1	-0.0588	(0.0146)
HOH_RELIG_MUSLIM_1	0.0513	(0.0203)
STRATA_1	0.3166	(0.0796)
STRATA_2	0.4951	(0.0722)
STRATA_3	0.5221	(0.0714)
STRATA_4	-0.0796	(0.0465)
STRATA_5	0.2127	(0.0272)
TA_REMOTE100K	-0.017	(0.0098)
_URBAN$HHD_HAS_BICYCLE_11	-0.125	(0.0353)
_URBAN$HHD_HAS_RADIO_11	-0.0983	(0.0339)
_URBAN$HHD_ROOF_GRASS_11	0.1054	(0.0475)
_URBAN$HHD_ROOMS_101	-0.4033	(0.0682)
_URBAN$HHD_ROOMS_102	-0.2877	(0.0563)
_URBAN$HHD_ROOMS_103	-0.2806	(0.0543)
_URBAN$HHD_ROOMS_104	-0.1993	(0.0535)
_URBAN$HHD_WALL_MUDBRCK_11	-0.0967	(0.0386)
Adjusted R^2		0.618
Observations		12262

Table 8: ELL household (log) consumption regression model

	Coefficient	Std. Err.
Intercept	-4.3293	(0.0726)
STRATA_1	0.0265	(0.1417)
STRATA_2	-0.2575	(0.1435)
STRATA_3	-0.2739	(0.1332)
STRATA_4	0.2134	(0.0946)
STRATA_5	0.2481	(0.064)
TA_REMOTE100K	-0.062	(0.0262)
Adjusted R^2		0.004
Observations		12262

Table 9: ELL heteroskedasticity model

Appendix 4: Remote-sensing variables and sources

***Theme* / Layer**	**Description**	**Approximate grid size**	**Source**
Initial conditions			
dist2cap	Distance to capital	N/A	NSO and Global Insights v6.1
dist2border	Distance to border	N/A	NSO
dist2major_river	Distance to major river (centerline)	N/A	NSO and Natural Earth 3.0 1:10 million (2017)
dist2rail	Distance to railway	N/A	NSO and Global Insights v6.1
elev30m	Elevation from SRTM	90m	Jarvis et al. (2008)
srtm_popmask	Elevation in populated areas in 2012 from Landscan	1km2	Jarvis et al. (2008); Bright et al. (2013)
Slope	Slope (degrees of slope (x100))	30-arc-seconds	Verdin et al. (2007)
Ruggedness	Ruggedness Index	3 arc-seconds	Nunn and Puga (2012)

***Theme* / Layer**	**Description**	**Approximate grid size**	**Source**
rugged_popmask	Ruggedness Index in populated areas	1km2	Nunn and Puga (2012); Bright et al. (2013)
Agriculture			
ag_hybrid	Agricultural Hybrid	1km2	Fritz et al. (2015)
lgp50_v6990	Length of Growing Period	10km2	FAO-GAEZ version 3 (Fischer et al. 2012)
rain_crop_sh	Share of area with rainfed crop	1km2	GFSAD1000 V1.0: (Teluguntla et al. 2014)
irrigated_crop_sh	Share of area with irrigated crop	1km2	GFSAD1000 V1.0: (Teluguntla et al. 2014)
agr_oppt_cost	Global map of gross economic rents from agricultural lands (2000 US dollars per hectare per year)	10km2	Naidoo and Iwamura (2007)
Climate			
bio1	Annual Mean Temperature	1km2	Bioclim (Hijmans et al. 2005)
bio4	Temperature Seasonality (standard deviation *100)	1km2	Bioclim (Hijmans et al. 2005)
bio12	Annual Precipitation	1km2	Bioclim (Hijmans et al. 2005)

***Theme* / Layer**	**Description**	**Approximate grid size**	**Source**
bio15	Precipitation Seasonality (Coefficient of Variation)	1km2	Bioclim (Hijmans et al. 2005)
bio16	Precipitation of Wettest Quarter	1km2	Bioclim (Hijmans et al. 2005)
bio17	Precipitation of Driest Quarter	1km2	Bioclim (Hijmans et al. 2005)
oxmap_daylst	Day time land surface temperature (Celcius)	5km2	MOD11A2 and Weiss et al. (2014)
oxmap_ngtlst	Night time land surface temperature (Celcius)	5km2	MOD11A2 and Weiss et al. (2014)
Aridity	Aridity Index	1km2	Trabucco and Zomer (2009)
Pet	Potential Evapotranspiration	1km2	Trabucco and Zomer (2009)
Environment			
NDVI	Normalized Difference Vegetation Index (NDVI) (1981-2011)	10km2	Zhu et al. (2013)
lndsatndvi_2008	Normalized Difference Vegetation Index (NDVI) (2008)		LandSat

Theme / Layer	Description	Approximate grid size	Source
EVI	Enhanced Vegetation Index (EVI)	5km2	MODIS BRDF-corrected imagery (MCD43B4) with gap-filled method by Weiss et al. (2014)
Forloss	Forest Loss (2008 – 2000)		Hansen et al. (2013)
PM2.5	Particulate Matter 2.5 at 35% RH (ug/m3) (no dust or seasalt) (2008)	1km2	van Donkelaar et al. 2015
ET_solrad	Extraterrestrial Solar Radiation by month	2.5km2	Zomer et al. 2008
Socio-economics			
NTL	Night Time Lights (1996, 2000, 2010)	1km2	NOAA DMSP-OLS
Population count and density			
lpop12	Landscan (2012)	1km2	Bright et al. (2013)
urban_lpop12	Urban population	1km2	Bright et al. (2013); WDI (2017); Authors' calculations
ghspop	GHS-Pop (1975, 1990, 2000, 2014)	1km2	Freire et al. (2016)
ghsbuiltup	GHS-Built up (1975, 1990, 2000, 2014)	300m	Freire et al. (2016)

***Theme* / Layer**	**Description**	**Approximate grid size**	**Source**
Travel time and accessibility			
dist2railway	Distance to railway	N/A	NSO and Global Insights v6.1
tt50k2000	Mean Travel time to major cities (circa 2000)	1km2	Uchida and Nelson (2008)
tt50k2010	Mean Travel time to major cities (circa 2010)	1km2	Berg et al. (in prep)
ai50k2000	Agglomeration Index (circa 2000)	1km2	Uchida and Nelson (2008)
ai50k2010	Agglomeration Index (circa 2010)	1km2	Roberts et al. (2017)

Table 10: Data source summary

References

Alderman, H., Babita, M., Demombynes, G., Makhatha, N. and Özler, B. (2002). How low can you go? combining census and survey data for mapping poverty in south africa. *Journal of African Economies*, **11,** number 2, 169–200.

Anselin, L. (2001). *Spatial econometrics*, A companion to theoretical econometrics, 310330 edn.

Araujo, M., Ferreira, F., Lanjouw, P. and Ozler, B. (2008). Local inequality and project choice: Theory and evidence from ecuador. *Journal of Public Economics*, **92,** 1022–1046.

Baird, S., McIntosh, C. and Ozler, B. (2013). The regressive demands of demand-driven development. *Journal of Public Economics*, **106,** 27–41.

Bazzi, S. (2017). Wealth heterogeneity and the income elasticity of migration. *American Economic Journal: Applied Economics*, **9,** 219–255.

Bedi, T., Coudouel, A. and Simler, K. (2007). *More than a pretty picture: using poverty maps to design better policies and interventions*. World Bank Publications.

Bell, K. and Bockstael, N. (2000). Applying the generalized-moments estimation approach to spatial problems involving microlevel data. *Review of Economics and Statistics*, **82,** 72–82.

Berg, C., Blankespoor, B., Li, Z. and Selod, H. (2017). Travel time to major cities. mimeo. The World Bank.

Blumenstock, J., Cadamuro, G. and On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science*, **350,** 1073–1076.

Blumenstock, J. E. (2016). Fighting poverty with data. *Science*, **353,** 753–754.

Bosco, C., Alegana, V., Bird, T., Pezzulo, C., Bengtsson, L., Sorichetta, A. and Wetter, E. (2017). Exploring the high-resolution mapping of gender-disaggregated development indicators. *Journal of The Royal Society Interface*, **14**.

Bourguignon, F. (2017). *The globalization of inequality*. Princeton and Oxford: Princeton University Press.

Bright, E. A., Rose, A. N. and Urban, M. L. (2013). *LandScan 2012: High Resolution Global Population Data*, US Department of Energy: http://www. ornl. gov/landscan edn. UT-Battelle, LLC. Oak Ridge National Laboratory.

Burke, M., Driscoll, A., Lobell, D. and Ermon, S. (2021). Using satellite imagery to understand and promote sustainable development. *Science*, **371**.

Chi, G., Fang, H., Chatterjee, S. and Blumenstock, J. (2022). Microestimates of wealth for all low- and middle-income countries. *Proceedings of the National Acadamy of Sciences*, **119,** number 3, 1–11.

Crost, B., Felter, J. and Johnston, P. (2014). Aid under fire: Development projects and civil conflict. *American Economic Review*, **104,** 1833–1856.

Demombynes, G. and Ozler, B. (2005). Crime and local inequality in south africa. *Journal of Development Economics*, **76,** 265–292.

Dijkstra, L. and Poelman, H. (2014). A harmonised definition of cities and rural areas: the new degree of urbanisation. Regional Working Paper. European Comission Directorate-General for Regional and Urban Policy.

Donaldson, D. and Storeygard, A (2016). The view from above: Applications of satellite data in economics. *Journal of Economic Perspectives*, **30,** number 4, 171–198.

Dreze, J., Lanjouw, P. and Sharma, N. (1998). *in Lanjouw, P. and Stern, N. (eds) Economic Development in Palanpur Over Five Decades (Oxford: Oxford University Press).*

Elbers, C., Fujii, T., Lanjouw, P., Ozler, B. and Yin, W. (2007). Poverty alleviation through geographic targeting: How much does disaggregation help? *Journal of Development Economics*, **83,** number 1, 198–213.

Elbers, C., Lanjouw, J. and Lanjouw, P. (2002). Micro-level estimation of welfare. *World Bank Policy Research Working Paper No. 2911.*

Elbers, C., Lanjouw, J. and Lanjouw, P. (2003). Micro–level estimation of poverty and inequality. *Econometrica*, **71,** number 1, 355–364.

Elbers, C., Lanjouw, J. and Lanjouw, P. (2005). Crime and local inequality in south africa. *Journal of Economic Geography*, **5,** 101–118.

Elbers, C., Lanjouw, P., and Leite, P. (2008). Brazil within brazil: Testing the poverty mapping methodology in minas gerais. Policy Research Working Paper 4513. The World Bank.

Elbers, C. and van der Weide, R. (2014). Estimation of normal mixtures in a nested error model with an application to small area estimation of poverty and inequality. Policy Research Working Paper 6962. The World Bank.

Engstrom, R., Hersh, J. and Newhouse, D. (2022). Poverty from space: Using high-resolution satellite imagery for estimating economic well-being. *World Bank Economic Review*, **36,** 382–412.

Ferreira, F. H. G., Chen, S., Dabalen, A. L., Dikhanov, Y. M., Hamadeh, N., Jolliffe, D. M., Narayan, A., Prydz, E. B., Revenga, A. L., Sangraula, P., Serajuddin, U. and Yoshida, N. (2015). A global count of the extreme poor in 2012 : data issues, methodology and initial results. Policy Research Working Paper 7432. The World Bank.

Filmer, D. and Pritchett, L. (2001). Estimating wealth effects without expenditure data – or tears: With application to educational enrollments in states of india. *Demography*, **38,** 115–132.

Filmer, D. and Scott, K. (2012). Assessing asset indices. *Demography*, **49,** number 1, 359–392.

Fischer, G., Nachtergaele, F. O., Prieler, S., Teixeira, E., Tóth, G., Van Velthuizen, H., Verelst, L. and Wiberg, D. (2012). Global agro-ecological zones (gaez v3. 0) - model documentation. Technical Report. FAO, Rome.

Freire, S., MacManus, K., Pesaresi, M., Doxsey-Whitfield, E. and Mills, J. (2016). *Development of new open and free multi-temporal global population grids at 250 m resolution*, Geospatial Data in a Changing World; Association of Geographic Information Laboratories in Europe (AGILE) edn. AGILE.

Fritz, S., See, L., McCallum, I., You, L., Bun, A., Moltchanova, E., Havlik, P. and Co. (2015). Mapping global cropland and field size. *Global change biology*, **21,** number 5, 1980–1992.

Fujii, T. (2010). Micro-level estimation of child undernutrition indicators in cambodia. *World Bank Economic Review*, **24,** number 3, 520–553.

Fujii, T. and van der Weide, R. (2020). Is predicted data a viable alternative to real data? *World Bank Economic Review*, **34,** 485–508.

Gibson, J., Le, T. and Kim, B. (2017). Prices, engel curves, and time-space deflation: Impacts on poverty and inequality in vietnam. *World Bank Economic Review*, **31,** 504–530.

Goldberger, A. S. (1962). Best linear unbiased prediction in the generalized linear regression model. *Journal of the American Statistical Association*, **57,** 369–375.

Hansen, M.C., Potapov, P.V., Moore, R., Hancher, M., Turubanova, S. A., Tyukavina, A., Thau, D., Stehman, S.V., Goetz, S.J., Loveland, T.R. and Kommareddy, A. (2013). High-resolution global maps of 21st-century forest cover change. *Science*, **342,** 850–853.

Henderson, C. (1953). Estimation of variance and covariance components. *Biometrics*, **9,** 226–253.

Henderson, J. V., Storeygard, A. and Weil, D. N. (2012). Measuring economic growth from outer space. *American Economic Review*, **102,** number 2, 994–1028.

Hentschel, J., Lanjouw, J., Lanjouw, P. and Poggi, J. (2000). Combining census and survey data to trace the spatial dimensions of poverty: A case study of ecuador. *World Bank Economic Review*, **14,** number 1, 147–165.

Hertel, T. W., Burke, M. B. and Lobell, D. B. (2010). The poverty implications of climate-induced crop yield changes by 2030. *Global Environmental Change*, **20,** number 4, 577–585.

Hijmans, R. J., Cameron, S., Parra, J., Jones, P. G., Jarvis, A. and Richardson, K. (2005). *WorldClim*, version 1.3 edn. University of California, Berkeley.

Imran, M., Stein, A. and Zurita-Milla, R. (2014). Investigating rural poverty and marginality in burkina faso using remote sensing-based products. *International Journal of Applied Earth Observation and Geoinformation*, **26,** 322–334.

Jarvis, A., Reuter, H. I., Nelson, A. and Guevara, E. (2008). *Hole-filled seamless SRTM data*, version 4, available at: http://srtm. csi. cgiar.org edn. International Center for Tropical Agriculture (CIAT).

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B. and Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, **353,** 790–794.

Kelejian, H. and Prucha, I. (2010). Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances. *Journal of Econometrics*, **157,** 53–67.

Kilic, T., Serajuddin, U., Uematsu, H. and Yoshida, N. (2017). Costing household surveys for monitoring progress toward ending extreme poverty and boosting shared prosperity. Policy Research Working Paper 7951. The World Bank.

Lee, K. and Braithwaite, J. (2022). High-resolution poverty maps in sub-saharan africa. *World Development*, **159**.

Leyk, S., Gaughan, A. E., Adamo, S. B., Sherbinin, A. D., Balk, D., Freire, S. and Comenetz, J. (2019). The spatial allocation of population: a review of large-scale gridded population data products and their fitness for use. *Earth System Science Data*, **11,** number 3, 1385–1409.

Maloney, W. and Caicedo, F. (2015). The persistence of (subnational) fortune. *Economic Journal*, **126,** 2363–2401.

Marx, B., Stoker, T. and Suri, T. (2019). There is no free house: Ethnic patronage in a kenyan slum. *American Economic Journal: Applied Economics*, **11,** 36–70.

Molina, I. and Rao, J. (2010). Small area estimation of poverty indicators. *Canadian Journal of Statistics*, **38,** 369–385.

Naidoo, R. and Iwamura, T. (2007). Global-scale mapping of economic benefits from agricultural lands: Implications for conservation priorities. *Biological Conservation*, **140,** number 1-2, 40–49.

Nelson, A. (2008). *Travel time to major cities: A global map of Accessibility*, Global Environment Monitoring Unit edn. Joint Research Centre of the European Commission, Ispra Italy.

Newhouse, D., Merfeld, J., Ramakrishnan, A., Swartz, T. and Lahiri, P. (2022). Small area estimation of monetary poverty in mexico using satellite imagery and machine learning. *World Bank Policy Research Working Paper No. 10175*.

Nhu, C. and Noy, I. (2020). Measuring the impact of insurance on urban earthquake recovery using nightlights. *Journal of Economic Geography*, **20,** 857–877.

NSO (2012). Third integrated household survey 2010-2011 - final report. Technical Report. Malawi National Statistical Office (NSO) and The World Bank.

Nunn, N. and Puga, D. (2012). Ruggedness: The blessing of bad geography in africa. *Review of Economics and Statistics*, **94,** number 1, 20–36.

Perez-Heydrich, C., Warren, J. L., Burgert, C. R. and Emch, M. (2013). Guidelines on the use of DHS GPS data edn. ICF International.

Pinkovskiy, M. and Sala-i Martin, X. (2016). Lights, camera ... income! illuminating the national accounts-household surveys debate. *The Quarterly Journal of Economics*, **131,** number 2, 579–631.

Pokhriyal, N. and Jacques, D. C. (2017). Combining disparate data sources for improved poverty prediction and mapping. *Proceedings of the National Academy of Sciences*, **114,** E9783–E9792.

Pratesi, M. and Salvati, N. (2008). Small area estimation: the eblup estimator based on spatially correlated random area effects. *Statistical Methods and Applications*, **17,** 113–141.

Ravallion, M. (2018). Inequality and globalization: A review essay. *Journal of Economic Literature*, **56,** number 2, 620–642.

Roberts, M., Blankespoor, B., Deuskar, C. and Stewart, B. (2017). Urbanization and development: is latin america and the caribbean different from the rest of the world? Policy Research Working Paper 8019. The World Bank.

Searle, S., Casella, G. and McCulloch, C. (1992). *Variance components*. New York: Wiley.

Smythe, I. and Blumenstock, J. (2022). Geographic microtargeting of social assistance with high-resolution poverty maps. *Proceedings of the National Acadamy of Sciences*, **119,** number 32, 1–10.

Steele, J. E., Sundsøy, P. R., Pezzulo, C., Alegana, V. A., Bird, T. J., Blumenstock, J. and Hadiuzzaman, K. N. (2017). Mapping poverty using mobile phone and satellite data. *Journal of The Royal Society Interface*, **14**.

Tarozzi, A. (2011). Can census data alone signal heterogeneity in the estimation of poverty maps? *Journal of Development Economics*, **95,** 170–185.

Tarozzi, A. and Deaton, A. (2009). Using census and survey data to estimate poverty and inequality for small areas. *Review of Economics and Statistics*, **91,** 773–792.

Teluguntla, P., Thenkabail, P., Xiong, J., Gumma, M.K., Giri, C., Milesi, C., Ozdogan, M., Congalton, R., Tilton, J., Sankey, T.R., Massey, R., Phalke, A. and Yadav, K. (2015). *Global Cropland Area Database (GCAD) derived from Remote Sensing in Support of Food Security in the Twenty-first Century: Current Achievements and Future Possibilities*, Vol. II edn. Land Resources: Monitoring, Modelling, and Mapping, Remote Sensing Handbook edited by Prasad S. Thenkabail. In Press.

Trabucco, A. and Zomer, R. J. (2009). *Global aridity index (global-aridity) and global potential evapo-transpiration (global-PET) geospatial database*. CGIAR Consortium for Spatial Information.

Uchida, H. and Nelson, A. (2008). Agglomeration index: towards a new measure of urban concentration. Background paper for the World Bank's World Development Report 2009. The World Bank.

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. Policy Research Working Paper 7028. The World Bank.

Van Donkelaar, A., Martin, R. V., Brauer, M. and Boys, B. L. (2015). Use of satellite observations for long-term exposure assessment of global concentrations of fine particulate matter. *Environmental health perspectives*, **123,** number 2, 135–143.

Van Veelen, M. and van der Weide, R. (2008). A note on different approaches to index number theory. *American Economic Review*, **98,** 1722–1730.

Van Velthuizen, H., Huddleston, B., Fischer, G., Salvatore, M., Ataman, E., Nachtergaele, F. O., Zanetti, M. and Bloise, M. (2007). Mapping biophysical factors that influence agricultural production and rural vulnerability. Woking Paper 11. FAO, Rome.

Watmough, G. R., Atkinson, P. M., Saikia, A. and Hutton, C. W. (2016). Understanding the evidence base for poverty–environment relationships using remotely sensed satellite data: an example from assam, india. *World Development*, **78,** 188–203.

Watmough, G. R., Marcinko, C. L., Sullivan, C., Tschirhart, K., Mutuo, P. K., Palm, C. A. and Svenning, J. C. (2019). Socioecologically informed use of remote sensing data to predict rural household poverty. *Proceedings of the National Academy of Sciences*, **116,** number 4, 1213–1218.

Weiss, D. J., Atkinson, P. M., Bhatt, S., Mappin, B., Hay, S. I. and Gething, P. W. (2014). An effective approach for gap-filling continental scale remotely sensed time-series. *ISPRS Journal of Photogrammetry and Remote Sensing*, **98,** 106–118.

Winsemius, H. C., Jongman, B., Veldkamp, T. I., Hallegatte, S., Bangalore, M. and Ward, P. J. (2015). Disaster risk, climate change, and poverty: assessing the global exposure of poor people to floods and droughts. Technical Report.

Zhao, X., Yu, B., Liu, Y., Chen, Z., Li, Q., Wang, C. and Wu, J. (2019). Estimation of poverty using random forest regression with multi-source data: A case study in bangladesh. *Remote Sensing*, **11,** number 4.

Zhu, Z., Bi, J., Pan, Y., Ganguly, S., Anav, A., Xu, L. and Myneni, R. (2013). Global data sets of vegetation leaf area index (lai) 3g and fraction of photosynthetically active radiation (fpar) 3g derived from global inventory modeling and mapping studies (gimms) normalized difference vegetation index (ndvi3g) for the period 1981 to 2011. *Remote sensing*, **5,** number 2, 927–948.

Zomer, Robert J, Trabucco, Antonio, Bossio, Deborah A and Verchot, Louis V (2008). Climate change mitigation: A spatial analysis of global land suitability for clean development mechanism afforestation and reforestation. *Agriculture, ecosystems & environment*, **126,** number 1-2, 67–80.