Australian Government

Australian Centre for International Agricultural Research

Final report

Small research and development activity

project	**Vulnerability in the Anthropocene: a prospective analysis of the need for social protection**
project number	LS/2020/206
date published	9/05/2023
prepared by	Dr Paulo Santos and Reiss McLeod
co-authors/ contributors/ collaborators	Stefan Meyer, Mukhammad Fajar Rakhmadi, Hai-Chau Le
approved by	Dr Anna Okello
final report number	FR2023-015
ISBN	978-1-922983-07-7
published by	ACIAR GPO Box 1571 Canberra ACT 2601 Australia

This publication is published by ACIAR ABN 34 864 955 427. Care is taken to ensure the accuracy of the information contained in this publication. However, ACIAR cannot accept responsibility for the accuracy or completeness of the information or opinions contained in the publication. You should make your own enquiries before making decisions concerning your interests.

© Commonwealth of Australia 2023 - This work is copyright. Apart from any use as permitted under the Copyright Act 1968, no part may be reproduced by any process without prior written permission from the Commonwealth. Requests and inquiries concerning reproduction and rights should be addressed to the Commonwealth Copyright Administration, Attorney-General's Department, Robert Garran Offices, National Circuit, Barton ACT 2600 or posted at http://www.ag.gov.au/cca.

Contents

1 Acknowledgments

We thank the Dr Anna Okello and Dr Todd Sanderson for the stimulating discussions as this study progressed. We are equally thankful to all participants in an online presentation of these results to ACIAR, in particular to Dr Eric Huttner and Dr Veronica Doerr for some of the preliminary results and their policy implications, particularly with respect to agricultural insurance.

Finally, we thank multiple staff members in the Statistical Offices of Cambodia, Timor-Leste and The Philippines for facilitating access to the data used in this study. Particular thanks are due to our friends in the Myanmar Country Office of CDE-Bern who facilitated access to statistical data for that country while in the middle of social convulsion.

2 Executive summary

The Covid-19 pandemic exposed how fragile were some of the progresses made towards reducing poverty, and raised the need to better understand economic vulnerability and resilience to shocks. This report presents the results of the relation between human wellbeing (measured as income/consumption), natural capital (in particular, biodiversity, primary forest and soil) and climate in the rural areas of seven Southeast Asian countries (Cambodia, Indonesia, Lao PDR, Myanmar, The Philippines, Timor-Leste and Vietnam).

The analysis combines data on income/consumption from recent representative Household Income and Expenditure Surveys with a rich set of other georeferenced data that measure natural capital and climate. Using Machine Learning techniques (in particular, regression forests and its interpretation through surrogate models) we classify rural households in eight groups, characterised by distinct between-group living standards, and reflecting different environmental characteristics. Several conclusions emerge from this analysis.

The first is the importance of poverty in rural Southeast Asia, with 41% of the households in our data being allocated to groups with average income below the poverty line. The second is the considerable heterogeneity of the natural environment of poor households, aggregated in five distinct groups, in contrast with the remaining 59% (non-poor) households who were aggregated in, essentially, two large clusters. The third is the importance of environmental risk in the first month of the rainy season (measured by the variance of temperature and variance of number of wet days): risk does create poverty in this data, suggesting a role to be played in the development of agricultural insurance (and, more generally, social safety nets).

The fourth conclusion is that these groups also exhibit significant differences in terms of income variability which, together with differences in average income, lead to significant differences in vulnerability to poverty. Finally, our analysis shows that biodiversity richness predicts both lower income variability and lower expected income, suggesting that biodiversity conservation may come at the cost of increased vulnerability to poverty.

3 Background

AFS [Agriculture and Food Systems] innovation feeds back into demographic transitions, income growth, and the climate and extinction crises. *Indeed, we face real climate, environmental, health, and social dangers today and in the decades ahead in part because the past century's AFS innovations have focused so tightly on boosting agricultural productivity, especially output per unit area cultivate (i.e., yields), to the exclusion of other objectives.* Nudging the coming generation of AFS innovations in better directions requires envisioning a broader set of shared objectives. (Barrett et al, 2020, pp. 34-35, our emphasis)

The Covid-19 pandemic exposed chronic development fault lines across the Asia-Pacific. Recent development gains have been undermined and the costs of the pandemic were not equally distributed. For example, the poor were disproportionately affected because they could not follow the World Health Organization's social isolation recommendations (Brown, Ravallion and van de Walle, 2020). The possibility of large fractions of the population descending into poverty saw, as a consequence, an unprecedented and perhaps surprising willingness of governments to transfer income to households as a way to smooth the economic consequences of the macroeconomic shock created by the pandemic.

Government transfers are one of several policy options available to address the consequences of shocks. For example, Bigio, Zhang and Zilberman (2020) discuss, from a macroeconomic perspective, the relative merits of increasing transfers versus expanding access to credit with the objective of managing the business cycle. From a microeconomic perspective, closer to the analysis in this report and more focused on how to reduce persistent poverty, Barrett (2005) distinguishes between "safety nets", intended to protect the asset base of the poor from shocks and prevent a slide into poverty from which escape may be difficult, and "cargo nets", directed at building up productive assets and supporting poverty reduction.

From a planner's perspective, safety nets are attractive given the slowly built evidence regarding the impacts of cash transfers, how easy it is to roll-out such transfers even in contexts of low penetration of financial markets and, in the case of the UBI, the little information needed for its implementation given its untargeted nature. From a budgetary perspective, safety nets such as UBI capture a large part of public funds, which need to be raised via taxation with distortionary effects that are not accounted in this very simplified illustration. This consideration may make technological change attractive, given the perceived high returns to investment in R&D for example, although everything that makes safety nets attractive is now absent: long lags between development of technologies and widescale adoption (even if successful), lower evidence regarding "what works" and for whom.

Although useful as a guide, it is important to recognise that, in reality, the distinction between cargo and safety nets is less clear. The presence of multiple market failures makes uninsured risk (ie, the likelihood of negative shocks against which agents cannot effectively protect themselves) a potential driver of the "Faustian bargain" where lower expected income is accepted as the price to pay to avoid catastrophic reductions in living standards (Wood, 2003): in this context, self-insurance breeds poverty and market failures may blur the metaphorical difference between "giving a fish" and "teaching to fish".

The hope then, as made clear in the recent reviews of the experience with index insurance (eg, Carter et al, 2017s) or cash transfers (Hanlon, Barrientos and Hulme, 2012) is that the direct reduction of the risk of being poor (conceptually, a "safety net") can lead to dynamic adaptations, such as investment in more productive but riskier technologies (conceptually, a "cargo net") that contribute to reduce poverty. Similarly, the

large gains in mean yield associated with the development of GM technology may lead to improvements in welfare, and reduce the need for safety nets, even if their effect on yield risk is ambiguous (Nolan and Santos (2019)). Or, perhaps more sensibly, rather than and "either/or" discussion, the different types of policies can be combined, as in the analysis of Boucher et al (2021).

The point we will want to emphasise here, and keep in mind in the remaining of this report, is that despite their largely positive record, safety nets are but one instrument in the arsenal of instruments intended to address the importance of risk as a driver of poverty. However, while safety nets may "only" require the identification of the poor (and this, in itself, is no small requirement), the development of effective cargo nets requires the preliminary identification of who is 'at risk" of becoming poor, ie, vulnerable (World Bank, 2001).

This report presents the results of an analysis of the relative importance of different predictors of vulnerability to poverty. We frame that discussion in terms of what has come to be known as the Anthropocene, an umbrella for the large set of human driven changes in the natural environment. Among those changes, we focus on climate and natural capital, in particular biodiversity, both of which are increasingly perceived to be changing at rates that threaten Human survival (Steffen, 2015).

Before we present the approach used in this study and our results, a few preliminary points are in order. The first is what Caro et al (2021) call the "inconvenient misconception" that the accelerated reduction in biodiversity is being driven by climate change – something that, at least until now, is not true. This is not to say that they are independent: importantly, ecosystem services provided by a healthy environment seem to be among the most cost-effective ways to mitigate or adapt to climate change, driving much effort into the design of ways to design solutions that may address both crises (eg, Zhu et al (2021))

The second is that it is unlikely that society will not demand changes in the way that the agricultural sector contributes to human welfare, reflecting the sector's contribution to ongoing environmental degradation. This is particularly evident in the case of changes in biodiversity, mostly driven by land use change induced by the expansion of agricultural production (Pendrill et al, 2022) as well as harvesting of wildlife in rural areas (reference), potentially as a coping strategy to agricultural production shocks (Kader and Santos, 2022). The emphasis on expanding the area devoted to conservation, written into ambitious claims of reserving up to "Half the Earth" (Wilson, 2016) for Nature, has important implications for agricultural production (Mehrabi, Ellis and Ramankutty (2018)) and is an obvious illustration of the challenges posed to agricultural development, particularly in developing countries given their overlap with biodiversity hotspots (Myers, 2000). A similar argument can be made about the potential impacts of carbon-forestry.

Because agriculture is at the intersection of all SDGs, other challenges, for example with respect to diet and its implications for human health (Willet et al (2019)) or reducing the risk of contact with new infectious diseases (Roth et al (2019)), are likely to add to the need for change in the way that farming contributes to human welfare. The implication, which is also the main message of the quote with which we opened this section, is that the traditional emphasis on the contributions of agriculture to economic development and poverty reduction, dating back to the Green Revolution and somewhat repeated, with geographic nuances, in the most recent World Development Report on Agriculture (World Bank, 2007), is likely to be challenged given that society demands are also changing. A mathematical truth is that optimisation subject to multiple binding constraints leads to values of a single objective that are lower than in a less constrained problem: that explains why win-win solutions are elusive (Hegwood, Lagendorf and Burgess (2022)), and why a recent comprehensive review of agriculture's contribution to sustainable development suggests that only bundles of interventions, addressing multiple constraints, are likely to succeed (Barrett et al, 2020).

Finally, one comment about the role of smallholder farmers. As Hayami (2002) influential analysis of plantation economy makes clear, the choice between large- and small-scale production was always a political decision. The perceived superiority of smallholders in terms of land productivity, which reflected the capacity to overcome labour market failures, favoured the emphasis on smallholders in many contexts. Smallholder farmers do not have an advantage in overcoming credit and information market failures (eg, Collier and Dercon (2014)). The need to steadily reduce the land and water footprint of food production through substituting capital for land and water inputs, a process that Barrett (2021) labels the deagrarianization of food production, will only reinforce the importance of those limitations, calling into question the viability of smallholders as the central actors in the development of a multifunctional agriculture. This seems a fundamental tension between the poverty reduction and environmental sustainability objectives of the Sustainable Development Goals which merits further attention.

4 Objectives

This project contributes to the overall objective in ACIAR's ten-year strategy of creating options for sustainable development that address the needs of smallholder farmers in the countries where it operates. This project aims to address one key research question: which factors predict vulnerability to poverty in rural areas of Southeast Asia?

5 Methodologies

5.1 Data

To quantify the relative importance of predictors of vulnerability to poverty we compiled data from representative national household income and expenditure surveys, designed to measure and monitor poverty, from seven Southeast Asian countries: Cambodia, Indonesia, Lao PDR, Myanmar, The Philippines, Timor-Leste and Vietnam. All surveys were representative at rural (vs urban) level, and in the analysis we will focus on the rural strata only. The information available in these surveys is described in detail in section 5.1.1.

One important feature of this data is the availability of information about spatial location of households in the national income and expenditure surveys. This feature enables the linking of economic data on consumption and wealth with other datasets that contain information of a wide array of environmental variables that may capture the two dimensions of the Anthropocene in which we planned to focus: climate change and biodiversity loss. These environmental variables are described in section 5.1.2.

5.1.1 Income and Expenditure Data

We use the latest publicly available survey data for each country. The year and sample size of each of the surveys used are presented in Table 1. In Appendix 9.1, Table 17 to Table 23 we present information about each of the Household Income and Expenditure Surveys (HIES) in the different countries. Besides information like the number of survey rounds and sample size, we report the survey sampling strategy when there are multiple survey rounds (ie, either panel or repeated cross-section), as well as strata and sampling, given the implications of survey design for the analysis (Deaton, 2018).

Table 1 Survey round selected for the analysis by country

Country	Year	Sample size
Cambodia	2014	11,622
Indonesia	2019	181,981
Lao PDR	2012/2013	4,385
Myanmar	2017	11,915
The Philippines	2018	79,850
Timor-Leste	2014/15	4,920
Vietnam	2018	25,071

The main variable of interest is household consumption, defined as the sum of goods and services consumed by a household within a predefined period, and typically seen as an accurate measure of wellbeing in developing countries with large informal employment (Deaton, 2018). Data for the consumption aggregate is collected by National Statistical Offices (NSOs) through nationally representative household income and expenditure surveys (HIES). Well known examples are household surveys collected by the Living Standard Measurement Study (LSMS) (see Grosh and Glewwe (2000) for an overview) with a lasting methodological influence, inclusively in a direct way in some of the surveys used in this study (Myanmar, Timor Leste, Vietnam).

NSOs collect and make available the aggregated household consumption variable and data for different consumption modules, These modules typically cover four categories, food (e.g. purchased, in-kind, home-produced and food away from home), non-food (e.g. education, health and other non-food), durables and housing (Deaton and Zaidi, 2002), with data for several items collected for each category. The data collection method differs by survey and module. Some data is collected through diaries, where households' record their consumption, while other data is collected via recall. The length of the period of data collection varies by category. For consumption goods, the period is usually shorter (typically., one week to one month) than for durables (e.g., 3 months, 6 months or 1 year). The aggregate for each category is then extrapolated to one year.

Aggregate consumption is calculated by summing the value of goods consumed in each category. For purchased consumption goods, the values are directly registered as collected in the household survey. Goods that were received by the households' as in-kind payment or self-produced (for which, typically, only quantities are collected) are valued using average local prices. For durables, the usage values (current value depreciated for the total time of usage) are estimated. Some aggregates also contain housing expenditure (ie, rent, which are estimated through hedonic regressions and imputed in the case of households owning their residence).The final consumption aggregate accounts for cost-of-living or spatial differences by deflating expenditure by a Paasche Price Index (Deaton and Zaidi, 2002). The prices used to calculate the price index are either unit values, separately collected through community surveys fielded with the household survey or regional data from other sources.

We summarize the data collected for each country in Table 2. Data on durables is typically not available and housing are not included in all of the consumption aggregates due to data limitations. For example, the income aggregate produced by the NSO in Lao PDR was calculated without housing rents because the renting market information was very thin, as only 1.4% of the respondents were tenants (Pimhidzai et al., 2014). Given the differences in the data collected across the four consumption modules (food, non-food, rent for housing and durables), the consumption indicator used in this study is based on value of consumption of food and non-food items only, to ensure comparability across countries. As a result, our estimates (for example, of poverty) will differ slightly from official national estimates.

Table 2 Consumption data

	Food consumption	Non-food consumption	Rent for housing	Durables (use value)
Cambodia	X	X	X	
Indonesia	X	X	X	
Lao PDR	X	X		
Myanmar	X	X	X	X
The Philippines	X	X	X	
Timor-Leste	X	X	X	
Vietnam	X	X	X	

In addition, we make a number of adjustments to the expenditure variables. Given we have data from different years (see Table 1), we convert the consumption aggregate per adult equivalent to a common year (2019) and common currency (USD) to allow for comparability.

We also rely on a common adult equivalent adjustment for all surveys. We follow the approach suggested by Deaton and Zaidi (2002):

$$AE = (A + \alpha C)^{\theta}$$

where adult equivalent is the sum of the number of adults (A) and children (C) adjusted by a parameter that accounts for differences in the relative expenditure of children when compared to adults (α), which we assume to be 0.5 (as the cost of children is relatively low in an agricultural economy), and a parameter that accounts for economies of scale, ie shared consumption within a household, (θ) which we assume to be 0.9, reflecting the fact that economies of scale are usually low in developing countries, given that food is the main consumption good in households in developing countries.

The scope of the HIES data is quite wide, with most surveys collecting data through multi-modular surveys that, in addition to consumption, include household roster (with information on demographic characteristics such as gender, age and ethnicity of household head, household composition, education), assets, labour and employment (Grosh and Glewwe, 2000). We use data from these modules to construct predictors of vulnerability to poverty, grouped into two categories: human capital and physical capital (see Table 3). While the general scope and coverage of each survey is similar, there are some important differences. Some HIES datasets only contain a limited set of variables for poverty analysis (for example, The Philippines, Timor Leste, Vietnam) while others include a broad set of modules covering multiple topics (for example, Myanmar). Village surveys are also administered, however, it was not possible to get access to the village survey data in all countries (for example, Philippines and Myanmar) due to data protection policies.

Table 3 Household Characteristics

Category	Variable	Description	Comments
Human capital	Female household head	Dummy indicating whether the head of the household is a female	All
	Age household head	Variable representing the age of the household head	All
	Schooling household head	Variable representing the level of education of the household head – presented as dummy variables for primary, secondary, university	All
	Household size	Variable representing the size of the household	All
	Majority group	Variable indicating whether the household belongs to a majority ethnic group	Philippines and Indonesia excluded
Physical capital	Housing index	Multiple component analysis applied to construct an index that picks up variation across the different variables on housing assets	All
	Large ruminants	Quantity of large livestock numbers converted to a common unit.	Indonesia excluded
	Small animals	Quantity of small livestock numbers converted to a common unit.	Indonesia excluded
	Productive asset index	Principal component analysis applied to construct an index that picks up variation across the different variables about productive assets	All
Village characteristics	Market in village	Variable representing whether there is a market in the village	Laos only
	Road access all year	Variable representing whether there is a road access all year	Laos, Cambodia and Timor Leste only

5.1.2 Spatial data

The National Statistical Offices provide information about the different administrative levels at which we can place surveyed households (hereon, spatial identifiers). As Table 4 makes clear, the resolution of the spatial identifiers varies in each country, ranging from the village level to the provincial level. We have access to polygonal data which indicate the boundaries of each of the spatial identifiers, allowing us to link different layers of spatial data with the survey data using the common spatial identifier.

Table 4 List of the spatial identifiers

Country	Level of the spatial identifier
Cambodia	Commune
Indonesia	District
Lao PDR	Village
Myanmar	Village
The Philippines	Province
Timor-Leste	Suco (group of villages)
Vietnam	Village

We collect spatial data across three categories of variables, including natural capital and climate, physical capital, and shocks to health. Table 5 to Table 12 provide an overview of the different variables, grouped by category, and provide the resolution of the data. Most of the data is in raster format, with cells differing in size (for example, 100 by 100 metres, or 250 by 250 metres). We compile several variables by extracting the mean value of the spatial data listed in Table 5 to Table 8 for each spatial identifier described in Table 4. Each variable is calculated as:

$$a_j = \frac{1}{N_j} \sum_{i=1}^{N_j} g_{ij}$$

where a_j is the mean of the variable in geographical location j (e.g. district), g_{ij} is the value of the observation i (at cell level) within the geography, N_j is the total number of cells within geography j.

Natural capital and climate

We use several recently constructed datasets that have global coverage and, as such, define our variables of interest in a common way across countries.

We proxy for natural capital using variables that measure stocks of natural resources, both in terms of quantity (for example, soil depth) and, whenever possible, its condition (for example, erosion). We focus on variables that contribute to the production of biomass (for example, available water capacity and climatic variables) that can be interpreted as inputs in an agricultural production function.

Data on natural capital comes from a range of sources (see Table 5) while data on climate (Table 6) comes from CRU TS (Climatic Research Unit gridded Time Series) presented in Harris et al (2020).

Table 5 Natural capital variables

Variable	Description	Resolution	Year
Forest cover	We use data on forest cover presented in Hansen et al. (2013). Forest cover is defined as the percentage of a pixel covered with forest, where forest is defined as any vegetation that exceeds 5m in height. The raw data are the images collected by NASA's Landsat satellites.	30m	2000
Biodiversity Intactness	We use the global Biodiversity Intactness Index (BII) presented in Newbold et al. (2016), who follow the definition of Scholes and Biggs (2005). BII is the average abundance of a species (originally) present divided by the pre-anthropogenic abundance. The global index was calculated in several steps. First, biodiversity data was used from the PREDICTS (Projecting Responses of Ecological Diversity In Changing Terrestrial Systems) Project database (year of download 2015). The data consists of nearly 40,000 Species for 14 terrestrial biomes and is sufficiently representative. It contains information on both plants and vertebrate species. In a second step, the count data is projected using four variables measuring anthropogenic activities: land-use, the intensity of land-use, density of human population and distance between the location and a road. To determine the baseline value of species in an area, expected values were calculated for areas with minimal influence from humans. Scholes and Biggs (2005) defined the BII as: $$BII = 100\,x\,\frac{\sum_i \sum_j \sum_k R_{ij} A_{jk} I_{ijk}}{\sum_i \sum_j \sum_k R_{ij} A_{jk}}$$ where I stands for the relation of the taxonomic group's population i in the terrestrial biome j and land-use category k to the pre-anthropogenically influenced baseline. The variables R and A represent the species richness and the size of the terrestrial biome, respectively. The index is expressed in percentage and a larger value is interpreted as a more intact biodiversity.	~1km	2005
Soil Erosion	We use data on erosion, estimated using the Revised Universal Soil Loss Equation (RUSLE), provided in the original Global Soil Erosion map, available from the European Soil Data Centre (ESDAC). The risk of erosion on arable land in the RUSLE model is expressed as: $A=R*K*LS*C*P$ where A = soil loss (mg ha−1 year−1), R = rainfall erosivity factor (mm ha−1 year−1), K = soil erodibility factor (mg ha−1 year−1), LS = slope-length and slope steepness factor (dimensionless), C = land management factor (dimensionless), and P = conservation practice factor (dimensionless).	25 km	2012
Ruggedness	We use the high-resolution global Terrain Ruggedness Index data compiled by Nunn and Puba (2012) following the approach suggested by Riley et al. (1999) and calculated with data from GTOPO30 (USGS 1996) elevation data.	30 Arc Seconds	1996

Variable	Description	Resolution	Year
	GTOPO30 is a global elevation data set developed through a collaborative international effort led by staff at the US Geological Survey's Center for Earth Resources Observation and Science (EROS). The Terrain Ruggedness Index (TRI) is calculated as: $$TRI_{r,c} = \sqrt{\sum_{i=r-1}^{r+1} \sum_{i=c-1}^{c+1} (e_{i,j} - e_{r,c})^2}$$ where $e_{r,c}$ is the elevation in row r and cell c of the global elevation matrix of cells. TRI is the square root of the sum of all squared differences of the elevation of a grid from the elevation of its 8 surrounding grids.		
Slope	We use data of the average uphill slope of the polygon surface, constructed using the GTOPO30 elevation data (USGS 1996). For each point on the elevation grid, the absolute value of the difference in elevation between this point and the point on the Earth's surface 30 arc-seconds North of it is calculated. This is then divided by the sea-level distance between the two points to obtain the uphill slope. The same calculation is performed for each of the eight major directions of the compass (North, Northeast, East, Southeast, South, Southwest, West, and Northwest), and the eight slopes obtained are then averaged to calculate the mean uphill slope for the 30 by 30 arc-second cell centred on the point.	30 Arc seconds	1996
Available water capacity	We use data on the amount of water that can be stored in a soil profile and be available for growing crops, made available by SoilData for the soil depth interval 0-100 cm (Hengl et al., 2017).	5x5 arc-minutes	2000
Soil depth	We use the predicted absolute depth of the topsoil (surface to bedrock in cm), predicted using machine learning algorithms trained on soil ground observations, as described in Shangguan et al. (2017). Original data comes from a compilation of soil profile data (ca. 1,300,000 locations) and borehole data (ca. 1.6 million locations).	250 m	?
Elevation	We use altitude above sea level, provided by the Global Multi-resolution Terrain Elevation Data (GMTED2010) published by USGS and the National Geospatial-Intelligence Agency (NGA) (Danielson et al., 2011).	7.5-arc-seconds	2010
Latitude	We use the latitude of the centroid of each polygon.	point data	--

Contrary to variables presented in Table 5, which are snapshots of stocks of natural conditions in specific locations, the climate data is, as expected, dynamic: the CRU –TS, described in Harris et al. (2020) provides a high-resolution, monthly grid of land-based (excluding Antarctica) observations going back to 1901. We use eight observed and derived variables, including average temperature, average rainfall and the number of wet days for each month, as described in Table 6. We exclude several climatic variables from the analysis given their high correlation (between 0.88-0.96) with included variables (mean temperature in the wet season, variance of temperature in all months of the wet season, variance of wet days in the wet season, mean rainfall per month in the rainy season and variance of rainfall in the wet season). Given their almost perfect collinearity

with included variables, this decision improves prediction and substantially reduces estimation time.

Table 6 Climate variables

Temperature in first month of rainy season	Air temperature in degrees Celsius, at 2 meters above the surface, in the first month of the rainy season in the year of the survey**.**	**55km**
Variance of temperature in first month of rainy season	This variable is a measure of the variance of temperature (air temperature in degrees Celsius at 2 meters above the surface) in the first month of the rainy season across 2010 to 2020.	55km
Number of wet days in in first month of rainy season	This variable is a measure of the number of wet days (a wet day is one receiving ≥0.1 mm precipitation) in the first month of the rainy season in the year of the survey.	55km
Variance of wet days in first month of rainy season	This variable is a measure of the variance of number of wet days (a wet day is one receiving ≥0.1 mm precipitation) in the first month of the rainy season for the period 2010-2020.	55km
Total Number of wet days in the rainy season	This variable is a measure of the number of wet days (a wet day is one receiving ≥0.1 mm precipitation) in the wet season in the year of the survey.	55km
Rainfall in first month of rainy season	This variable is a measure of rainfall (in mm) in the first month of the rainy season in the year of the survey.	55km
Variance of rainfall in first month of rainy season	This variable is a measure of the variance of rainfall in the first month of the rainy season for the period 2010-2020.	55km
Total rainfall in the rainy season	This variable is a measure of total rainfall in the wet season in the year of the survey.	55km

Physical Capital

Although physical capital at community level is a potentially important predictor of income in rural areas, its importance reflects agro-ecological conditions (natural capital and climate), as they shape agricultural profitability. We only include data on irrigation in our analysis as it is both globally available and is expected to moderate the effect of climatic conditions, which are of primary interest to this analysis. Although it would be potentially interesting to include data on road access, as it proxies for access to markets, a different way to manage production shocks, we did not have access to this data measured in a comparable way across countries. Finally, in some countries, as mentioned above, it may be possible to complement this data with information collected by village surveys, conducted simultaneously with household surveys, but its existence or availability is far from common.

Table 7 Physical capital

Variable	Description	Resolution	Year
Area equipped for Irrigation	We use the global dataset of area equipped for irrigation compiled by Siebert et al. (2013). For this process, they relied on two datasets sub-national irrigation statistics from national statistics as well as from international organizations (e.g. FAO and World Bank). To	5 arc-minutes	2005

	identify the geospatial locations of the irrigation schemes irrigation maps of the reports were digitalized. Additionally, information from other sources (e.g. atlases and inventories) were utilized. The data is on the grid cell level. Each grid cell is the share of the total area equipped with irrigation.

Health shocks

We include measures of disease prevalence, as indicative of the likelihood of health shocks that may reduce productivity. , or any unexpected event that has a large- impact on households, businesses and the government is distinct from measures of human, produced and natural capital. Examples are floods, cyclones, extreme heat events, oil spills, outbreaks of disease and terrorist events.

Table 8 Shock variables

Variable	Description	Resolution	Year
Malaria	We use the malaria Stability Index presented in Kiszewski et al. (2005), relying on data of the most important vector mosquito in a region. The projected Index includes the share of human (vs. animal) blood meals of the vector, the average survival time of a vector (in days), the length of the main malaria transmission season as well as the time period how long it takes for a anopheles mosquito to develop parasites after a infested blood meal.	55km	Unclear
Dengue	We use data from the global high-resolution map of dengue transmission intensity developed by Cattarino et al. (2020) by fitting environmentally driven geospatial models to geolocated force of infection estimates derived from cross-sectional serological surveys and routine case surveillance data.	18.5 km	Unclear

5.2 Methods

5.2.1 Vulnerability to poverty

Following the last World Development Report (WDR) on Poverty (World Bank, 2001), vulnerability to poverty is defined as the probability of being poor in the future, where poverty is defined as having an income below a certain threshold (eg, a poverty line). Although much progress has been made since the WDR2001 in the measurement and understanding of the nature of poverty last (eg, its multidimensionality), progress in understanding and quantifying vulnerability has lagged, perhaps reflecting the relatively demanding nature of this concept, namely its prospective and probabilistic nature, on the data.

This report builds on previous empirical applications that addressed these two demands by using past data on income or consumption to infer the *probability* of *future* deprivation, under the assumption that the "production function" underlying income generation is stationary (ie, independent of the time period at which income and other variables are measured). In practice, this means that we will concern ourselves with obtaining estimates of both a household's expected (i.e., mean) consumption and of its variance. The intuition for the need to move beyond expected income should be clear: for example, a salaried public servant with an expected level of consumption roughly similar to that of a farmer may nevertheless be (and feel) much less vulnerable to poverty because of the relative stability of the former's income.

The characterization of the income generation process in terms of mean and variance allows us, in a second stage, to predict the probability that household consumption will be below the poverty line, ie, empirical estimates of

$$vh_t = \Pr\left(ch_{t+1} \leq z\right)$$

where ch_{t+1} is the household's per-capita consumption level at time $t+1$ and z is the appropriate poverty line. Note that the level of vulnerability at time t is defined in terms of the household's consumption prospects at time $t+1$. Throughout this analysis, we set z as equal to 1.90 USD per day per person in 2011 dollars and adjust it to 2019 dollars (1.6735 USD per day) to ensure it aligns with the consumption aggregate, which is also adjusted to 2019 values.

Obtaining estimates of vulnerability to poverty requires consistent estimates of the mean and variance of income. In a regression context, the first step is to characterize household consumption as a function of its observable characteristics, X_h, as:

$$lnc_h = X_h\hat{\beta} + \mu_h \quad (1)$$

Using the estimates $\hat{\beta}$ we are able to directly estimate the expected (log) consumption which, conditional on X_h, is now a deterministic component of the distribution of consumption:

$$\hat{E}[lnc_h|X_h] = X_h\hat{\beta} \quad (2)$$

and the variance of log consumption, conditional on X_h:

$$\hat{V}[lnc_h|X_h] = \hat{\sigma}^2_{\mu,h} = X_h\hat{\theta} \quad (3)$$

for each household h.

Under the usual assumption that consumption is log-normally distributed (i.e., that lnc_h is normally distributed), we can use these estimates to fully characterise the distribution of income and obtain estimate of the probability that a household with characteristics X_h, will be poor, i.e, estimate the household's vulnerability level. Letting Φ(.) denote the cumulative density function of the standard normal distribution, this estimated probability is given by:

$$\hat{v}_h = \widehat{Pr}(\ln c_h < \ln z\,|X_h) = \Phi\left(\frac{\ln z - X_h\hat{\beta}}{\sqrt{X_h\hat{\theta}}}\right) \quad (4)$$

5.2.2 Empirical application

A large literature in production economics, building on Just and Pope (1978, 1979) shows how to analyse the conditional variance of an outcome variable (in our case, income) as a function of observable characteristics of the household. Pritchett et al. (2000) and Chaudhuri et al. (2002) are two early examples of the use of a conceptually similar approach to estimate vulnerability to poverty, with numerous more recent applications (e.g. Novignon et al. (2012), Imai et al. (2015), Cahyadi and Waibel (2016), Sharaunga et al. (2016) and Azeem et al. (2018)).

In a first step, this approach requires that we estimate the conditional mean of income through a regression of the type:

$$ln\ C_{ij}\ =\ \beta_0 +\ \beta_1 X_i +\ \beta_2 C_i +\ \beta_3 S_j + \beta_4 I_j +\ \varepsilon_{ij} \quad (5)$$

where C_i is the per capita consumption of household i living in community j, X_i are natural capital variables (including biodiversity and forest cover) C_i are climatic variables and S_j are vectors of communal shock variables (e.g. dengue and malaria), respectively, and I_j is a vector of community physical capital variables (e.g., irrigation). The effect of observable characteristics on consumption are reflected on our estimates of the parameters β, while ε_i is an idiosyncratic error term that captures the unobserved determinants of consumption.

In a next step, we estimate a regression of the variance of consumption on covariates (typically, but not necessarily, the same as in equation (5)):

$$\hat{\varepsilon}_{ij}^{\,2} =\ (ln\ C_{ij}\ -\ E(Ln\ C_{ij}|X_i, S_i, S_j, I_j))^2\ =\ \theta_0 +\ \theta_1 X_i +\ \theta_2 C_i +\ \theta_3 S_j + \theta_4 I_j +\ v_{ij} \quad (6)$$

The estimated θ parameters allow us to quantify the main correlates of income risk and, in conjunction with the estimates of their effect on mean income, obtain estimates of vulnerability to poverty. In a last step, we re-estimate equation (5) using the estimated weights from the second regression to adjust for heteroskedasticity. The weights are the absolute values of the *m*-th root of the fitted values of equation (6).

Empirically, there are three central choices in modelling and interpreting the income generation process using this approach. The first is how to account for fundamental differences in the "production technology". The inclusion of variables, for example measures of soil quality, which act as "technology shifters" that contribute linearly to income generation is the simplest and standard way to relax the assumption of a homogeneous, common, income generation process. An alternative, which we explore in more detail in the next section, is to hypothesize different production functions that are optimised for different and specific values of the technology shifter, with both the variables and its critical threshold being an empirical question – ie, letting the data "tell" what the best description of the production technology is.

The second question concerns the interpretation of the estimates from equations (5) and (6), which reflects the set of explanatory variables included in our estimates. Many of earlier applications rely, like us, on the use of cross-sectional data, with limitations that are well known. First, vulnerability is a dynamic poverty concept and an analysis of cross-sectional data may not adequately capture changes in poverty over time. Against this criticism, Chaudhuri et al. (2002) argued that estimates using a large cross-sectional data set covering a sufficiently large variation in consumption and observable household characteristics can be a good proxy of poverty dynamics estimates. Second, the suspicion that we are leaving out of the estimation of equations (5)-(6) relevant explanatory variables. Although we have a rich set of both household and environmental variables that may explain individual wellbeing, we cannot credibly claim to include all variables that may matter to explain income. In particular, and as mentioned above, several datasets do not include information on the full set of public services (roads, health services, etc) that we would like to account for. This omission naturally affects the interpretation of those variables for which we consistently have information across surveys and that we are able to include when estimating equations (5)-(6). Hence, we interpret these estimates as predictors of vulnerability, which has implications for the type of policy implications that can be drived from our analysis. We discuss this issue in more detail in section 7.

A third, and final, question relates with the relatively high correlation between different included explanatory variables. Such multicollinearity has consequences for the magnitude and precision of the OLS estimates (typically very large, sometimes with counter-intuitive directions of the effect). Several solutions exist to this problem, among them the use of the Least Absolute Shrinkage and Selection Operator (LASSO) regression (Tibshirani, 1996). The LASSO minimises the residual sum of squares subject

to the sum of the absolute value of the coefficients being less than a constant. Given this constraint, some estimates are exactly zero (ie, they are effectively dropped from the results). Although the resulting models are easier to interpret and typically exhibit better predictive behaviour, this approach effectively increases the bias of the fitted model.

Finally, one comment on inference. The household survey data is typically clustered at the village level. Due to similarities between households in villages, estimates of variance in a clustered sample underestimate the true variance, requiring the use of clustered standard errors.

5.2.3 Creating homogeneous cohorts

Implicit in the above framework is the assumption that, conditional on observable characteristics *X*, all observations fit the same income "production function". One way to avoid such a strong assumption is to use approaches such as regression trees to identify the hierarchy of importance of different constraints and, in particular, the possibility of thresholds in this relation. Equation (5) can then be rewritten as

$$ln\ C_{ij} = \mathrm{f1}\left(X_i, S_i, S_j, I_j\right)\ if\ W \geq w0 \qquad (7.1)$$

$$ln\ C_{ij} = \mathrm{f2}\left(X_i, S_i, S_j, I_j\right)\ otherwise \qquad (7.2)$$

where depending on whether a specific variable *W* is above or below a certain cut-off (w_0) implies that the effect of other variables *(X, S, I)* is better expressed by function f_1 or f_2, respectively, rather than a common function as in equation (5). The selection of variables *W* and associated threshold levels, w_0, leads to the identification of a hierarchy of importance of those variables in predicting income.

In the empirical application, the set of *W* will include those variables that most closely measure the effect of humans on the environment (climate, biodiversity), while also accounting for variables that may moderate that effect (ruggedness, access to irrigation). This choice implies that we do not include variables that measure human investment in infrastructure, human or physical capital, as they likely reflect environmental conditions (eg, R&D and associated extension services are directed to more productive agro-ecosystems while households react to the changes in production possibilities embodied in those new technologies by investing in education and agricultural assets). A characterisation of the income of rural households as a function of their physical environment is therefore useful to understand poverty, and consequently the utility and the need of social protection programs even if, as it should be clear, the variables that potentially split the sample are typically not the focus of social protection policies, either as safety or cargo nets.

A large (and increasing) number of statistical approaches, under the label of machine learning, aim to capture the basic intuition underlying equations (7.1) and (7.2): that it is better (in a predictive sense) to account for heterogeneity rather than assume homogeneity. This improvement in predictive power comes at the cost of increased complexity (the model captured by these equations is less parsimonious than the one described by equation (5)), which needs to be (negatively) weighted against the gain in predictive accuracy.

Generally speaking, machine learning covers algorithmic approaches to predicting outcomes (e.g. consumption) based on a number of variables (e.g. stocks of physical, natural, and human capital). Machine learning methods aim to produce the best predictions by finding a balance between bias (the fit of the model to the data) and variance (the fit of the model to other data). As such, machine learning algorithms are typically trained on a subset of data and then tested on the remaining data.

There are a range of machine learning models available. Beyond ordinary least squares, we can consider decision trees and linear trees (the intuition of which was briefly presented above), gradient boosted trees, random forests, and neural networks. Each

model differs in terms of complexity and computational requirements. To select the best model, we calculate the R squared for each model on test data to compare performance (see section 6).

Random forests

We focus on the Random Forest (RF) algorithm to construct different cohorts as it was the best predictor of consumption within a feasible running time. A RF generates multiple decision trees, each constructed by minimising the Gini index, and where each decision tree considers only a random subset of the data, leading to a different set of parameters that are individually biased. The final model is determined by an average voting scheme among individual trees (Kim and Kim, 2022). As our outcome variable is a continuous variable, each decision tree in the random forest is a regression tree.

Figure 1 Random forest estimators

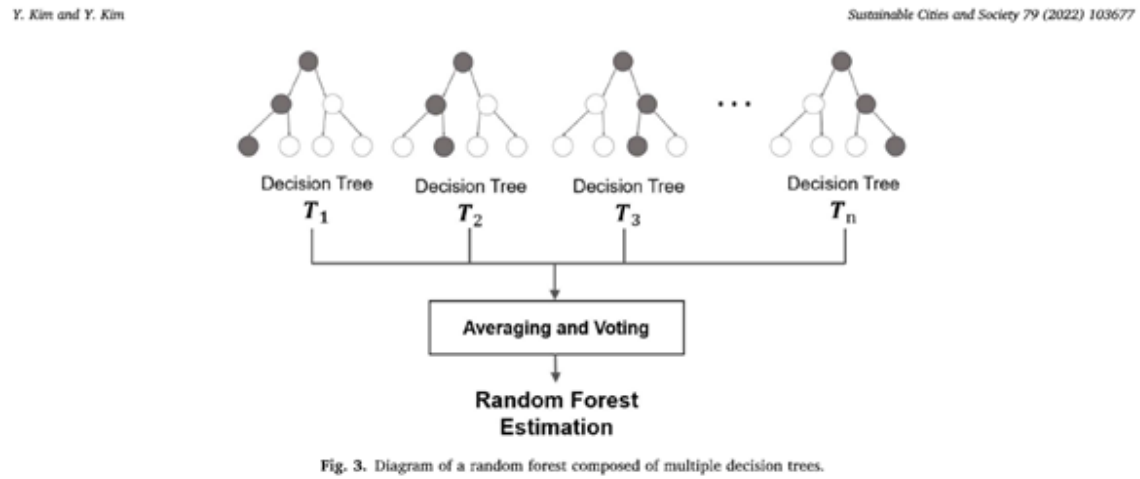
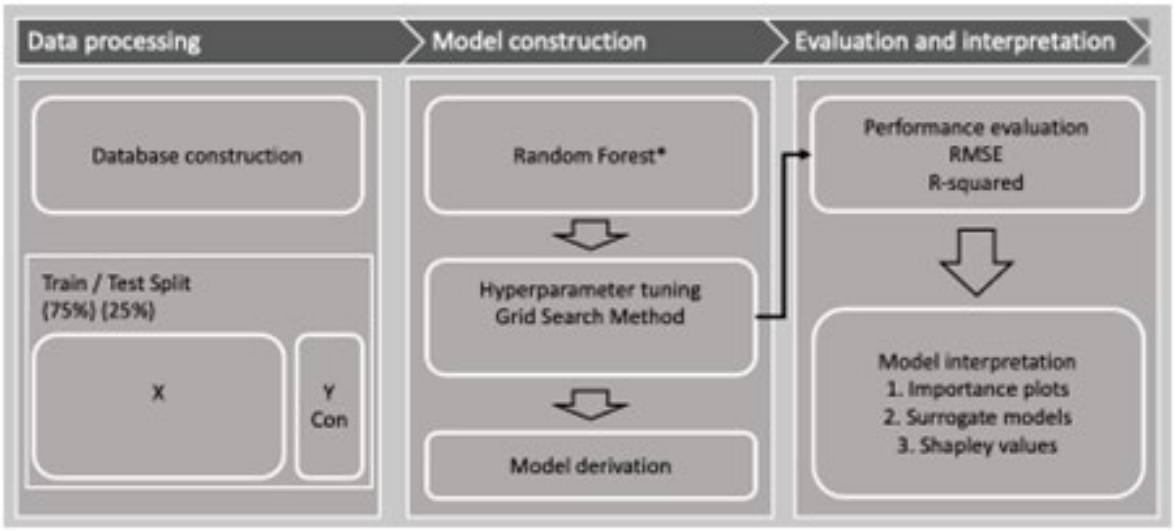


Fig. 3. Diagram of a random forest composed of multiple decision trees.

The development of the RF model (overview provided in Figure 2) was performed using Ranger and Tidy Models packages in R. The data was split into a training and a test set, with the training set consisting of 60% of the sample, and the test set consisting of the remaining 40% of the sample. We tuned a set of hyperparameters to get the best random forest model, including the number of predictors that will be randomly sampled at each split and the minimum number of data points in a node that is required for the node to split further. The number of trees contained in the random forest was set to 1000.

Figure 2 Model development



Note: we also explored other machine learning techniques – see appendix 9.2.

To tune the hyperparameters in random forest, we split the data into 10 folds of equal size. We then computed a set of performance metrics (RMSE and Rsquared) for the set of tuning parameters across the 10 resamples of the data. To do this, we specified a grid with tuning combinations of number of predictors and number of data points in a node. We then created a refined grid based on the best performing sections of the grid. For example, we found that the number of predictors were best between 0 and 5, and number of data points in a node were best between 30 and 40. We calculated the performance metrics for the refined grid. The best model is when number of predictors is 1 and number of data points in a node is 31. The chosen model has a R-squared of .36 and a RMSE of .52 for the test data.

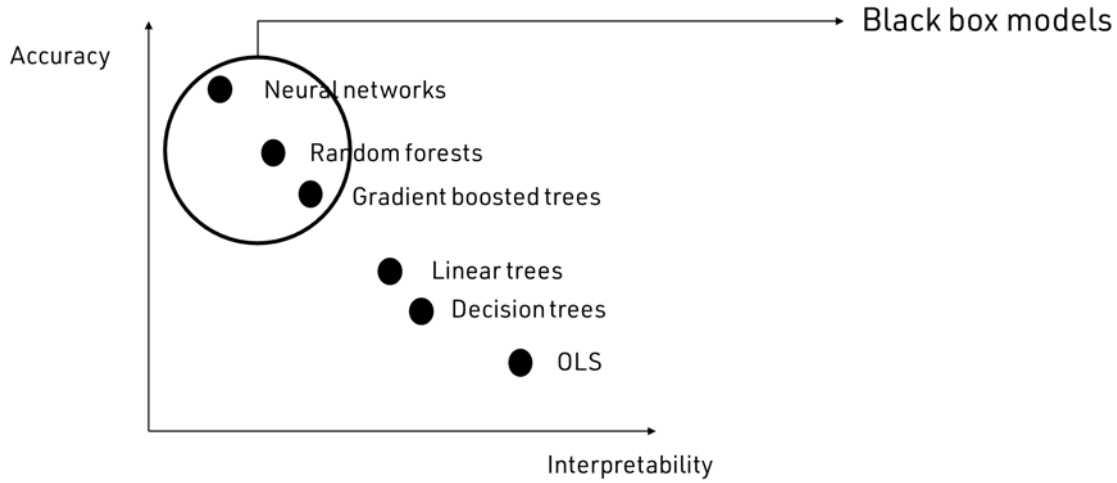
Interpretability

We define an interpretable technique as any that results in a model that humans can understand in terms of the reasoning behind the predictions and decisions made by the model. Unfortunately, there is a trade-off between accuracy and interpretability, as illustrated in Figure 3.

The simplest models, for example OLS and Decision trees are extremely interpretable, meaning we can “read” their outputs and quickly understand its main message. For example, the regression models defined in section 5.2.2 can be considered interpretable as it is possible to predict the value of the dependent variable for any set of independent variable values, as the model outputs include the coefficients for each variable and the structure of the relationship is known. As a result, we can understand which variables are important in making the prediction, as we can judge their relative magnitude and direction.

However, and as mentioned above, these interpretable models can sometimes lack accuracy: their simplicity comes at the cost of higher bias or variance. Other models, such as gradient boosted trees and random forest, have been shown to be more accurate but this improvement comes at the cost of an increase difficulty in interpreting their results (hence the label “black boxes”).

Figure 3 Black box machine learning algorithms



Several techniques have been developed to make the black box models interpretable, ie, to obtain some understanding of the model and the relationships between input and output variables.

The first approach we use is the quantification of variable importance scores, which provide an indication of a variables importance in making a prediction and are calculated by removing the variable from the model and observing the change in predictive accuracy (ie the error term). A large increase in the error term (large reduction in prediction

accuracy) means that the variable is important in making the prediction. In other words, it is a measure of how much the accuracy of the model decreases when a variable is removed. When the “black box” model is a Random Forest, importance scores are computed by averaging the difference in prediction error when a variable is included compared to when it is excluded across each of the trees in the model.

One of the shortcomings of measures of variable importance is that they cannot be easily interpreted, even with respect to the direction of their contribution to prediction the outcome. Additive Shapley Values overcome that gap.

Unlike variable importance scores, which are a global approach to interpreting variables – they describe average behaviour of a machine learning model – Shapley values are a local approach to interpreting variables as they reflect the empirical estimate of the contribution of each variable to the prediction made. A Shapley value represents the average marginal contribution of the variable to the prediction made for one observation: for example, in the case of three variables (A, B and C), calculating the Shapley value for variable A would involve estimating its effect on prediction for each subset of variables (ie, subgroup (A only), subGroup(A and B), subGroup(A and C) and subGroup(A, B and C), and then averaging its marginal contribution to the prediction across all possible subgroups of variables.

Because Shapley values are calculated for each observation in the sample their number is equal to the number of observations in the sample. By averaging individual Shapley values across all observations, we can then get a global measure of variable importance.

A final, complementary approach to interpreting variables is the global surrogate model. The model, which we estimate, is trained to approximate the predictions of the underlying black box model as accurately as possible while being interpretable (https://christophm.github.io/interpretable-ml-book/global.html).

We selected a regression tree as our interpretable surrogate model. The intuition of this approach was illustrated above: regressions trees are a type of decision tree model that split data according to cut-off values for different variables and generate predictions based on these splits and the different subsets of data they create. These splits occur where the sum of squared errors across variables is minimised. The final subsets each observation ends up in are the terminal or leaf nodes.

One way to measure how well the surrogate replicates the black box model is by calculating the R-squared measure:

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \sum_{i=1}^{n} \frac{(\hat{y}_*^{(i)} - \hat{y}^{(i)})^2}{(\hat{y}^{(i)} - \check{y})^2}$$

Where $\hat{y}_*^{(i)}$ is the prediction for the i-th instance of the surrogate model, $\hat{y}^{(i)}$ the prediction of the black box model and $\check{y}$ is the mean of the black box model predictions, while SSE stands for sum of squares error and SST for sum of squares total. The R-squared measure can be interpreted as the percentage of variance in the predictions that is captured by the surrogate model. If R-squared is close to 1 (= low SSE), then the interpretable model approximates the behaviour of the black box model very well, in which case we may replace the complex model with the interpretable model. If the R-squared is close to 0 (= high SSE), then the interpretable model fails to explain the black box model. The interpretation of the surrogate model becomes irrelevant if the black box model is bad, because then the black box model itself is irrelevant.

6 Results

We present three main sets of results. The first provides an analysis of the importance of environmental variables as predictors of consumption, using random forests to explore the underlying heterogeneity in the income generating process. The results are made interpretable using variable importance scores, Shapley values and surrogate models.

In the second set of results we describe the clusters formed by the surrogate model, including differences in their vulnerability to poverty. The third set of results presents the lasso estimates of the effect of each of selected predictors of vulnerability to poverty for each of the clusters defined by the surrogate model.

6.1 Environment and poverty in the Anthropocene: a machine learning approach

We use machine learning algorithms to predict consumption based on a set of environmental variables: erosion, forest cover, ruggedness, biodiversity integrity, slope, soil water capacity, soil depth, elevation, importance of irrigation (in %), exposure to health shocks (dengue and malaria) and several climatic variables (temperature in first month of rainy season and its variance of temperature over 2010-20, number of wet days in first month of rainy season and its variance over 2010-20, number of wet days in rainy season, precipitation in first month of rainy season, precipitation in the rainy season). We estimate 3 different machine learning models, of which two were interpretable and one is a black box models. Each model was estimated on a training set (60 percent of the sample, used to tune the different parameters of each model, as discussed above) and its accuracy in terms of key performance metrics such as Rsquared and the Root Mean Squared Error (RSME) evaluated in a test set (40 percent of the sample).

The performance metrics for each of the machine learning algorithms is presented in Table 9. The best performing model was the random forest and, in the rest of this report, we will present the results from using this model to estimate vulnerability to poverty, and to interpret predictors of vulnerability.

Table 9 Rsquared and Root Mean Squared Error, Random Forest Specification

	RSquared	RMSE
OLS	0.22	0.57
Regression tree	0.19	0.58
Random Forest	0.34	0.51

6.1.1 Interpreting predictors of poverty

Random forest algorithms are complex and are not interpretable (when compared, for example, with OLS regression or regression trees), hence they do a poor job at identifying the main predictors of poverty. As mentioned in the previous section, we use three approaches to understanding which environmental variables are most important in predicting consumption: variable importance scores, additive Shapley values, and surrogate models.

Variable importance scores

The variable importance scores, defined in the previous section, are presented in Table 11, which ranks the ten most important variables in terms of predictive power. The main conclusion is that the most important variables in predicting income are climatic conditions in the first month of the rainy season and, importantly, two of the top variables (variance of temperature and variance of number of wet days, measured over 2010-20) are measures of climatic risk, rather than weather realizations in the year of the survey. Other variables (soil properties, forest cover) are much less important.

Table 10 Top ten variables in terms of their variable importance score

Variable	Importance Score
Variance of temperature in first month of rainy season	3,753
Dengue	2,768
Number of wet days in first month of the rainy season	2,670
Rainfall in first month of the rainy season	2,277
Temperature in first month of rainy season	1,860
Variance of number wet days in first month of rainy season	1,849
Forest Cover	1,731
Mean soil depth	1,680
Slope	1,455
Elevation	1,443

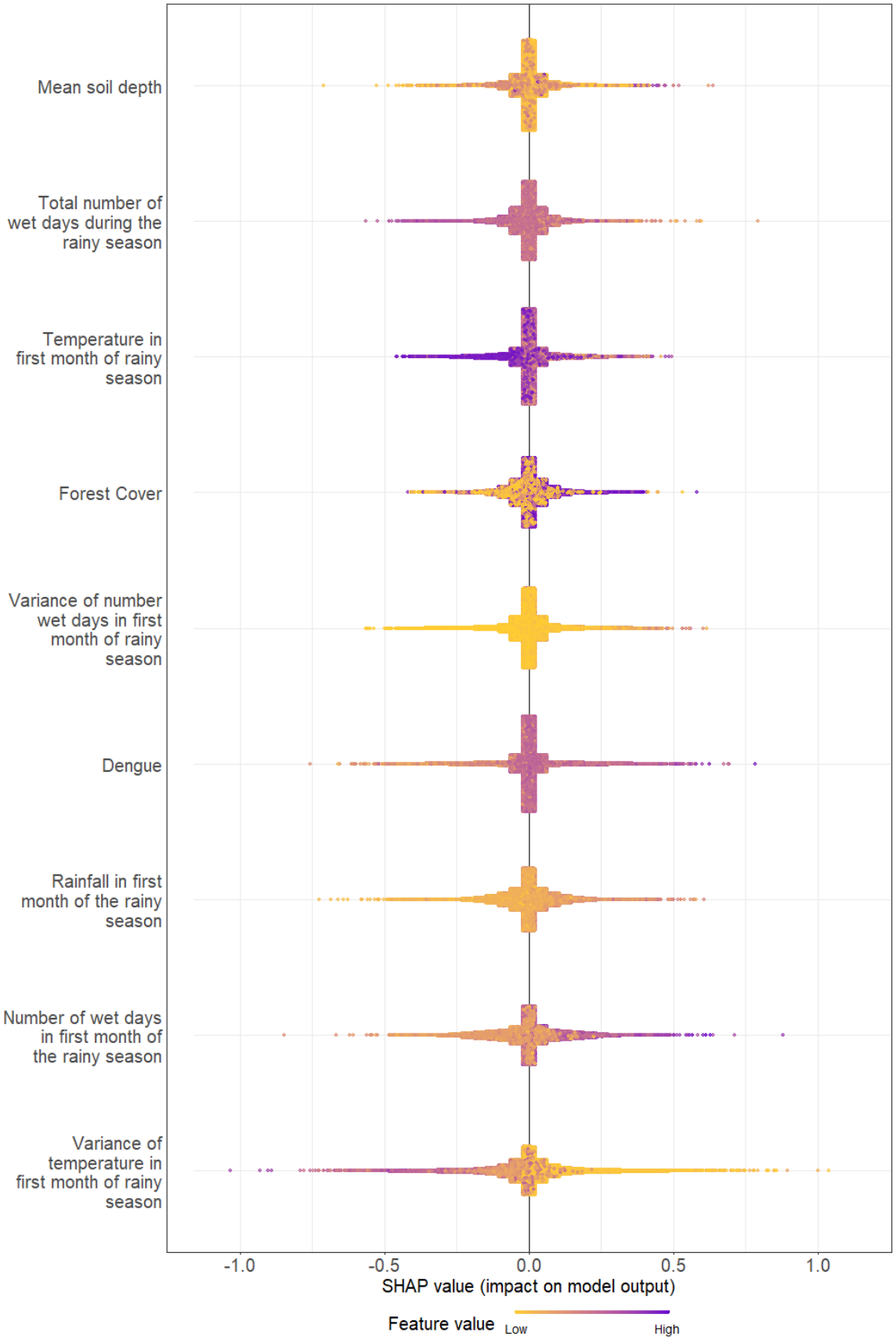
Additive Shapley values

Figure 4 presents additive Shapley values for the 10 most important variables, as ranked by their global Shapley value.

Because Shapley values are plotted for every observation in the dataset, we have a distribution of estimates of the effect of each variable, which are either positive or negative (as revealed by their position with respect to the axis at 0) and either high (in purple) or low (in yellow). The negative importance of variance of temperature in first month of rainy season (ie, a negative effect of climatic risk) reflects a larger frequency of individual high estimates (in purple) with a negative sign (ie, to the left of the 0-axis).

Figure 4 allows us some conclusions. In addition to the negative effect of variability of temperature in the first month of the rainy season, just used as an example to facilitate the interpretation of the results, high values of the number of wet days and precipitation in the first month of the rainy season contribute positively to consumption. Summarizing, the ranking of variables following the Shapley values conveys the same message as the ranking provided in Table 10: in particular, the first five top variables reflect the importance of weather in the first month of the rainy season, and climatic risk (as measured by the variance of the two climatic variables listed above) remains important. The added value of this approach is that we now have an indication of the direction of the effect of each variable.

Figure 4 Additive Shapley values (top 10 variables, from least (top) to most important (bottom))



Surrogate models

The final approach used to interpret the random forest model is the surrogate model. As explained in the previous section, this approach involves using the random forest prediction of consumption (rather than observed consumption) as the outcome variable, and then estimating an interpretable model which, in our case, is a regression tree which separates data into a number of cohorts based on different conditions. The results of this approach are presented in Figure 5.

The algorithm splits the sample by minimising a loss function – in this case, minimising the root of the sum of squared errors (RSME) – and continues to partition the sample until there is no further improvement in predictive power that is large enough to more than compensate for the added complexity of the tree. Starting at the node at the top of the tree, the sample is continuously split by different conditions until reaching a leaf node (the nodes at the bottom of the tree). The estimates of consumption for each leaf node are the average of subsample.

The performance of the surrogate model (decision tree) relative to the black box model (regression forest) is relatively high (see Table 11). This result suggests that the surrogate model fits the random forest reasonably well, rendering the interpretation of the splits in the decision tree presented in Figure 5 potentially informative.

Table 11 surrogate model: parameterization and fit

	Parameters	R^2
Surrogate	cost/complexity parameter = .01 max depth of tree = 30 min number of data points for it to be split further = 30	0.55

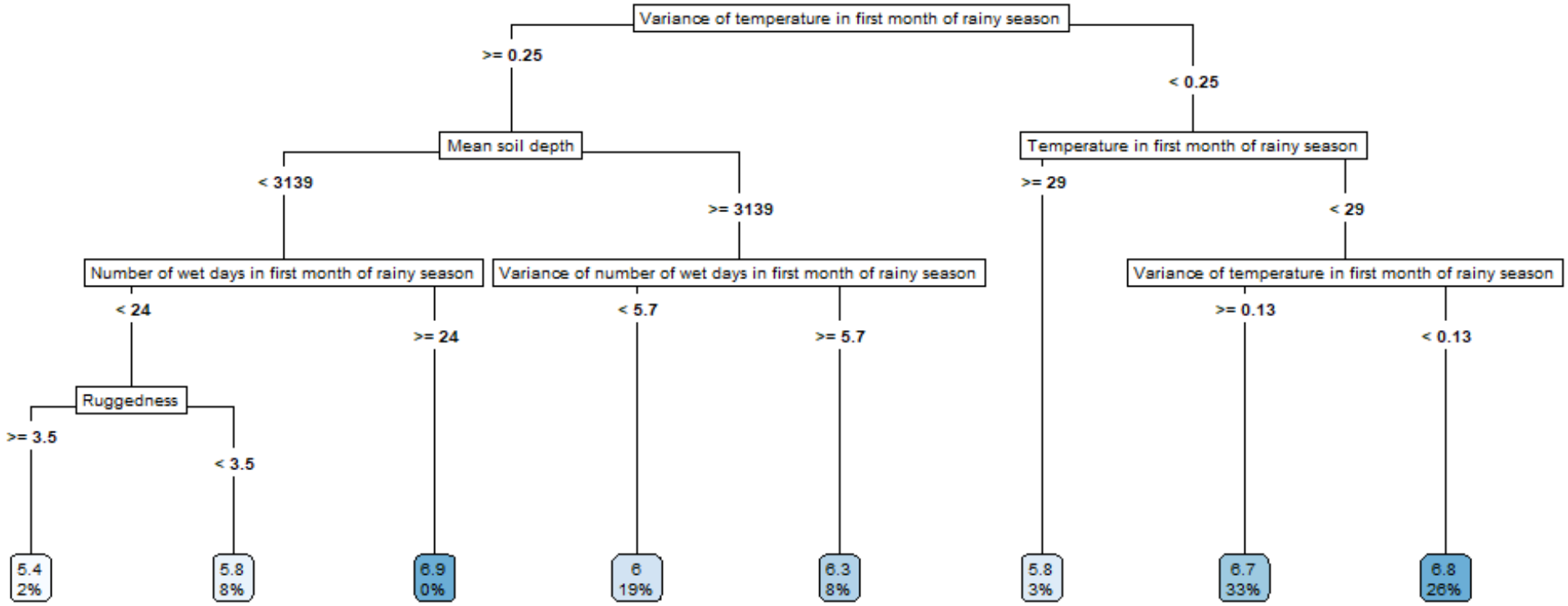
The relative importance of different predictors is largely in line with the other two approaches. The full sample is first split into two groups as a function of the variance of temperature in the first month of the rainy season. Observations in areas with values of this variable greater than or equal to 0.25 (ie, higher climatic risk) form one group (39 percent of the sample) that will have, on average, lower income than those that exhibit lower values of the split variable, who will form a different group (61 percent of the sample). This logic can be followed until the leaf nodes are reached.

Turning our attention to the leaf nodes, we can make the following observations: First, the regression tree is successful in creating groups with meaningful differences in consumption: the mean consumption per capita per year in the poorest group is 270 USD (=e^(5.6)) which is less than 1/3 of the average consumption of the better-off group, at almost 900 USD (=e^(6.8)). Second, production conditions in the first month of the rainy season seem to matter most among the weather conditions, either in terms of temperature (and its variance) or number of wet days (and its variance). It appears that our findings from the Shapley values and also the original decision tree are corroborated here.

Third, the two better-off groups, which together include approximately 60% of the sample and are the only ones that are, on average, above the poverty line, are characterized by a small number of splits: low risk in terms of temperature, and relatively low temperatures. On the contrary, average consumption below poverty can be associated with a diversity of paths/splits although, in most cases, climatic risk (high variability in terms of temperature in the first month of the rainy season) seems to be the common characteristic.

.

Figure 5 Predicting income: surrogate model, using a regression tree



6.2 Characterising income groups

The surrogate model created eight groups (corresponding to the final leaf nodes). Figure 6 shows their distribution across the seven Southeast Asian countries included in the analysis, with lighter tones of blue denoting higher values of mean consumption per capita, while Table 12 presents some descriptive statistics for each of the groups.

The first conclusion is that while poverty is substantially different across groups, its prevalence is reduced in a linear way with increases in income. Although care must be placed when interpreting regression results when units are substantially different (and, by construction, these are substantively different units), this result is supportive of earlier analyses that claim that agricultural growth has “special powers” in poverty reduction.

In addition to climate and ruggedness (which split the observations into homogeneous groups, and as such are expected to differ between groups), the analysis of Table 12 allows us two additional conclusions. The first conclusion is that biodiversity degradation and forest cover seems to follow an inverse-U relation with income, they begin quite high in the lower groups and then decrease before rising again in the latter groups. Interestingly, G1 is quite high in terms of slope and elevation and ruggedness compared to Groups 6 and 7, which may be contribute to the differences in income between the groups, despite otherwise similar characteristics.

Secondly, in terms of the human capital, there is no distinct pattern with respect to the age of the household head, nor household size. Similarly, in terms of physical capital we are not able to identify a clear distinct pattern in terms of the housing index or the productive asset index.

Figure 7 shows the vulnerability profile for households in each of the eight cohorts. Along the y axis is the percentage of total observations in the sample, and along the x axis is the probability that the household will fall below the poverty line. We can compare the vulnerability profile of cohort 3 to other cohorts, for example cohort 2 and cohort 4.

There are different distributions of vulnerability to poverty for each cohort. In our analysis our focus is on how we can shift the probability distribution, and how the distribution is influenced by shocks, and then developing social protection which keeps them from falling below the poverty line, and ultimately changing the vulnerability profile of these groups.

These distributions can be changed by increasing consumption. For example, to move from the cohort 1 to cohort 2, we need to change mean consumption by approximately 72 USD, and to move from C4 to C5, we need to change mean consumption by approximately 123 USD.

We can inspect the surrogate model, and observe key threshold. If we move a household from one side of the threshold to the other, we may see a different vulnerability profile. Unfortunately, those variables are fixed exogenous factors. Thus, the story may be trying to shift the household to the left of the distribution through social protection programs.

Figure 6 Spatial distributon of consumption groups, as defined by the surrogate regression tree

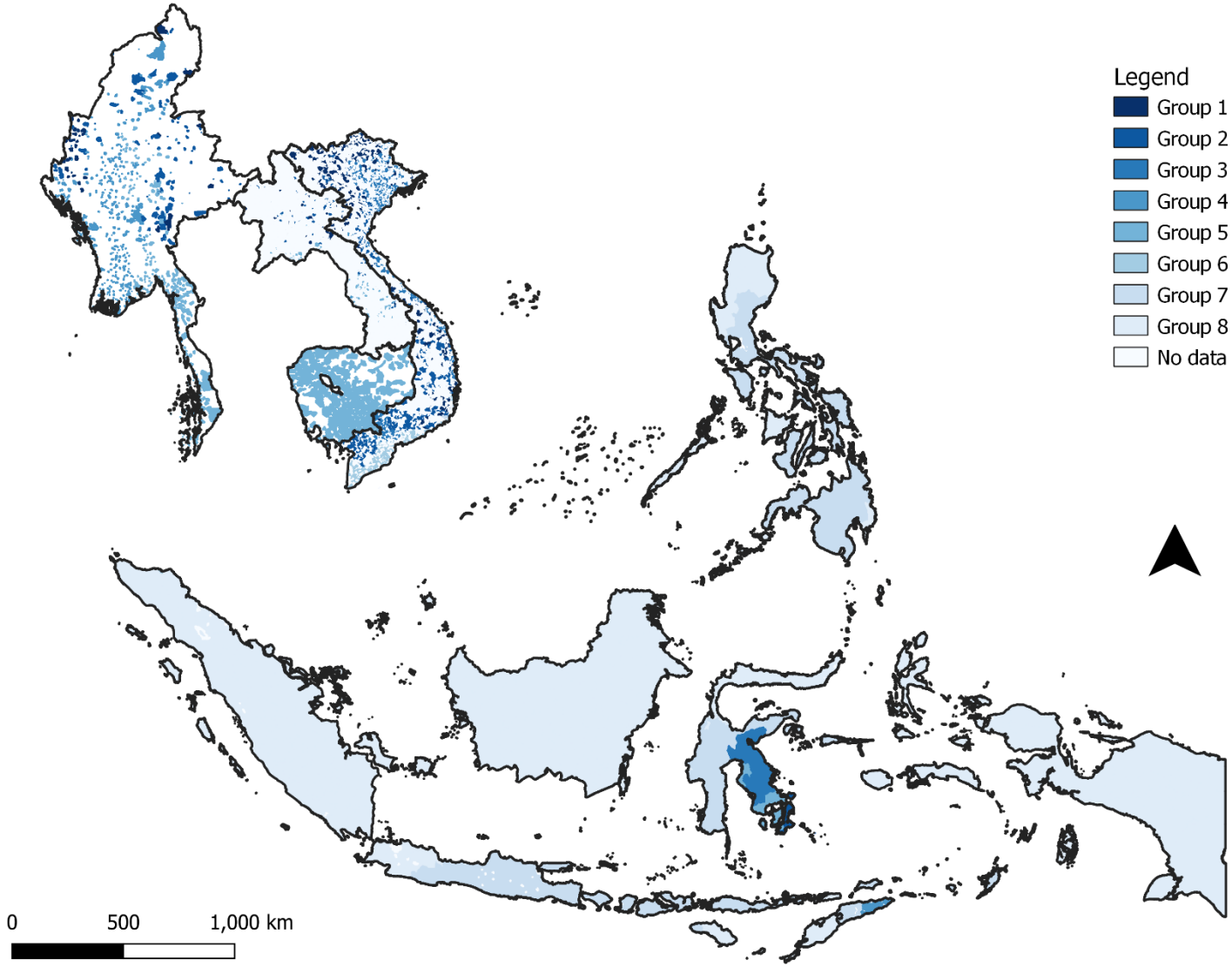
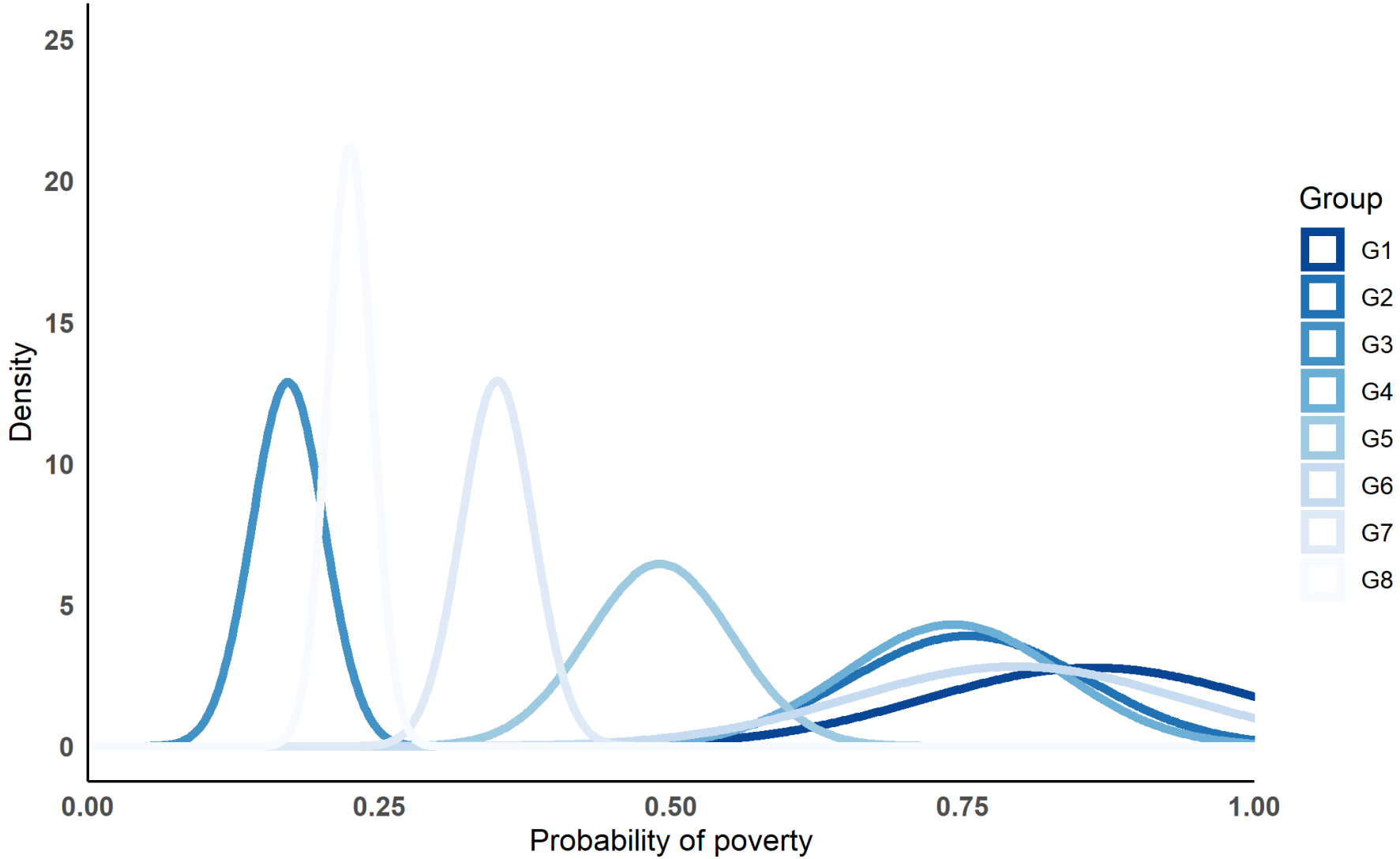


Table 12 Characterizing income groups

Variable \ group	G1	G2	G3	G4	G5	G6	G7	G8
Node in surrogate model (left to right)	1	2	6	4	5	7	8	3
Households (1000)	1,726	6,366	2,079	14,672	6,380	24,599	19,385	229
Proportion	2%	8%	3%	19%	8%	33%	26%	0%
Consumption per capita (USD 2019)	230	337	337	412	561	773	915	1012
Poverty rate	0.91	0.83	0.83	0.76	0.61	0.36	0.22	0.19
Age household head	46.06	51.54	54.93	52.67	49.35	50.90	48.13	46.70
Household size	4.76	3.94	3.66	3.83	4.39	4.00	3.95	4.12
Household asset index	-1.08	-0.22	-0.18	0.02	-0.05	-0.11	-0.14	-0.32
Productive asset index	-0.63	0.18	-0.16	0.38	-0.02	-0.22	-0.21	-0.08
Elevation (masl)	786.21	371.40	23.10	82.95	119.65	306.91	308.88	536.15
Ruggedness	4.44	1.52	0.17	0.25	0.31	1.33	1.02	1.69
Slope	12.29	4.11	0.45	0.67	0.81	3.57	2.75	4.50
Erosion	27.46	21.38	14.11	13.43	7.66	20.25	11.42	2.96
Mean soil depth (cm)	1,278.59	1,867.62	8,454.54	7,952.31	9,950.43	2,770.45	2,134.91	1,928.82
Soil water capacity	248.83	242.04	227.40	232.04	230.78	235.23	231.90	235.64
Proportion irrigated	2.97	10.66	34.74	30.68	11.28	12.96	9.87	2.56
Dengue	0.01	0.01	0.03	0.02	0.02	0.03	0.03	0.02
Malaria	1.87	3.06	9.47	6.52	7.43	2.15	3.23	2.19
Forest cover	57.62	36.52	9.46	7.48	12.92	45.19	59.08	82.97
Biodiversity integrity	0.99	0.92	0.84	0.82	0.85	0.88	0.91	1.00

Figure 7 Vulnerability to poverty for different cohorts



6.3 Predicting vulnerability to poverty

We use LASSO regression to identify the best predictors of the different moments of the distribution of income (and, consequently, vulnerability to poverty) in each of the eight groups identified in the previous section. Table 15 shows the results when the outcome variable is mean consumption and Table 16 shows the results when the outcome variable is the variance of consumption. The explanatory variables are normalised to get a sense of their relative magnitude.

Table 12 Summary of variables by moments of the distribution

		Variance			
		Negative (low risk)	Inconclusive	Positive (high risk)	Total
Mean	Positive	0	2	3	5
	Inconclusive	1	3	1	5
	Negative	4	2	2	8
	Total	5	7	6	18

Table 13 Summary of variables by moments of the distribution

		Variance		
		Negative (low risk)	Inconclusive	Positive (higher risk)
Mean	Positive (higher expected income)		Variance in the number of wet days in the first month of the rainy season Rainfall in the first month of the rainy season	Forest cover Soil Depth Number of wet days in the first month of the rainy season
	Inconclusive	Elevation	Erosion Irrigation Rainfall in the rainy season	Slope
	Negative	Biodiversity Available Water Capacity Temperature first month of rainy season Number of wet days in the wet season	Ruggedness Malaria	Dengue Variance of temperature in the first month of the rainy season

6.3.1 Mean consumption

Table 15 shows output from 8 individual lasso regressions for the different cohorts where mean consumption is the outcome variable. The results show that the size and the magnitude of the effect can be different for each variable across each of the groups. 4 out of the 10 variables have the same direction across all 8 groups meaning the remaining 6

variables have varied direction across the 8 groups. This suggests that the cohorts may respond differently to different interventions.

In terms of the variables that are consistent in direction across all groups, the interpretation is that older household heads and bigger household sizes are related with lower mean consumption, university education and productive assets are related with higher mean consumption. These results are consistent with an economic theory, that better inputs (produced and human capital) will produce better outputs (income and consumption).

In terms of the variables that are inconsistent in direction across all groups, female household head and erosion are perhaps the hardest to understand. The coefficients of primary and secondary education dummies are more or less consistent with the coefficient of the university dummy. The general findings from the environmental variables are that biodiversity has a positive effect on mean consumption (a mechanism for resilient production) and forest cover has a negative effect on mean consumption because for small holder farmers there is less agricultural land available.

6.3.2 Variance of consumption

Table 16 shows output from 8 individual lasso regressions for the different cohorts where variance of consumption is the outcome variable. The results show that the size and the magnitude of the effect can be different for each variable across each of the groups. 2 out of the 10 variables have the same direction across all 8 groups meaning the remaining 8 variables have varied direction across the 8 groups.

In terms of the variables that are consistent in direction across all groups, the interpretation is that larger household sizes are related with lower variance in consumption, and higher productive assets are related with higher variance in consumption. These results suggest that larger households can support each other to reduce the variance of income, but supporting each other comes at a cost for some of the members, as shown in the mean consumption estimates. Higher productive assets are likely to enable farmers to increase their income when the conditions are right, but bad investment decisions can cripple a smallholder farmer.

In terms of the variables that are inconsistent in direction across all groups, female household head appears to reduce variance of consumption in the really poor cohorts (group 1 and group 2), but is then related with increased variance of consumption in the other groups. The general finding from education is that higher education increases the variance in income. For the environmental variables for the most part forest cover reduces variance of income and biodiversity increases the variance of income.

There appears to be a fine line between risk and return here – some variables give farmers the opportunity to get higher income (which we want) but there needs to be some form of protection so they can take these risks, because sometimes they won't come off.

Table 14 Mean consumption: LASSO regression

name	G1	G2	G3	G4	G5	G6	G7	G8	+	-	Direction
(Intercept)	224.89	797.01	510.76	613.75	933.59	2,816.68	977.09	1,025.06			
Erosion	2.09	-14.46	6.87	X	13.89	X	-25.11	-0.78	3	3	?
Forest cover	-8.22	28.95	-18.48	X	90.19	X	79.67	31.29	4	2	+
Ruggedness	-30.08	X	X	X	X	-1.95	-599.71	-106.86	0	4	-
Biodiversity integrity	-67.13	-16.72	-25.00	-4.46	-86.19	X	-66.56	42.73	1	6	-
Slope	X	-2.21	-14.13	X	20.73	X	624.50	-8.12	2	3	?
Soil water capacity	3.14	-26.63	-9.19	-15.76	-39.48	-13.66	-58.17	-14.83	1	7	-
Soil depth	3.61	19.29	3.30	14.13	86.00	X	73.01	-32.49	6	1	+
Elevation	17.37	-14.00	12.11	-3.24	-40.88	X	-66.76	69.05	3	4	?
Percent irrigated	20.16	-11.86	-13.17	-0.12	1.85	-13.78	48.81	48.86	4	4	?
Dengue	14.92	-31.67	X	36.50	-7.43	X	-52.63	-17.22	2	4	-
Malaria	-30.96	-24.63	X	-3.61	-13.13	X	-50.07	31.72	1	5	-
Temperature in first month of rainy season	29.88	-13.83	-38.24	-20.16	-64.35	X	12.49	-21.92	2	5	-
Variance of temperature in first month of rainy season	23.04	-87.94	-26.27	-25.21	-20.26	-13.98	201.86	-30.23	2	6	-
Number of wet days in first month of rainy season	-303.38	X	31.51	14.35	430.46	X	215.16	-9.07	4	2	+
Number of wet days in rainy season	158.42	-24.02	-31.48	X	-230.27	-96.36	-184.51	-21.57	1	6	-
Variance of Number of wet days in first month of rainy season	65.83	220.58	-111.70	-96.79	19.57	2,825.38	-138.96	44.83	5	3	+
Precipitation in first month of rainy season	92.58	49.96	X	32.00	-122.43	54.28	37.16	-27.09	5	2	+
Precipitation in the rainy season	-19.47	X	-32.60	-15.76	16.32	X	58.41	106.45	3	3	?

Note: X= excluded variable; + = positive effect, - = negative effect, ? = inconclusive

Table 15 Variance of consumption: LASSO regression

name	G1	G2	G3	G4	G5	G6	G7	G8	+	-	Direction
(Intercept)	149,758	237,733	82,688	87,797	504,737	235,185	565,345	613,363			
Erosion	X	X	X	X	X	X	X	X	0	0	0
Forest cover	X	X	X	X	21,929	X	138,595	X	2	0	+
Ruggedness	X	X	155,787	X	37,565	X	-54,242	-3,695	2	2	?
Biodiversity integrity	X	-28,680	X	X	-24,824	X	-35,334	46,964	1	3	-
Slope	X	X	X	12,564	56,391	X	59,330	X	3	0	+
Soil water capacity	X	X	X	-8,563	-54,523	X	-26,491	26,970	1	3	-
Soil depth	X	X	X	12,434	53,138	X	117,316	X	3	0	+
Elevation	X	X	-112,757	8,279	-138,687	X	-35,471	X	1	3	-
Percent irrigated	X	X	X	X	-818	X	21,151	-12,988	1	2	?
Dengue	X	X	X	8,889	18,594	X	-68,306	51,742	3	1	+
Malaria	X	3,348	X	X	-680	X	-126,361	125,557	2	2	-
Temperature in first month of rainy season	X	X	X	-11,731	-165,367	X	157,484	-67,718	1	3	-
Variance of temperature in first month of rainy season	X	X	55,460	X	-41,710	X	237,193	77,771	3	1	+
Number of wet days in first month of rainy season	X	X	X	2,543	312,996	X	296,763	-27,713	3	1	+
Number of wet days in rainy season	X	-33,132	X	X	-214,753	X	-178,393	936	1	3	-
Variance of Number of wet days in first month of rainy season	X	X	-435,834	X	26,864	X	-219,695	85,200	2	2	?
Precipitation in first month of rainy season	X	X	67,155	X	-91,421	X	2,544	X	2	1	?
Precipitation in the rainy season	X	X	X	-3,300	-18,999	X	87,416	61,607	2	2	?

Note: X= excluded variable; 0= no effect, + = positive effect, - = negative effect, ? = inconclusive

7 Conclusions and recommendations

7.1 Conclusions

The first conclusion of our analysis is that heterogeneity with respect to natural production conditions (natural capital and climate) matters in terms of predicting income and vulnerability to poverty in rural Southeast Asia.

The different approaches used to estimate and interpret the main predictors of these differences offer similar rankings regarding the importance of the variables included in the analysis. Taken together, they suggest that greater ruggedness is a major determinant of poverty (confirming the perception on ongoing differences between uplands and plains) and that, conditional on this difference, production conditions in the first month of the wet season is of central importance in determining income.

Although some of the climatic variables are year-to-year levels, and as such they are akin to weather shocks (eg, number of wet days in the first month of the wet season), others reflect underlying climatic risk (eg, variance of number of wet days in the first month of the wet season, estimated over 10 years period). Hence, both shocks and risk matter to explain poverty in our cross-sectional data, the latter presumably because it shapes investment decisions, the former because it reflects limited capacity to smooth income/consumption. Linking these results with spatially explicit models of changes in climate may provide some guidance regarding future demand for both safety and cargo nets.

The discussion of vulnerability to poverty has the advantage of focusing the attention on more than snapshots of welfare, as measured by expected income, by forcing us to discuss other moments of its distribution. In our data, this characterization suggests two conclusions.

Firstly, that increasing expected income (ie, growth, as traditionally defined, or in graphical terms, moving to a different cohort, with higher) is still, at least conceptually, an important insurance strategy: households in cohorts with higher expected income have much lower probability of being poor. Secondly, that no environmental condition seems to simultaneously predict higher expected income and lower variance of income, suggesting that bundles of solutions to potentially important trade-offs may be required to reduce vulnerability to poverty in rural Southeast Asia.

7.2 Recommendations

The central importance of the first month of the wet season, which we believe reflects the ongoing central importance of rice production in the rural economies studied in this report, raises one obvious question: how to cope with negative changes in temperature and rainfall during that critical period?

Addressing this question, which may require either new technologies (eg, drought resistant varieties, or varieties with a different production cycle), new institutions (eg, better functioning labour or machine rental markets, perhaps using digital technologies), or a combination of both, seems central in insuring minimal disruptions to the production of what remains the staple food in this part of the world.

How to cope with risk is a longstanding question in agricultural economics, given the reliance of agricultural production on weather, and in development economics, given the perceived importance of risk as a determinant of poverty (which is supported in our analysis). Given the perceived increase in variability in climate, understanding the scope for insurance markets to function seems increasingly more relevant, even if changes in climate make the definition of such products harder than under stationary conditions.

Ongoing work on index insurance, typically directed at one crop/activity at the time, seems to be successful when that crop/activity largely dominates the livelihood portfolio (eg, livestock among East African pastoralists, insured against weather shocks through the Index Based Livestock Insurance, IBLI). Nevertheless, uptake of such insurance products remains, in most cases, disappointingly low, undermining their capacity to effectively reduce poverty. One possible direction for future research is whether the limited scope of the insurance product makes it less attractive in more diversified rural economies – in which case, “livelihood insurance” may be a much more attractive proposition.

The relationship between poverty reduction and biodiversity conservation seems particularly difficult to address. Although our results are correlations (and need to be interpreted as such) a plausible interpretation is that households in areas that are still relatively rich in biodiversity were bypassed by past agricultural R&D (that, plausibly, has looked “under the light” and focused on increasing yields in areas of greater return to such investments, eg, alluvial plains). If correct, existing biodiversity then reflects the lack of similar technologies.

Given that, in a time of re-wilding and of protecting “half the Earth”, societal constraints make it implausible that agricultural growth will follow the same technological path, how to increase income while minimising negative impacts on remaining nature remains a question that is of especial importance for much of the poorest areas of rural Southeast Asia. Again, how to develop different technologies, that aim a explicitly addressing agriculture multifunctionality (admittedly an imprecise concept), create new markets designed to reward the provision of environmental services or a combination of both, seem fruitful directions of research, particularly where poverty-environment trade-offs seem most relevant.

The reliance on subsidies typically looms large in discussing the interest of these solutions, all of which we could consider as cargo nets. In our view, that is a misguided approach.

In the absence of such technological or institutional changes, and in a world where migration remains limited, the forecasted increased in climate variability and the frequency of shocks and of degradation in natural capital, is expected to drive a larger share of rural population into unacceptable levels of welfare. Safety nets, themselves the clearest form of a subsidy, would then be more needed than ever, either as an ongoing poverty alleviation strategy or as emergency payments intended to minimise the consequences of shocks. Hence, we suggest that a better way to think about the interest of cargo vs safety nets is simpler, and perhaps less ideological: which one is more effective in terms of achieving society’s objectives, given what we expect about their short and long run impacts?

It is important to recognize however that, in this debate, the two approaches are now at very different starting points in terms of the strength of evidence supporting their use. While the credibility of cash transfers benefit from the continued and rigorous evaluation of their impacts, since the inception of Mexico’s Progresa, in 1997, very few examples of cargo nets can claim similar support, which is almost non-existent in some cases (eg, nature conservation parks).

Finally, and methodologically, the conclusions presented and discussed here are obviously more credible the lower “we can go” in terms of linking socio-economic data on household consumption with the environmental data. Because in two countries (Indonesia and The Philippines) we are somewhat limited by the size of the spatial units at which we can “locate” the households, our conclusions about the identification of distinct income cohorts in those countries has to be interpreted with more care than in other contexts (eg, mainland Southeast Asia).

Future work may explore whether it is possible to go lower in terms of locating households in space. If feasible, it may also be interesting to explore the higher frequency of HIES data in those two countries (both collect yearly data on income/consumption) to study

poverty dynamics through the use of approaches such as pseudo-panels, rather than the more limited approach of relying on cross-sectional data used in our analysis.

8 References

8.1 References cited in report

Amemiya, T. (1977) The maximum likelihood estimator and the non-linear three stage least squares estimator in the general nonlinear simultaneous equation model, Econometrica, 45, 955-968.

Antle, J. M. (1983). Testing the stochastic structure of production: a flexible moment-based approach. Journal of Business & Economic Statistics, 1(3), 192-201.

Azeem, M. M., Mugera, A. W., & Schilizzi, S. (2018). Vulnerability to multi-dimensional poverty: An empirical comparison of alternative measurement approaches. The Journal of Development Studies, 54(9), 1612–1636.

Bai, Z. G., Dent, D. L., Olsson, L., & Schaepman, M. E. (2008). Proxy global assessment of land degradation. Soil use and management, 24(3), 223-234.

Banks, James, Richard Blundell, and Agar Brugiavini. (2001). Risk Pooling, Precautionary Saving and Consumption Growthǁ. Review of Economic Studies, 68(4): 757-779.

Barrett, C.B. (2005), Rural poverty dynamics: development policy implications. *Agricultural Economics*, 32: 45-60.

Barrett, C.B. (2021), Overcoming Global Food Security Challenges through Science and Solidarity. *American Journal of Agricultural Economics*, 103: 422-447.

Barrett CB, Benton TG, Fanzo J, … and Wood S. 2020. *Socio-technical Innovation Bundles for Agri-food Systems Transformation*, Report of the International Expert Panel on Innovations to Build Sustainable, Equitable, Inclusive Food Value Chains. Ithaca, NY, and London: Cornell Atkinson Center for Sustainability and Springer Nature.

Battle, K. E., Lucas, T. C., Nguyen, M., Howes, R. E., Nandi, A. K., Twohig, K. A., ... & Gething, P. W. (2019). Mapping the global endemicity and clinical burden of Plasmodium vivax, 2000–17: a spatial and temporal modelling study. The Lancet, 394(10195), 332-343.

Bigio, S, Zhang, M and Zilberman, E (2020) Transfers vs Credit Policy: Macroeconomic Policy Trade-offs during Covid-19, National Bureau of Economic Research Working Paper Series, No. 27118.

Boucher, S, Carter, M, Flatnes, J E, Lybbert, T, Malacarne, J, Marenya,P, Paul, L A (2021), Bundling Genetic and Financial Technologies for More Resilient and Productive Small-scale Agriculture, National Bureau of Economic Research Working Paper Series, No. 29234

Brown, C S, Ravallion, M and van de Walle, D (2020) Can the World's Poor Protect Themselves from the New Coronavirus? National Bureau of Economic Research Working Paper Series, No. 27200

Burke, L., Reytar, K., Spalding, M., & Perry, A. (2011). Reefs at risk revisited. World Resources Institute.

Burke, M., & Lobell, D. B. (2017). Satellite-based assessment of yield variation and its determinants in smallholder African systems. *Proceedings of the National Academy of Sciences*, 114(9), 2189-2194.

Caro, T., Rowe, Z., Berger, J., Wholey, P., & Dobson, A. (2022). An inconvenient misconception: Climate change is not the principal driver of biodiversity loss. *Conservation Letters*, *15*(3), e12868

Carter, M, de Janvry, A, Sadoulet, E and Sarris, A (2017), Index Insurance for Developing Country Agriculture: A Reassessment, *Annual Review of Resource Economics*, 9, 421-438.

Cahyadi, E. R., & Waibel, H. (2016). Contract farming and vulnerability to poverty among oil palm smallholders in Indonesia. The Journal of Development Studies, 52(5), 681–695.

Canavire-Bacarreza, G., & Robles, M. (2017). Non-parametric analysis of poverty duration using repeated cross section: an application for Peru. Applied Economics, 49(22), 2141-2152.

Chantarat, S.C., Mude, A.G., Barrett, C.B., & Carter M.R., (2013). Designing index based livestock insurance for managing asset risk in northern Kenya. The Journal of Risk and Insurance, 80(1), pp. 205-237.

Chaudhuri, S., J. Jalan, and A. Suryahadi (2002) Assessing household vulnerability to poverty from cross-sectional data: a methodology and estimates from Indonesia. Unpublished, Department of Economics, Discussion Paper Series 0102-52, Columbia University.

Collier, P and Dercon, S (2014), African Agriculture in 50Years: Smallholders in a Rapidly Changing World? *World Development*, 63, 92-101.

Dang, H. A., & Dabalen, A. L. (2019). Is poverty in Africa mostly chronic or transient? Evidence from synthetic panel data. The Journal of Development Studies, 55(7), 1527-1547.

Dang, H. A., & Lanjouw, P. (2013). Measuring poverty dynamics with synthetic panels based on cross-sections.

Dang, H. A., Lanjouw, P., Luoto, J., & McKenzie, D. (2014). Using repeated cross-sections to explore movements into and out of poverty. Journal of Development Economics, 107, 112-128.

Deaton, Angus (1985) Panel Data from Time Series of Cross-Sections‖, Journal of Econometrics 30: 109-216.

Deaton, A. (2018). The Analysis of Household Surveys: A Microeconometric Approach to Development Policy. Reissue Edition with a New Preface. Washington, DC: World Bank.

Deaton, Angus and Christina Paxson. (1994). Intertemporal Choice and Inequality‖. Journal of Political Economy, 102(3): 437- 467.

Deaton, A., & Zaidi, S. (2002). Guidelines for constructing consumption aggregates for welfare analysis (Vol. 135). World Bank Publications.

Devereux, P. J. (2007). Improved errors-in-variables estimators for grouped data. Journal of Business & Economic Statistics, 25(3), 278-287.

Fischer, G., F. Nachtergaele, S. Prieler, H.T. van Velthuizen, L. Verelst, D. Wiberg, 2008. Global Agro-ecological Zones Assessment for Agriculture (GAEZ 2008). IIASA, Laxenburg, Austria and FAO, Rome, Italy.

Giri, C., E. Ochieng, L. L. Tieszen, Z. Zhu, A. Singh, T. Loveland, J. Masek, and N. Duke. 2010. Status and Distribution of Mangrove Forests of the World Using Earth Observation Satellite Data. Global Ecology and Biogeography: A Journal of Macroecology 20(1): 154-159. https://doi.org/10.1111/j.1466-8238.2010.00584.x.

Grosh, M., & Glewwe, P. (2000). Designing household survey questionnaires for developing countries. Washington, DC: World Bank.

Hanlon, J, Barrientos, A and Hulme, D (2012) *Just give money to the poor*, Kumarian Press

Hayami, Yujiro (2002), Family farms and plantations in tropical development, *Asian Development Review*, 19 (2), 67-89.

Hegwood, M., Langendorf, R.E. & Burgess, M.G. (2022) Why win–wins are rare in complex environmental management. *Nature Sustainability,* 5, 674–680.

Hengl, T., Mendes de Jesus, J., Heuvelink, G. B., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., ... & Kempen, B. (2017). SoilGrids250m: Global gridded soil information based on machine learning. PLoS one, 12(2), e0169748.

Imai, K. S., Gaiha, R., & Thapa, G. (2015). Does non-farm sector employment reduce rural poverty and vulnerability? Evidence from Vietnam and India. Journal of Asian Economics, 36, 47–61.

Just, R. E., & Pope, R. D. (1978). Stochastic specification of production functions and economic implications. Journal of econometrics, 7(1), 67-86.

Just, R. E., & Pope, R. D. (1979). Production function estimation and related risk considerations. American Journal of Agricultural Economics, 61(2), 276-284.

Kader, S and Santos, P (2022), Income and wildlife hunting in the Anthropocene: Evidence from Cambodia, Monash University Economics Working Paper

Mehrabi, Z., Ellis, E.C. & Ramankutty, N. (2018) The challenge of feeding the world while conserving half the planet. *Nature Sustainability,* **1**, 409–412

Meijer, J.R., Huijbegts, M.A.J., Schotten, C.G.J. and Schipper, A.M. (2018): Global patterns of current and future road infrastructure. *Environmental Research Letters*, 13-064006. Data is available at www.globio.info

Nolan E, Santos P (2019) Genetic modification and yield risk: A stochastic dominance analysis of corn in the USA. *PLoS ONE* 14(10): e0222156.

Novignon, J., Mussa, R., & Chiwaula, L. S. (2012). Health and vulnerability to poverty in Ghana: Evidence from the Ghana living standards survey round 5. Health Economics Review, 2(1), 11.

Pencavel, John. (2007). A Life Cycle Perspective on Changes in Earnings Inequality among Married Men and Women‖. Review of Economics and Statistics, 88(2): 232-242.

Pendrill, F, Gardner, T, Meyfroidt, P … West, C (2022), Disentangling the numbers behind agriculture-driven tropical deforestation, *Science*, 377 (6611), eabm9267.

Pimhidzai, O., Fenton, N. C., Souksavath, P., & Sisoulath, V. (2014). *Poverty profile in Lao PDR: poverty report for the Lao consumption and expenditure survey 2012–2013* (No. 100120, pp. 1-74). The World Bank.

Pritchett, L., A. Suryahadi, and S. Sumarto (2000) Quantifying vulnerability to poverty: a proposed measure, applied to Indonesia." Unpublished, Policy Research Working Paper No. 2437, The World Bank, Washington, DC.

Rohr JR, Barrett CB, Civitello DJ, … and Tilman D (2019) Emerging human infectious diseases and the links to global food production. *Nature Sustainability*. 2(6):445-456.

Sharaunga, S., Mudhara, M., & Bogale, A. (2016). Efects of 'women empowerment' on household food security in rural KwaZulu-Natal province. *Development Policy Review*, 34(2), 223–252.

Steffen, W, Richardson, K, Rockström, J … and Sörlin, S (2015), Planetary boundaries: Guiding human development on a changing planet, *Science*, 347 (6223), 1259855

Verbeek, M., & Nijman, T. (1992). Can cohort data be treated as genuine panel data? *Empirical Economics*, 17(1), 9–23.

Weiss, D. J., Lucas, T. C., Nguyen, M., Nandi, A. K., Bisanzio, D., Battle, K. E., ... & Gething, P. W. (2019). Mapping the global prevalence, incidence, and mortality of

Plasmodium falciparum, 2000–17: a spatial and temporal modelling study. *The Lancet*, *394*(10195), 322-331.

Wilson, E (2016) *Half-Earth: our planet's fight for life*, New York: W W Norton

Willett W, Rockström J, Loken B ... and Murray CJL (2019). Food in the Anthropocene: the EAT-Lancet Commission on healthy diets from sustainable food systems. *Lancet.* 393(10170):447-492.

Wood, G (2003) Staying Secure, Staying Poor: The "Faustian Bargain", *World Development*, 31 (3), 455-471

World Bank (2001) *World Development Report 2000/2001: Attacking Poverty*. New York: Oxford University Press.

World Bank (2007) *World Development Report 2008: Agriculture for Development*. Washington, DC. World Bank.

Zhu, L, Hughes, A C, … Watson, J E M (2021), Regional scalable priorities for national biodiversity and carbon conservation planning in Asia, Science Advances, 7 (35), eabe4261.

9 Appendixes

9.1 Appendix 1: Survey information

Table 16 Summary of household income and expenditure survey, Cambodia

Component	Description
Region	Southeast Asia
Survey name	Cambodia Socio-Economic Survey (CSES)
Rounds	2009, 2014, 2019
Sample size	CSES 2009: 12,000 CSES 2014: 12,096 CSES 2019: 10,075
Data structure	Repeated cross-section
Strata	Province (24), urban and rural
Sampling	Stage 1: Proportional to size (PPS) sampling (by number of households) of villages from each stratum Stage 2: Random sampling of one EA per village (large villages more than one EA) Stage 3: Random sampling of households per village (CSES 2009: 10 and 20 households in urban and rural villages, respectively, CSES 2014/2019: 12 hh per village)
Modules	Demographic characteristics, Housing, Agriculture, Education, Labour Force, Health and Nutrition, Victimization, Household Income and Consumption
Expenditure aggregate	Food (recall): consumed at home or outside the home (purchased, produced, received as gifts, or otherwise), Non-food (mostly recall): housing services (firewood, electricity, gas, water, and so forth), transportation and communication, purchase values of selected durable goods, personal use goods, recreation and entertainment, education and health Housing: rent (or imputed rent)
Normalization	Per capita

Table 17 Summary of household income and expenditure survey, Indonesia

Component	Description
Region	Southeast Asia
Survey name	National Socioeconomic Survey (SUSENAS)
Rounds	2010 – 2019 (yearly)
Sample size	~300,000
Data structure	Repeated cross-section (also a panel segment)
Strata	district
Sampling	Stage 1: PPS sampling of census blocks Stage 2: Random sampling of 16 households from each census block
Modules	Modules are collected in 3 year turns: First year, household income and expenditure Second year, household welfare socio-culture, trips and criminality module Third year, health, nutrition, education and housing
Expenditure aggregate	Food, non-food, rent
Normalization	
Note	in 2015 the reference period for certain items (health) was extended

Table 18 Summary of household income and expenditure survey, Lao PDR

Component	Description
Region	Southeast Asia
Survey name	Lao Expenditure and Consumption Survey (LECS)
Rounds	LECS 4 (2007/08), LECS 5 (2012/13)
Sample size	LECS 4: 8,226 LECS 5: 4938 (only 60% of the data publicly available)
Data structure	Panel (~ 4000 households)
Strata	Province and village type (urban, rural with road and rural without road)
Sampling	Stage 1: 518 (LECS 5: 515) PPS sampling of villages within each strata Stage 2: 16 Randomly sampling of 16 households (8 from earlier round and other 8 randomly selected from village roster)
Modules	Household characteristics, Consumption, Assets, Agriculture, Shocks, village characteristics
Expenditure aggregate	Food (30 days diary): purchased, own consumption, gifts and meals in restaurants and hotels Non-food (30 days diary): education, medical expenses, clothing, fuel and utilities, transportation and communication, personal care, recreation, accommodation, alcohol and tobacco, traditional and cultural expenses, household sundries and operating expenses and other miscellaneous items Housing: no information on rent
Normalization	Per capita (household members)

Table 19 Summary of household income and expenditure survey, Myanmar

Component	Description
Region	Southeast Asia
Survey name	Myanmar Living Condition Survey (MLCS)
Rounds	MLCS 2017
Sample size	13,730
Data structure	Cross-section
Strata	State/Region, urban and rural
Sampling	Stage 1: PPS sampling of Enumeration Areas (EA) Stage 2: 12 households were randomly selected in each EA The sample covers all districts and 296 townships (total 330 townships)
Modules	Household Roster, Education, Health, Housing, Food Consumption, Non-food purchases, Household durables, Labour & employment, Agriculture, Non-farm business, Finance, Shocks & coping strategies, Remittances and Other income
Expenditure aggregate	Food (weekly): food, consumption of home-produced food and food received in kind (self-reported or imputed market price), Non-food (past 30 days, 6 months or 12 months: tobacco and alcohol, education, clothes and footwear, energy, water and sanitation, personal care, transport and communication (excluding purchase of vehicles), recreation, leisure and cultural expenses, entertainment materials and consumables) Housing: rent and imputed rent for owners durables: usage value of durable goods (e.g. cars)
Normalization	Scales to calculate adult equivalents = 0.55 (<1 year); 0.67 (1-3 years); 0.79 (4-6 years); 0.83 (7-9 years); 0.97 (10 – 12 years); 1.04 (13 – 15 years); 1,1 (16 – 19 years); 1 (20+ years)
Note	For Kayin State and Rakhine State, total food consumption was imputed due to data quality issues

Table 20 Summary of household income and expenditure survey, Timor Leste

Component	Description
Region	Southeast Asia
Survey name	Timor-Leste Survey of Living Standards (TLSLS)
Rounds	TLSLS 2 (2006/07) , TLSLS (2014/15)
Sample size	TLSLS 2: 4,477 TLSLS 3: 5,916
Data structure	Repeated cross-section
Strata	TLSLS 2: Urban and rural strata of 5 regions TLSLS 3: Urban and rural strata of 13 districts
Sampling	TLSLS 2: Stage 1: PPS sampling of 60 Enumeration Areas (EAs) from each region (total 300EAs) Stage 2: Randomly selection of 15 households (clustered at EAs) TLSLS 3: Stage 1: For the 2010 Census, the total population was disaggregated in 1809 EAs; Sampling followed the 2012 Labor Force Survey (LFS) with a sample size of 472 EAs (pps sampling); 400 EAs were randomly selected for TLSLS 3 (with same probabilities at strata level) Stage 2: Random of 15 households (clustered at EAs)
Modules	Consumption expenditures, health and education status of households, anthropometric measurements of children, assets, agriculture, and occupational and employment status of household members
Note	Due to violent conflicts during the data collection of TLSLS2, a second survey (detailed questions about the conflict) was collected in a subsample of 1789 households. The name of this survey was TLSLS2X.
Expenditure aggregate	Food, non-food, rent
Normalization	Per capita

Table 21 Summary of household income and expenditure survey, Philippines

Component	Description
Region	Southeast Asia
Survey name	Family Income and Expenditure Survey (FIES)
Rounds	2006, 2009, 2012, 2015, 2018
Sample size	FIES 2006: 38,483 FIES 2009: 38,400 FIES 2012: 40,171 FIES 2015: 41,544 FIES 2018: 170,917
Data structure	Repeated cross-section
Strata	Major domains(Region(33)/province(81)/other areas(3) (and highly urbanized cities (HUC)))
Sampling	FIES 2018 (similar for FIES 2006 – 2015): Stage 1: 87,098 Primary Sampling Units (PSUs) are formed from 42,036 barangays. PSU size ranges from 100 to 400 households. PSUs were ordered according to the following criteria: (1) Geographic location (NS/WE); (2) Proportion of HHs with Overseas Worker; and (3) Wealth Index. Counting and selecting PSUs Stage 2: Random selection of households. Selected number of households varies with respect to PSU size (Mean: Urban: 12 hh; Province: 16 hh)
Modules	Identification and Other Information; Expenditures and Other Disbursements; Housing Characteristics; Income and Other Receipts; Entrepreneurial Activities; Social Protection; Evaluation of the Household Respondent by the Interviewer.
Expenditure aggregate	food, non-food, gifts, support, assistance (by the family to friends), rent (and imputed rent of owner-occupied dwelling unit), own-produced goods consumed by the family
Normalization	Per capita

Table 22 Summary of household income and expenditure survey, Vietnam

Component	Description
Region	Southeast Asia
Survey name	Vietnam Household Living Standards Survey (VHLSS)
Rounds	2014, 2016, 2018
Sample size	VHLSS2014: 46,995 (expenditure data collected on a subsample of 9,399 households)
Data structure	Rolling panel (50% of the households are revisited)
Strata	Regions (8), provinces (63), rural and urban
Sampling	Stage 1: PPS sampling of communes (stratified for province and urban/rural) Stage 2: PPS sampling of 3 EAs for each commune Stage 3: Selection of households
Modules	Household survey: household roster, education, employment, health, income and household production, expenditure, durable goods and assets, housing, participation in poverty reduction programs,
Consumption module	Demographics, education, health and health care, labour – employment, income, consumption expenditure, durable goods, housing, electricity, water, sanitation facilities, participation in poverty alleviation programmes, household businesses, commune general characteristics
Expenditure aggregate	Food, non-food, rent
Normalization	

9.2 Appendix 2: Machine Learning

Machine learning is algorithmic approach to predicting outcomes (e.g. consumption) based on a number of variables (e.g. stocks of produced, natural, and human capital). There are techniques which are more complex than regression techniques which may improve the accuracy of predictions when linearity does not hold. Although predictions of the outcome variable are not the focus of this analysis, we are interested in understanding how the different variables contribute to the predictions made by these models.

Machine learning techniques vary in their degree of interpretability. There are several definitions of interpretability available in the literature. Here we define an interpretable technique as any that results in a model which operates in a manner such that humans can understand the reasoning behind the predictions and decisions made by the model. For example, the regression models defined above in section 3.1 can be considered interpretable as it is possible to predict the value of the dependent variable for any set of independent variable values, as the model outputs include the coefficients for each variable and the structure of the relationship is known

There are techniques for making the black box models interpretable. That is, it may still be feasible to obtain some understanding of the model and the relationships between input and output variables. This may be via supplementation with other models known as

surrogate models, visualisation of coefficient relationships, development of variable importance scores, and/or understanding of some subset of rules and relationships inherent in the model.

Five different machine learning techniques were applied here, with varying levels of interpretability, including regression trees, random forests, gradient boosted trees, linear tree models, and cubist models. Details on each of the models we utilised, and their respective level of interpretability is included below:

Regression trees: Regressions trees are a type of decision tree model that split data according to cut-off values for different variables and generate predictions based on these splits and the different subsets of data they create. These splits occur where the sum of squared errors across variables is minimised. The final subsets each observation ends up in are the terminal or leaf nodes. Regression trees are useful for determining important splits in variables and overall importance of features in a tree.

Linear tree models: These are an extension of regression trees with linear models at of the leaves. This enables prediction of output for each observation rather than the average outcome calculated in regression trees.

Random forests: Random forests are ensemble methods (methods that combine a number of models) consisting of a large number of decision trees, in this case regression trees. The concept of bootstrap aggregation (bagging) is employed by selecting random samples (bags) from the data for each tree. Given regression trees are based on different samples of data, each may give a different prediction. The prediction random forests reach is the average of the predictions from the regression trees inherent within the forest. By employing bootstrap aggregation and then taking an average, model performance is improved as variance of the model is decreased without increasing bias. This is particularly pertinent given the sensitivity of regression trees to training data. Importance scores for variables can then be computed by averaging the difference in out-of-bag (those observations not included in a tree) error before and after the permutation over all trees. The before out-of-bag error is recorded for each data point and averaged over the forest. To calculate the after out-of-bag error, the values of the feature are removed among the training data and the out-of-bag error is again computed on this perturbed data set. Features which produce large values for the difference are ranked as more important than features which produce small values (Breiman, 2001). This importance is a measure of by how much removing a variable decreases accuracy, and vice versa.

Gradient boosted trees: Gradient boosted trees are another ensemble method. Similar to other boosting methods they are built step-wise by combining multiple models to reduce variance without adding additional bias. Gradient boosted trees are initialised with a weak learner (prediction is the average outcome) and then supplemented with additional trees until the predictive ability is optimised (at this point adding an additional tree does not reduce the error). Similar to random forests data is placed in random subsets as each tree is produced, where gradient boosted trees differ is that where data is poorly modelled it is prioritised in new trees. This approach of continuously taking account of the fit of previous trees that are built to improve accuracy is achieved by weighting throughout the boosting processes and improves the likelihood of all relevant variables being included. For each tree, the gain on each node can then be calculated for each variable and the contribution summed across trees to gain a measure of variable importance.

Cubist models: Cubist models are rule-based models that are used to create trees with a linear regression model in the leaves that is based on a set of rules developed to subset the data. They contain intermediate linear models at each step of the tree. Cubist models partition data into subsets with characteristics similar to the target variable and covariates, and then establish a series of rules to define the partitions. Each of these rules can be based on one or more covariates. This results in a set of regression equations that are general in form but local to the subsets of data partitioned, which decreases the overall error. In the cubist models developed here we also employ a scheme similar to boosting

called committees, where iterative model trees are created in sequence and all trees produced after the first use adjusted versions of the training set outcome. Unlike boosted trees, weights are not used to average the prediction from each model tree, the final prediction is a simple average from each tree. In addition, a nearest neighbour algorithm is applied to the leaf nodes and an ensemble approach combining the cubist prediction and nearest neighbour prediction used. Given the rules used can be directly observed the interpretability of cubist trees is higher relative to random forests and gradient boosted trees. However, where supplementary committee and nearest neighbour approaches are employed, this interpretability does decrease.