Australian Government
Australian Centre for International Agricultural Research

Vulnerability in the Anthropocene: a prospective analysis of the need for social protection

102
ACIAR TECHNICAL REPORTS SERIES

Vulnerability in the Anthropocene

A prospective analysis of the need for social protection

Paulo Santos
Reiss McLeod
Stefan Meyer
Hai-Chau Le
Mukhammad Fajar Rakhmadi
Editors

2023

The Australian Centre for International Agricultural Research (ACIAR) was established in June 1982 by an Act of the Australian Parliament. ACIAR operates as part of Australia's international development assistance program, with a mission to achieve more productive and sustainable agricultural systems, for the benefit of developing countries and Australia. It commissions collaborative research between Australian and developing-country researchers in areas where Australia has special research competence. It also administers Australia's contribution to the International Agricultural Research Centres.

The Chief Executive Officer of ACIAR reports directly to the Australian Government Minister for Foreign Affairs. ACIAR operates solely on budget appropriation from Australia's Official Development Assistance.

The use of trade names constitutes neither endorsement of nor discrimination against any product by ACIAR.

ACIAR TECHNICAL REPORTS SERIES

This series of publications contains technical information resulting from ACIAR-supported programs, projects and workshops (for which proceedings are not published); reports on ACIAR-supported fact-finding studies; or reports on other topics resulting from ACIAR activities. Publications in the series are available as hard copy, in limited numbers, and published on the ACIAR website at aciar.gov.au

Santos P, McLeod R, Meyer S, Le HC, and Rakhmadi MF (2023) *Vulnerability in the Anthropocene: a prospective analysis of the need for social protection* ACIAR Technical Report No. 102, Australian Centre for International Agricultural Research, Canberra.

ACIAR Technical Report Series No. 102 (TR102)

ISBN 978-1-922983-27-5 (print)

ISBN 978-1-922983-28-2 (PDF)

Editing and proofreading by Neat Copy

Design by Elliott Street Typesetting & Design

Printing by Instant Colour Press

Cover: A livestock farmer and her child in northwest Vietnam

Photo: Vu Khanh Long

Foreword

The COVID-19 pandemic was a global health and economic crisis that disrupted the lives and livelihoods of diverse communities worldwide and will continue to have impacts for many years to come. Developmental challenges in the Asia-Pacific region were magnified as a result of the Covid-19 pandemic. Hard-won advancements in recent development efforts were eroded, and the burdens of the pandemic were borne disproportionately, albeit with a high degree of variability, by people experiencing poverty across the region. This raised questions about why particular groups were more vulnerable to particular types of shock than others.

The Australian Centre for International Agricultural Research (ACIAR) is mandated under the ACIAR Act (1982) to work with partners across the Indo-Pacific region to generate the knowledge and technologies that underpin improvements in agricultural productivity, sustainability and food systems resilience. We do this by funding, brokering and managing research partnerships for the benefit of partner countries and Australia.

ACIAR supported partners from Monash University to analyse secondary data on household-level poverty in 7 Southeast Asian countries: Cambodia, Indonesia, Lao PDR, Myanmar, The Philippines, Timor-Leste, and Vietnam. Utilising machine learning techniques, the analysis identified key factors shedding light on potential underlying drivers of acute vulnerability to shocks.

This report presents the results of the analysis, illustrating the relationship between human well-being, household assets (including natural capital like biodiversity and climate), physical capital, and exposure to health shocks such as Malaria and Dengue. It concludes a three-stage assessment process that began with ACIAR Technical Report 95, titled 'Food systems security, resilience, and emerging risks in the Indo-Pacific in the context of COVID-19: a rapid assessment.'

The technical report underscores the significance of natural capital and prevailing climate as predictors of vulnerability to shocks and subsequent poverty in rural Southeast Asia. It also offers recommendations to ACIAR for further research and suggests avenues for governments and policymakers to enhance household resilience to future shocks.

Wendy Umberger
Chief Executive Officer
ACIAR

Authors

Paulo Santos
Economics, Monash University

Reiss McLeod
Economics, Monash University

Stefan Meyer
Monash University

Tran Minh Hang
Institute of Anthropology, Vietnam Academy of Social Sciences, Vietnam

Hai-Chau Le
Independent researcher

Mukhammad Fajar Rakhmadi
Independent researcher

Contents

List of tables

List of figures

Acknowledgements

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

Finally, we thank 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.

Acronyms and abbreviations

HIES	Household income and expenditure surveys
LASSO	Least Absolute Shrinkage and Selection Operator
OLS	Ordinary Least Squares regression
RMSE	Root mean square error
SSE	Sum of squares error
SST	Sum of squares total
UBI	Universal basic income

Executive summary

The COVID-19 pandemic exposed the fragility of some of the progress made towards reducing poverty. This 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:

- Cambodia
- Indonesia
- Laos
- Myanmar
- the Philippines
- Timor-Leste
- Vietnam.

The analysis combines:

- data on income/consumption from recent representative Household Income and Expenditure Surveys (HIES)
- rich 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 8 groups. Groups are characterised by living standards and environmental characteristics.

Several conclusions emerge from this analysis.

1. The importance of poverty in rural South-East Asia – 41% of the households in our data were allocated to groups with average income below the poverty line.
2. The high diversity of the natural environment of low-income households, aggregated in 5 distinct groups, in contrast with the remaining 59% households in 2 large clusters.
3. 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 these data, suggesting agricultural insurance (and, more generally, social safety nets) may play a role in reducing poverty.
4. These groups are also very different in terms of income variability. Together with differences in average income, this leads to high disparity in vulnerability to poverty.
5. Our analysis shows that biodiversity richness predicts both lower income variability and lower expected income. This suggests that biodiversity conservation may come at the cost of increased vulnerability to poverty.

1 Background

credit: Alex Tilley and Ruby Grantham

1 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, our emphasis)*

The COVID-19 pandemic exposed chronic development fault lines across the Asia-Pacific. For example, people living in poverty were more affected than others because they could not follow the World Health Organization's social isolation recommendations (Brown et al. 2020). It was possible that large parts of the population would go into poverty. Because of this, governments transferred income to households to smooth the economic outcomes of the shock created by the pandemic.

Government transfers are one of several policy options available to address the consequences of shocks. For example, Bigio et al. (2020) discuss, from a macroeconomic perspective, the relative merits of more transfers versus more access to credit with the aim of managing the business cycle. Closer to the analysis in this report, and more focused on how to reduce persistent poverty, Barrett (2005) distinguishes between:

- 'safety nets', which aim to protect the asset base of people experiencing poverty from shocks and prevent a slide into poverty that is hard to escape
- '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 on the impacts of cash transfers
- how easy it is to roll-out such transfers even in contexts of low penetration of financial markets
- in the case of the universal basic income (UBI), the little information needed for its implementation given its untargeted nature.

In terms of budgets, safety nets such as UBI capture a large part of public funds. These must be raised through taxes with distortionary effects that are not shown in this very simple illustration.

This may make technological change attractive, given the perceived high returns to investment in research and development, for example, even if 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, the distinction between cargo and safety nets is less clear in reality. The presence of multiple market failures makes uninsured risk (the likelihood of negative shocks against which agents cannot effectively protect themselves) a potential driver of the 'Faustian bargain'. This means 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 difference between 'giving a fish' and 'teaching to fish'.

As made clear in the recent reviews of the experience with index insurance (Carter et al. 2017) or cash transfers (Hanlon et al. 2012), the hope is that directly reducing the risk of experiencing poverty (a 'safety net') can lead to dynamic adaptations. This could be investment in more productive but riskier technologies (a 'cargo net') that help to reduce poverty. Similarly, the large gains in mean yield that go 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 an either/or discussion, the different types of policies can be combined, as in the analysis of Boucher et al. (2021).

The point to keep in mind in this report is that despite their largely positive record, safety nets are just one of many ways to address the importance of risk as a driver of poverty. However, while safety nets may 'only' require identifying people experiencing poverty (in itself, not easy), developing effective cargo nets requires first identifying who is at risk of experiencing poverty (or '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 with what has come to be known as the Anthropocene: 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 these are seen to be changing at rates that threaten human survival (Steffen et al. 2015).

Before we present the approach used in this study and our results, there are a few initial points. The first is what Caro et al. (2022) call the 'inconvenient misconception' that climate change is behind the accelerated reduction in biodiversity. At present, this 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. This drives much effort into design solutions that may address both crises (Zhu et al. 2021).

The second is that it is likely that society will demand changes in how the agricultural sector contributes to human welfare, reflecting the sector's contribution to ongoing environmental degradation. This is particularly evident through changes in biodiversity, mostly driven by:

- land use change induced by the expansion of agricultural production (Pendrill et al. 2022)
- harvesting of wildlife in rural areas, which could be a coping strategy in agricultural production shocks (see Kader and Santos 2022 and references therein).

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 et al. 2018) which emphasise 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 intersects with all of the United Nations' 17 goals for Sustainable Development other challenges are likely to add to the need for change in the way farming contributes to human welfare, for example with:

- diet and its effect on human health (Willet et al. 2019)
- reducing the risk of contact with new infectious diseases (Roth et al. 2019).

The implication is that the traditional emphasis on the contributions of agriculture to economic development and poverty reduction, which date back to the Green Revolution and is 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. It is also the main message of the quote we opened this section with.

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. This explains why:

- win-win solutions are elusive (Hegwood et al. 2022)
- 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, a comment about the role of smallholder farmers. As Hayami's (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 land productivity reflected the capacity to overcome labour market failures and favoured the emphasis on smallholders in many contexts. However, smallholder farmers do not have an advantage in overcoming credit and information market failures (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 deagrarianisation of food production) will only reinforce the importance of those limitations and it will call into question the viability of smallholders as the central actors in developing 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.

2 Objective

credit: Massimo Munnichi

2 Objective

This project contributes to the overall objective in the ACIAR 10 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 South-East Asia?

3 Methodologies

credit: Conor Ashleigh

3 Methodologies

3.1 Data

We compiled data from representative national household income and expenditure surveys (HIES) designed to measure and monitor poverty from 7 South-East Asian countries. The 7 countries were:

- Cambodia
- Indonesia
- Laos
- Myanmar
- the Philippines
- Timor-Leste
- Vietnam.

All surveys were representative at rural and urban level. The analysis focused on the rural strata only. The detailed information in these surveys is in section 3.1.1.

One important feature of these data was the availability of information about spatial location of households. This feature enabled economic data on consumption and wealth to link with other datasets, which contain information on a wide array of environmental variables that may capture the 2 dimensions of the Anthropocene we planned to focus on:

- Climate change
- Biodiversity loss

These environmental variables are described in section 3.1.2.

3.1.1 Income and expenditure data

We used the latest publicly available survey data for each country. Table 3.1 shows the year and sample size of each of the surveys used. Appendix 7.1 has information about each of the 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 (either panel or repeated cross-section) as well as strata and sampling, given the implications of survey design for the analysis (Deaton 2018).

Table 3.1 Survey round selected for the analysis by country

Country	Year	Sample size
Cambodia	2014	11,622
Indonesia	2019	181,981
Laos	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. This is typically seen as an accurate measure of wellbeing in developing countries with large informal employment (Deaton 2018).

National statistical offices collect data for the consumption aggregate through nationally representative HIES. Well-known examples are household surveys collected by the Living Standard Measurement Study (Grosh & Glewwe 2000) with a lasting methodological influence, including in some of the surveys used in this study (Myanmar, Timor-Leste, Vietnam).

National statistical offices collect and can give the aggregated household consumption variable and data for different consumption modules. These modules typically cover 4 categories (Deaton and Zaidi 2002): food (purchased, in-kind, home-produced and food away from home), non-food (education, health and other non-food), durables and housing.

Data for several items were collected for each category. The data collection method differs by survey and module. Some data were collected through diaries, where households recorded their consumption. Other data were collected via recall. The length of the period of data collection varies by category. For goods that were consumed, the period was usually shorter (one week to one month) than for durables (3 months, 6 months or 12 months). The aggregate for each category was then extrapolated to one year.

Aggregate consumption was calculated by summing the value of goods consumed in each category. For bought goods, the values were recorded in the household survey. Goods received as in-kind payment or that were self-produced (for which, typically, only quantities were collected) were valued using average local prices. The usage values (current value depreciated for the total time of usage) for durables were estimated. Some aggregates also contain housing spending (like rent, estimated through hedonic regressions and imputed in the case of households owning their residence). The final intake aggregate accounts for cost-of-living or spatial differences by deflating spending by a Paasche Price Index (Deaton & 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.

Table 3.2. summarises the data collected for each country. Data on durables was typically not available. Data limitations meant housing was not included in all of the consumption aggregates. For example, the income aggregate produced by the national statistical offices in Laos was calculated without housing rents. This was because the rental market information was very thin, as only 1.4% of the respondents were tenants (Pimhidzai et al. 2014).

Data differed across the 4 consumption modules. Because of this, the consumption indicator used in this study is based on value of consumption of food and non-food items only. This is so data can be compared across countries. As a result, our estimates (for example, of poverty) differ slightly from official national estimates.

We also make some adjustments to the expenditure variables. Given we have data from different years (see Table 3.1), so they can be compared, we convert the consumption aggregate per adult equivalent to a common year (2019) and a common currency (US$).

Table 3.2 Consumption data

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

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 (shared consumption within a household) (θ) which we assume to be 0.9, reflecting that economies of scale are usually low in developing countries given that food is the main consumption good in households.

The scope of the HIES data is quite wide. Most surveys collected data through multi-modular surveys that, in addition to consumption, included:

- household roster (with information on demographics such as gender, age and ethnicity of household head, household composition, education)
- assets
- labour
- employment (Grosh & Glewwe 2000).

We use data from these modules to build predictors of vulnerability to poverty, grouped into 2 categories (see Table 3.3):

- human capital
- physical capital

While the general scope and coverage of each survey is similar, there are some important differences. Some HIES datasets have a limited set of variables for poverty analysis (for example, the Philippines, Timor-Leste, Vietnam). Others include a broad set of modules covering multiple topics (for example, Myanmar).

Village surveys were also administered. However, data protection policies made it impossible to access the village survey data in the Philippines and Myanmar.

Table 3.3 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 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

3.1.2 Spatial data

The national statistical offices provided information about the different administrative levels for surveyed households (hereon, spatial identifiers). Table 3.4 shows that spatial identifier levels vary in each country, ranging from the village level to the provincial level. Polygonal data showed the boundaries of each of the spatial identifiers. We could then link different layers of spatial data with the survey data using the common spatial identifier.

We collected spatial data across 3 categories of variables:

- natural capital and climate
- physical capital
- shocks to health.

Table 3.5 to Table 3.8 give an overview of the different variables, grouped by category, and provide the resolution of the data. Most of the data are in raster format, with cells differing in size (for example, 100 by 100 m, or 250 by 250 m). We compiled several variables by extracting the mean value of the spatial data listed in Table 3.5 to Table 3.8 for each spatial identifier described in Table 3.4.

Table 3.4 Consumption data

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

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 (such as district), g_{ij} is the value of the observation i (at cell level) within the geography and N_j is the total number of cells within geography j.

Natural capital and climate

We used several recently constructed datasets with global coverage. These define our variables of interest in a common way across countries.

We substituted natural capital using variables that measure stocks of natural resources, both in terms of:

- quantity (for example, soil depth)
- its condition, where possible (for example, erosion).

We focussed on variables that contributed to biomass production (for example, available water capacity and climatic variables) that can be interpreted as inputs in an agricultural production function.

Data on natural capital came from a range of sources (see Table 3.5). Data on climate (Table 3.6) came from Climatic Research Unit gridded Time Series shown in Harris et al. (2020).

Variables shown in Table 3.5 are snapshots of stocks of natural conditions in specific locations. Contrary to these variables the climate data are, as expected, dynamic. The Climatic Research Unit gridded Time Series, described in Harris et al. (2020) shows a high-resolution, monthly grid of land-based (excluding Antarctica) observations going back to 1901.

Table 3.5 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 5 m in height. The raw data are the images collected by NASA's Landsat satellites	30 m	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 were 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 are projected using 4 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.	~1 km	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

Table 3.5 Natural capital variables *(continued)*

Variable	Description	Resolution	Year
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. 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.	30 arc seconds	1996
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 2 points to obtain the uphill slope. The same calculation is performed for each of the 8 major directions of the compass (north, northeast, east, southeast, south, southwest, west, and northwest), and the 8 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 come from a compilation of soil profile data (ca. 1,300,000 locations) and borehole data (ca. 1.6 million locations).	7.5 arc seconds	2010
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	--

Table 3.6 describes 8 observed and derived variables, including:

- average temperature
- average rainfall
- the number of wet days for each month.

We excluded several climatic variables from the analysis because they had a high correlation (between 0.88 and 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
- variance of rainfall in the wet season.

Given their almost perfect alignment with included variables, this decision improves prediction and greatly reduces estimation time.

Table 3.6 Climate variables

Variable	Description	Resolution
Temperature in first month of rainy season	Air temperature in degrees Celsius, at 2 m above the surface, in the first month of the rainy season in the year of the survey.	55 km
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 m above the surface) in the first month of the rainy season across 2010 to 2020.	55 km
Number of wet days 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.	55 km
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 to 2020.	55 km
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.	55 km
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.	55 km
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 to 2020.	55 km
Total rainfall in the rainy season	This variable is a measure of total rainfall in the wet season in the year of the survey.	55 km

Physical capital

Physical capital at community level can be an important predictor of income in rural areas. Its importance reflects agro-ecological conditions (natural capital and climate), as they shape agricultural profitability. However, we only include data on irrigation in our analysis as it is both globally available and expected to moderate the effect of climatic conditions, which are of primary interest to this analysis.

It may be interesting to include data on road access as it proxies for access to markets – a different way to manage production shocks. However, we did not have access to this data measured in a comparable way across countries.

In some countries it may be possible to complement these data with information collected by village surveys, conducted at the same time as household surveys. These data however are uncommon.

Health shocks

We include measures of disease prevalence, as indicative of the likelihood of health shocks that may reduce productivity.

Table 3.7 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 2 datasets sub-national irrigation statistics from national statistics as well as from international organisations (such as FAO and World Bank). To identify the geospatial locations of the irrigation schemes, irrigation maps of the reports were digitalised. As well, information from other sources (such as atlases and inventories) was used. The data are on the grid cell level. Each grid cell is the share of the total area equipped with irrigation.	5 arc minutes	2005

Table 3.8 Health 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 for how long it takes for an anopheles mosquito to develop parasites after an infected blood meal.	55 km	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

3.2 Methods

3.2.1 Vulnerability to poverty

Following the last World Development Report on Poverty (World Bank 2001):

- vulnerability to poverty is defined as the probability of experiencing poverty in the future
- poverty is defined as having an income below a certain threshold (such as a poverty line).

Measuring and understanding the nature of poverty has progressed since the 2001 World Development Report (for example its multidimensionality). However, progress in understanding and quantifying vulnerability has lagged. This may reflect the relatively demanding nature of this concept on the data – namely its prospective and probabilistic nature.

This report builds on empirical applications that addressed these 2 demands. We used past data on income or consumption to infer the *probability* of *future* deprivation, assuming the 'production function' for income generation is independent of the period when income and other variables are measured. In practice, this means we estimated both a household's expected (or mean) consumption and its variance.

The need to move beyond expected income should be clear. For example, a salaried public servant with an expected level of consumption similar to a farmer's may still be (and feel) much less vulnerable to poverty because of the relative stability of their income.

Characterising the income generation process in terms of mean and variance allowed us, in a second stage, to predict the probability that household consumption would be below the poverty line. This means empirical estimates of $vh_t = \Pr(ch_{t+1} \leq z)$ where:

- ch_{t+1} is the household's per-capita consumption level at time $t + 1$
- z is the appropriate poverty line.

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 US$1.90 per day per person in 2011 dollars and adjusted it to 2019 dollars (US$1.6735 per day) so it aligned with the consumption aggregate, which was also adjusted to 2019 values.

Estimating vulnerability to poverty required consistent estimates of the mean and variance of income. In a regression context, the first step was to characterise 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 could directly estimate the expected (log) consumption which, conditional on X_h, became 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:

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

for each household h.

Assuming that consumption is log-normally distributed (equivalently, that lnc_h is normally distributed), we could use these estimates to:

- characterise the distribution of income
- estimate the probability that a household with characteristics X_h, would be experiencing poverty – that is, estimate the household's vulnerability level.

Letting Φ(.) denote the cumulative density function of the standard normal distribution, this estimated probability is given by:

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

3.2.2 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 2 early examples of using a conceptually similar approach to estimate vulnerability to poverty. More recent applications include Novignon et al. (2012), Imai et al. (2015), Cahyadi and Waibel (2016), Sharaunga et al. (2016), and Azeem et al. (2018).

In this approach, we must first 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_{ij} 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
- S_j are vectors of communal shock variables (dengue and malaria for example)
- I_j is a vector of community physical capital variables (such as irrigation).

The effect of observable characteristics on consumption shows in our estimates of the parameters β and ε_i is an idiosyncratic error term that captures the unobserved determinants of consumption.

Then 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 obtain estimates of vulnerability to poverty (in conjunction with the estimates of their effect on mean income).

Finally, 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 3 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'. Including variables 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 example of these type of variables is measures of soil quality.

An alternative is to hypothesise 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. In other words, letting the data reveal the best description of the production technology. We explore this more in the next section.

The second question is about interpreting the estimates from equations (5) and (6), which reflects the set of explanatory variables in our estimates. Many earlier applications rely, like us, on cross-sectional data, with well-known limitations.

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 can be a good proxy of poverty dynamics estimates when they use a large cross-sectional data set covering a large variation in consumption and observable household characteristics.

The second question is about leaving relevant explanatory variables out of the estimation of equations (5) and (6). We have a rich set of both household and environmental variables that may explain individual wellbeing, but we cannot claim to include all variables that may explain income. In particular, several datasets do not include information on the full set of public services (such as roads or health services) that we would like to account for.

This limitation affects the interpretation of those variables for which we consistently have information across surveys and that we can include when estimating equations (5) and (6). Hence, we interpret these estimates as predictors of vulnerability, which affects the type of policy implications that can be derived from our analysis. We discuss this issue in more detail in section 5.

A third question relates to the high correlation between included explanatory variables. Such multicollinearity affects the size and precision of the OLS estimates (typically very large, sometimes with counter-intuitive directions of the effect). There are a few solutions to this problem. One is to use 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 (they are effectively dropped from the results). Although the resulting models are easier to interpret and typically exhibit better predictive behaviour, this approach increases the bias of the fitted model.

One final comment on inference. The household survey data are typically clustered at the village level. Because households in villages were similar, estimates of variance in a clustered sample underestimate the true variance, requiring the use of clustered standard errors.

3.2.3 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. These identify and sort the importance of different constraints and, in particular, the possibility of thresholds in this relation. Equation (5) can then be rewritten as

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

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

Where depends on whether a specific variable *W* is above or below a certain cut-off (w_o) 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). Selecting variables W and their 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, such as climate and biodiversity
- account for variables that may moderate that effect, such as ruggedness or 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. For example, research and development and associated extension services are directed to more productive agro-ecosystems while households react to the changes in production possibilities as a result of 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 useful to understand poverty, and consequently the utility of and need for social protection programs. This is the case even if 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 growing) 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.

Machine learning generally covers algorithmic approaches to predicting outcomes (such as consumption) based on variables (for example 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 usually 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 (briefly presented above)
- gradient boosted trees
- random forests
- neural networks.

Each model differs in terms of complexity and computational requirements. To select the best model, we calculated the R squared for each model on test data to compare performance. This is presented in the results section of this report.

Random forests

We used the random forest algorithm to construct different cohorts. It was the best predictor of consumption within a feasible running time.

A random forest algorithm generates multiple decision trees, where each tree is constructed by minimising the Gini index. Each tree considers only a random subset of the data and this leads to a different set of individually-biased parameters. An average voting scheme among individual trees determines the final model, as shown in Figure 3.1 (Kim and Kim, 2022). Our outcome variable is a continuous variable, so each decision tree in the random forest is a regression tree.

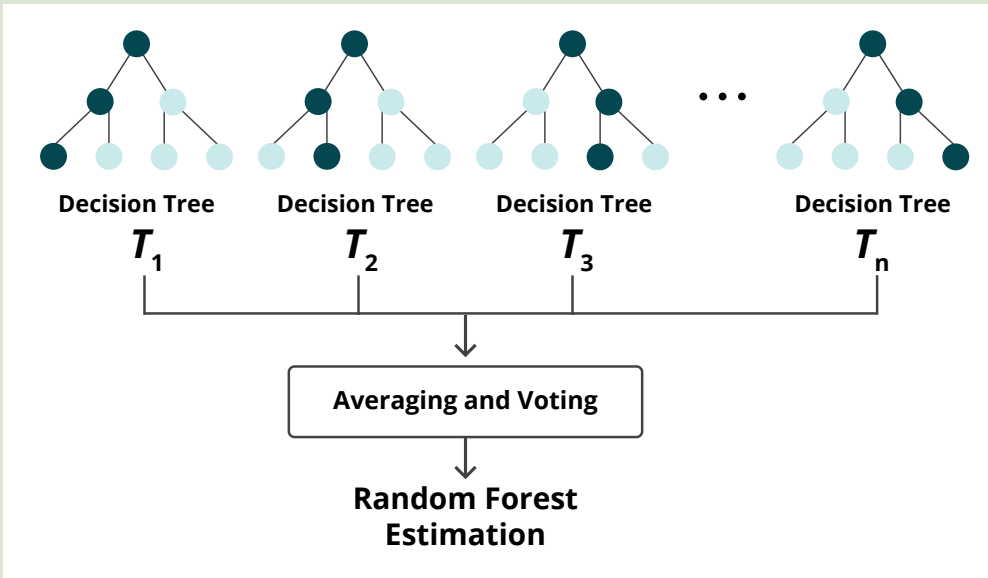

Figure 3.1 Random forest estimators (source: Kim and Kim 2022)

We used Ranger and Tidy Models packages in R to develop the random forest model (overview in Figure 3.2). The data were split into:

- a training set consisting of 60% of the sample
- a 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 1,000.

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 (root mean square error (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 made a refined grid based on the grid's best performing sections. For example, we found that the:

- number of predictors were best between 0 and 5
- 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 was when:

- number of predictors is 1
- number of data points in a node is 31.

The chosen model had an R-squared of .36 and an RMSE of .52 for the test data.

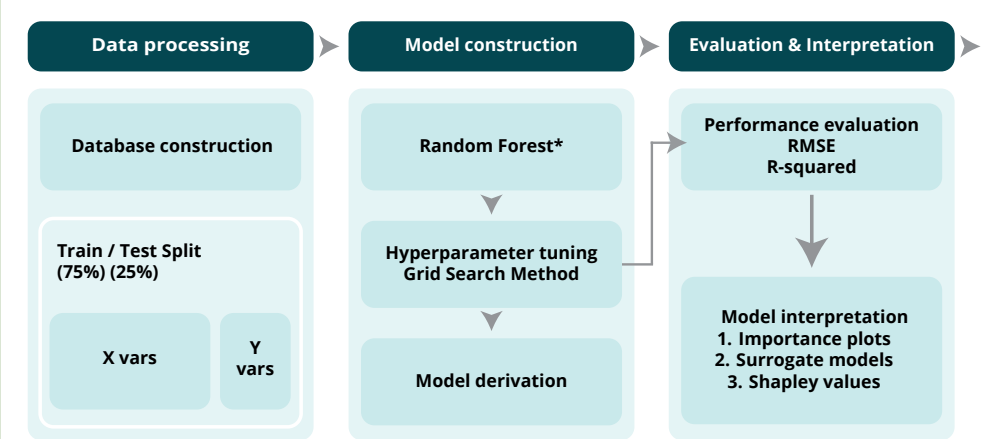

Figure 3.2 Model development

Note: we also explored other machine learning techniques – see appendix 7.2.

Interpretability

We define an interpretable technique as one that results in a model that humans can understand in terms of the reasoning behind the predictions and the decisions made by the model. Unfortunately, there is a trade-off between accuracy and interpretability, as shown in Figure 3.3.

The simplest models are very interpretable, for example, Ordinary Least Squares regression (OLS) and decision trees. We can 'read' their outputs and quickly understand the main message. For example, the regression models defined in section 3.2.2 are interpretable because it is possible to predict the value of the dependent variable for any set of independent variable values. As a result, we can understand which variables are important in making the prediction and judge their relative magnitude and direction.

However, 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 the results are harder to interpret (hence the label 'black boxes').

There are techniques to improve the understanding of the model and the relationships between input and output variables that help to make black box models more interpretable.

The first approach is to quantify variable importance scores. These indicate a variable's importance in making a prediction. They are calculated by:

- removing the variable from the model
- observing the change in predictive accuracy (the error term).

A large increase in the error term (large reduction in prediction accuracy) means 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 found 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.

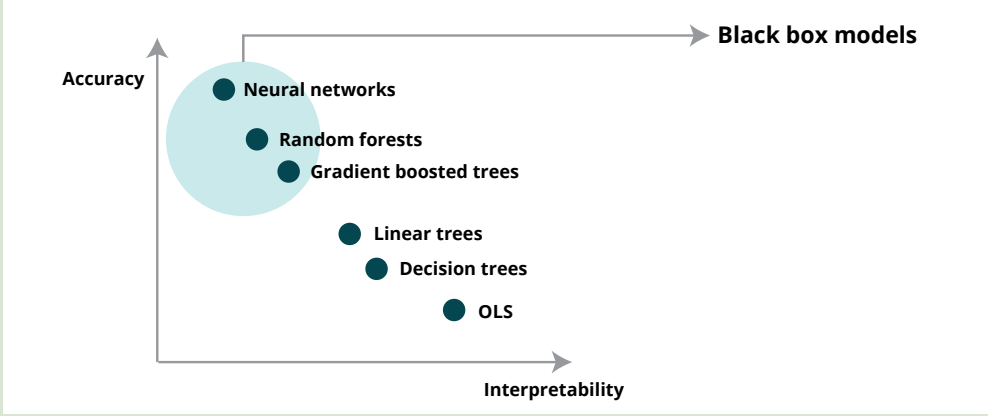

Figure 3.3 Black box machine learning algorithms

A problem with measures of variable importance is that they cannot be easily interpreted, even with respect to the direction of their contribution to predicting the outcome. Additive Shapley values overcome that gap.

A Shapley value represents the average marginal contribution of the variable to the prediction made for one observation. For example, in the case with 3 variables (A, B and C), calculating the Shapley value for variable A would involve estimating its effect on prediction for each subset of variables, and then average 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 (Molnar, 2023)

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
- 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
- $\hat{y}_*^{(i)}$ is the prediction of the black box model
- $\check{y}$ is the mean of the black box model predictions
- SSE stands for sum of squares error and SST for sum of squares total.

The Rsquared measure can be interpreted as the percentage of variance in the predictions that is captured by the surrogate model. If Rsquared is close to 1 (= low SSE), then the interpretable model approximates the behaviour of the black box model very well. In this case we may replace the complex model with the interpretable model. If the Rsquared 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.

4 Results

credit: Majken Søgaard

4 Results

We present 3 main sets of results. The first is an analysis of the importance of environmental variables as predictors of consumption. It uses random forests to explore the underlying heterogeneity in the income generating process. The results are made interpretable using:

- variable importance scores
- Shapley values
- surrogate models.

In the second set of results we describe the clusters formed by the surrogate model, including differences in their vulnerability to poverty.

Finally, we present the LASSO estimates of the effect of selected predictors of vulnerability to poverty for each of the clusters defined by the surrogate model.

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

We used 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).
- climatic variables:
  - temperature in first month of rainy season and its variance during 2010 to 2020
  - number of wet days in first month of rainy season and its variance during 2010 to 2020
  - number of wet days in rainy season
  - precipitation in first month of rainy season
  - precipitation in the rainy season.

We estimated 3 different machine learning models:

- 2 interpretable models (OLS and regression tree)
- one black box model (random forest).

Each model was estimated on:

- a training set (60% of the sample, used to tune the different parameters of each model, as discussed above)
- its accuracy with key performance metrics such as Rsquared and the RMSE evaluated in a test set (40% of the sample).

Table 4.1 shows the performance metrics for each of the machine learning algorithms. The best performing model was the random forest. We used this model to estimate vulnerability to poverty and to interpret predictors of vulnerability.

Table 4.1 Machine learning algorithm performance metrics

	Rsquared	RMSE
OLS	0.22	0.57
Regression tree	0.19	0.58
Random forest	0.34	0.51

4.1.1 Interpreting predictors of poverty

Random forest algorithms are complex and are not interpretable (when compared, for example, with OLS regression or regression trees). Consequently, they do not easily identify the main predictors of poverty. As mentioned, we used 3 approaches to understand which environmental variables are most important in predicting consumption:

- variable importance scores
- additive Shapley values
- surrogate models.

Variable importance scores

Table 4.2 shows the variable importance scores, defined in the previous section. The table ranks the 10 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. Importantly, 2 of the top variables are measures of climatic risk, rather than weather realisations in the year of the survey:

- Variance of temperature, measured during 2010 to 2020.
- Variance of number of wet days, measured during 2010 to 2020.

Other variables (such as soil properties, forest cover) are much less important.

Table 4.2 Top 10 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 4.1 presents additive Shapley values for the 10 most important variables, ranked by their global Shapley value.

Shapley values are plotted for every observation in the dataset. So, there is a distribution of estimates of the effect of each variable, which are:

- either positive or negative (shown by their position with respect to the axis at 0)
- either high (in purple) or low (in yellow).

The negative importance of temperature variance in the first month of the rainy season (a negative effect of climatic risk) reflects a larger frequency of individual high estimates (in purple) with a negative sign (to the left of the 0-axis).

Figure 4.1 allows us to draw some conclusions. For example, contributing positively to consumption are:

- the negative effect of variability of temperature in the first month of the rainy season
- high values of wet days and precipitation in the first month of the rainy season.

To summarise, ranking variables after the Shapley values has the same message as the ranking provided in Table 4.2. In particular:

- the first 5 top variables reflect the importance of weather in the first month of the rainy season
- climatic risk (as measured by the variance of the 2 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.

Surrogate models

The final approach used to interpret the random forest model is the surrogate model. Figure 4.3 shows the results of this approach. As explained, we:

- used the random forest prediction of consumption (rather than observed consumption) as the outcome variable
- estimated an interpretable model – in our case, a regression tree which separated data into cohorts based on different conditions.

The algorithm split the sample by minimising a loss function (the root of the sum of squared errors). Starting at the node at the top of the tree, the sample was continuously split by different conditions until reaching a leaf node (the nodes at the bottom of the tree). These splits continue until there was no further improvement in predictive power large enough to more than offset the added complexity 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, making the interpretation of the splits in the decision tree shown in 4.2 informative.

Table 4.3 Surrogate model: parameterisation 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

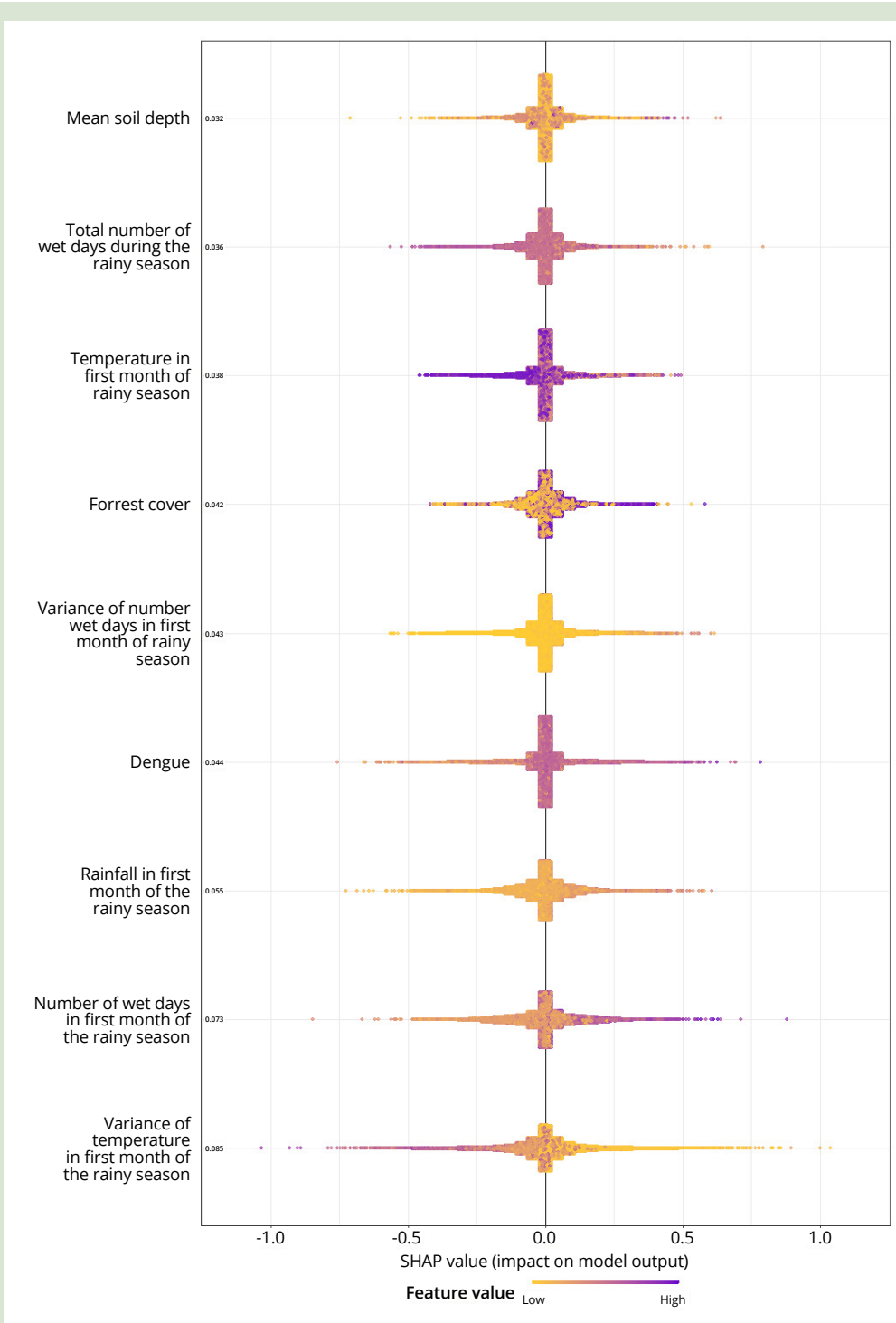

Figure 4.1 Additive Shapley values (top 10 variables, from least (top) to most important)

The relative importance of different predictors is largely in line with the other 2 approaches. The full sample is first split into 2 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 (for example, higher climatic risk) form one group (39% of the sample). This group will have, on average, lower income than those who exhibit lower values of the split variable. They form a different group (61% of the sample). This logic can be followed until reaching the leaf (final) nodes.

We can make some observations from the leaf nodes:

- First, the regression tree succeeds in creating groups with meaningful differences in consumption: the mean consumption per capita per year in the lowest-income group is US$270.00 which is less than one third of the average consumption per capita per year of the higher-income group of almost US$900.00.
- 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). Our findings from the Shapley values and the original decision tree corroborate this conclusion.
- Third, the 2 higher-income groups together include approximately 60% of the sample. They are the only groups generally above the poverty line. A small number of splits characterise these groups:
    - low risk in terms of temperature
    - 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.

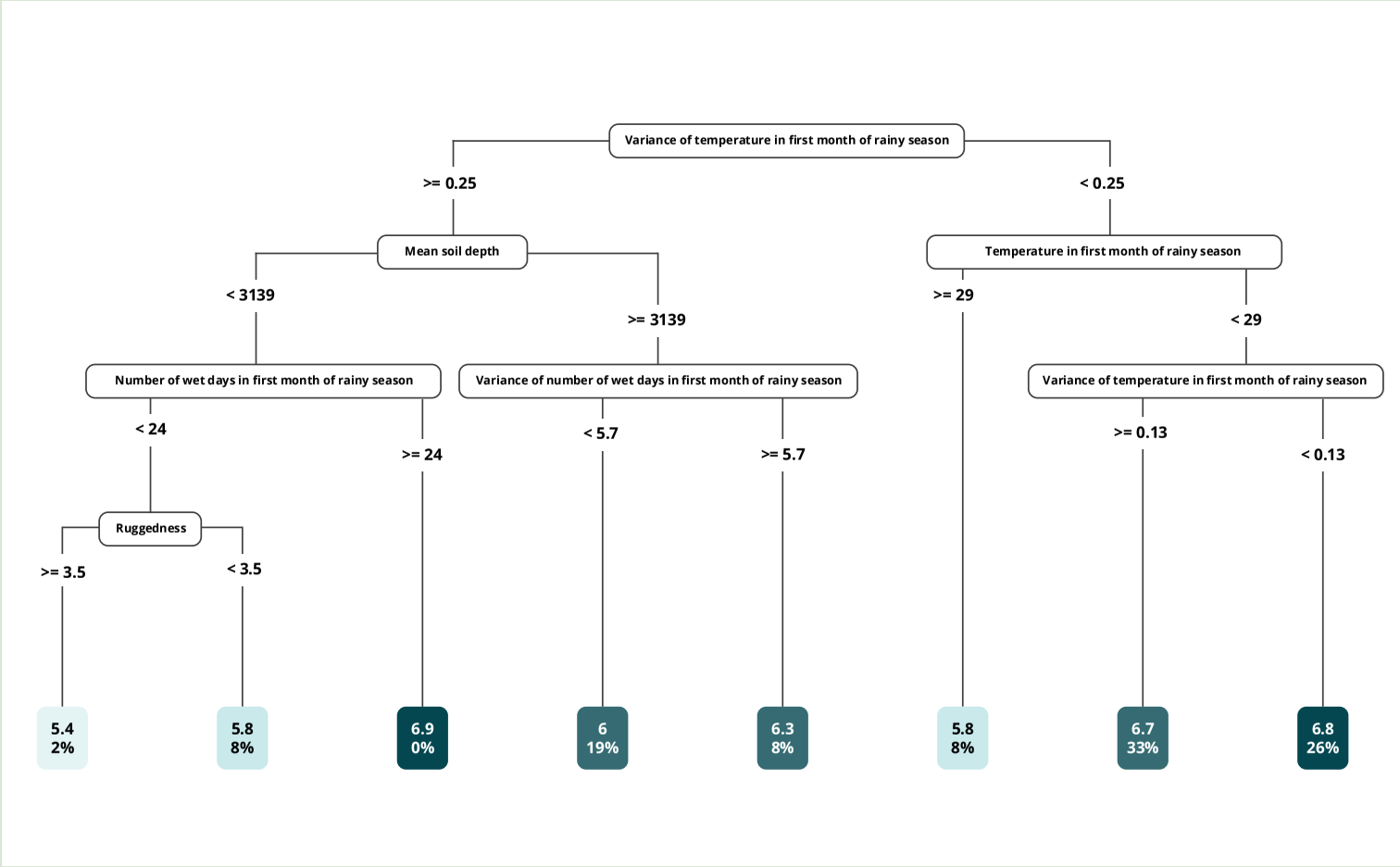

Figure 4.2 Predicting income: surrogate model, using a regression tree

4.2 Characterising income groups

The surrogate model created 8 groups (corresponding to the final leaf nodes). Figure 4.3 shows their distribution across the 7 South-East Asian countries included in the analysis. Lighter tones of blue show higher values of mean consumption per capita. Table 4.4 presents some descriptive statistics for each group.

The first conclusion is that while poverty is very different across groups, its prevalence is reduced in a linear way with rises in income. Care must be placed when interpreting regression results when units are substantially different (and these are substantively different units by construction). However, this result supports earlier analyses claiming 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 4.4 allows us 2 additional conclusions.

The first is that biodiversity degradation and forest cover seem to follow an inverse-U relation with income. They begin quite high in the lower groups then decrease before rising again in the latter groups. Interestingly, compared to groups 6 and 7, Group 1 is quite high in terms of slope and elevation and ruggedness. This may 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 or household size. Similarly, in terms of physical capital we are not able to identify a distinct pattern in terms of the housing index nor the productive asset index.

Figure 4.4 shows the vulnerability profile for households in each of the 8 groups:

- Along the y axis is the percentage of total observations in the sample.
- Along the x axis is the probability that the household will fall below the poverty line.

In our analysis our focus is on how to change the vulnerability profile of these groups, by understanding:

- how we can shift the distribution of the probability of being poor
- how this distribution is influenced by shocks
- what are the policy implications of this analysis in terms of developing social protection which keeps vulnerable households from falling below the poverty line.

Increasing consumption can change vulnerability to poverty. For example, to move from Group 1 to Group 2, we need to change mean consumption by approximately US$72.00. To move from Group 4 to Group 5, we need to change mean consumption by approximately US$123.00.

We can inspect the surrogate model and observe key thresholds that split households into different groups. 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, and it is unclear whether there are technological solutions that may support the development of 'cargo nets'. Thus, the immediate policy implication is that shifting the household to the left of the distribution (reduce their probability of falling into poverty) may require social protection programs.

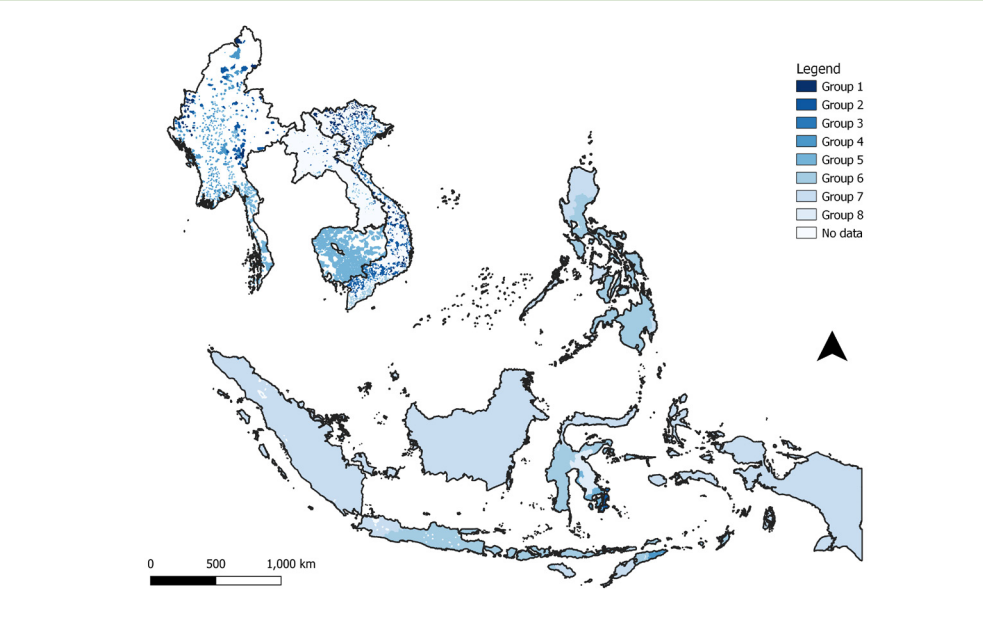

Figure 4.3 Spatial distributon of consumption groups, as defined by the surrogate regression tree

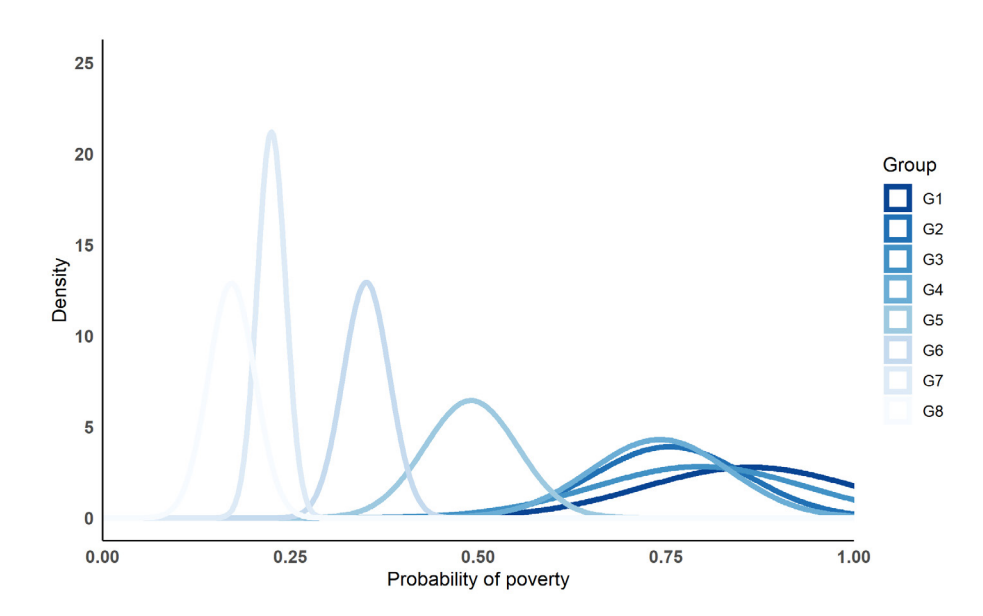

Figure 4.4 Vulnerability to poverty for different cohorts

Table 4.4 Characterising 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 (US$ 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	31.90	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

4.3 Predicting vulnerability to poverty

We used LASSO regression to identify the best predictors of the different moments of the distribution of income (and vulnerability to poverty as a result) in each of the 8 groups. Table 4.5 shows the results when the outcome variable is mean consumption. Table 4.6 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 4.5 Summary of variables by moments of the distribution: mean consumption

		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 4.6 Summary of variables by moments of the distribution: variance of consumption

		Variance		
		Negative (low risk)	Inconclusive	Positive (high 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

4.3.1 Mean consumption

Table 4.7 shows output from 8 individual LASSO regressions for the different groups 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.

Four out of the 10 variables have the same direction across all 8 groups. The remaining 6 variables have varied direction across the 8 groups. This suggests that households in different groups may respond differently to different interventions.

The interpretation of variables that are consistent in direction across all groups is that:

- older household heads and larger household are related with lower mean consumption
- university education and productive assets are related with higher mean consumption.

These results are consistent with the economic theory that having better inputs (physical and human capital) will lead to better outputs (income and consumption).

With variables that are inconsistent in direction across all groups, a female household head and erosion may be the hardest to understand. The coefficients of primary and secondary education are generally 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)
- forest cover has a negative effect on mean consumption, possibly because smallholder farmers have access to agricultural land.

Table 4.7 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

4.3.2 Variance of consumption

Table 4.8 shows output from 8 individual LASSO regressions for the different cohorts where variance of consumption is the outcome variable. The results show the size and the magnitude of the effect can be different for each variable across each of the groups. Two out of the 10 variables have the same direction across all 8 groups. The remaining 8 variables have varied direction across the 8 groups.

The interpretation of variables that are consistent in direction across all groups is that:

- larger household sizes are related with lower variance in consumption
- higher productive assets are related with higher variance in consumption.

These results suggest that members of larger households can support each other to reduce the variance of income. However, this insurance comes at a cost of lower mean consumption. Higher productive assets likely enable farmers to grow their income when the conditions are right but at the cost of exposing them to greater risk of lower incomes when they are not.

The interpretation of variables that are inconsistent in direction across all groups is that a female household head appears to reduce variance of consumption in the very low-income groups (Group 1 and Group 2). However, this variable relates 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, forest cover mostly reduces variance of income and biodiversity increases the variance of income.

The general conclusion is that there appears to be a fine line between risk and average return. Some variables give farmers the opportunity to get more income (which is positive) but there needs to be some form of protection so they can take these risks, because sometimes they will not succeed.

Table 4.8 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

5 Conclusions and recommendations

credit: Jeoffrey Maitem

5 Conclusions and recommendations

5.1 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 South-East Asia.

The different approaches used to estimate and interpret the main predictors of these differences ranked the importance of the variables included in the analysis in similar ways. Taken together, they suggest that:

- greater ruggedness is a major determinant of poverty (confirming the perception on ongoing differences between uplands and lowlands)
- conditional on this difference, production conditions in the first month of the wet season is key for determining income.

Some of the climatic variables are year-to-year levels, so are akin to weather shocks (for example the number of wet days in the first month of the wet season). Others reflect underlying climatic risk (such as the variance of the number of wet days in the first month of the wet season, estimated over 10 years). Hence, both shocks and risk matter to explain poverty in our cross-sectional data:

- Risk because it shapes investment decisions.
- Shocks because they reflect limited capacity to smooth income/consumption.

Linking these results with spatially-explicit models of changes in climate may guide future demand for both safety and cargo nets.

Discussing vulnerability to poverty has the advantage of focusing attention on more than snapshots of welfare, as measured by expected income. It forces us to discuss other characteristics of its distribution. In our data, this suggests 2 conclusions.

First, that increasing expected income (growing it or moving to a different group with higher income) is still, at least conceptually, an important insurance strategy. Households in groups with higher expected income have much lower probability of living with poverty.

Second, that no environmental condition seems to predict higher expected income and lower variance of income at the same time. This suggests that bundles of solutions to potentially important trade-offs may be needed to reduce vulnerability to poverty in rural South-East Asia.

5.2 Recommendations

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

Addressing this question seems central to ensuring minimal disruptions to the production of what remains the staple food in this part of the world. This may require either:

- new technologies (such as drought resistant varieties or varieties with a different production cycle)
- new institutions (such as better-functioning labour or machine rental markets, perhaps using digital technologies)
- a combination of both.

How to cope with risk is a longstanding question in both agricultural and development economics because agricultural production relies on weather, and agricultural production is still important in rural economies. The importance of this question is supported in our analysis.

Given the perceived increase in climate variability, understanding the scope for insurance markets to function better seems increasingly relevant, even if changes in climate makes defining such products more difficult.

Ongoing work on index insurance, typically directed at one crop or activity at the time, seems to be successful when that crop or activity dominates the livelihood portfolio. For example, East African pastoralists insuring their livestock against weather shocks through the Index Based Livestock Insurance.

Nevertheless, uptake of such insurance products remains disappointingly low in most cases. This undermines their capacity to 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 this case, 'livelihood insurance' may be much more attractive.

The relationship between poverty reduction and biodiversity conservation seems particularly difficult to address. Our results are correlations (and must be interpreted as such). However, a likely interpretation is that previous agricultural research and development bypassed households in areas that are still relatively rich in biodiversity. Previous research and development perhaps looked 'under the light' and focused on increasing yields in areas of greater return on such investments, for example on alluvial plains. If correct, existing biodiversity reflects the lack of similar technologies.

In a time of re-wilding and of protecting 'half the Earth', societal constraints make it unlikely that agricultural growth will follow the same technological path. Given this, it is important for the highest-poverty areas of rural South-East Asia to ask how to grow income while minimising negative impacts on remaining nature remains.

Particularly where poverty-environment trade-offs seem most relevant, research could be done on:

- how to develop different technologies that aim to address agriculture multifunctionality (although this is an imprecise concept)
- how to create new markets designed to reward the provision of environmental services
- a combination of both.

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

In the absence of technological or institutional changes, and with migration remaining a limited livelihood option, we should expect a larger share of the rural population to be driven into unacceptable levels of welfare with the future increases in:

- climate variability
- frequency of shocks
- degradation in natural capital.

Safety nets are the clearest form of a subsidy. They will be needed more than ever, either as an ongoing poverty alleviation strategy or as emergency payments intended to minimise the consequences of shocks. Hence, we suggest a better, simpler, and perhaps less ideological way to think about the interest of cargo vs safety nets. 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 recognise that, in this debate, the 2 approaches are now at very different starting points in terms of the strength of evidence supporting their use. The credibility of cash transfers benefit from the ongoing and rigorous evaluation of their impacts, starting with the evaluation of Mexico's Progresa in 1997. Very few examples of cargo nets can claim similar support. This evidence is almost non-existent in some cases, for example in the longstanding debates about the impact of nature conservation.

Conclusions about policy impacts are more credible the lower 'we can go' with linking socio-economic data on household consumption with the environmental data. In Indonesia and the Philippines, we are limited by the size of the spatial units at which we can 'locate' the households. So, our conclusions about the identification of distinct income cohorts in those countries must be interpreted with more care than in other contexts (in mainland South-East Asia for example).

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 2 countries – to collect yearly data on income and consumption – to study poverty dynamics using approaches such as pseudo-panels.

6 References

credit: Conor Ashleigh

6 References

Azeem MM, Mugera AW and Schilizzi S (2018) 'Vulnerability to multi-dimensional poverty: An empirical comparison of alternative measurement approaches', *The Journal of Development Studies*, 54(9):1612–1636.

Banks J, Blundell R and Brugiavini A (2001) 'Risk pooling, precautionary saving and consumption growth', *Review of Economic Studies*, 68(4):757–779.

Barrett CB (2005) 'Rural poverty dynamics: development policy implications', *Agricultural Economics*, 32:45–60.

Barrett, CB (2021) 'Overcoming global food security challenges through science and solidarity', *American Journal of Agricultural Economics*, 103:422–447.

Barrett CB, Benton TG, Fanzo J, Herrero M, Nelson RJ, Bageant E, Buckler E, Cooper K, Culotta I, Fan S, Gandhi R, James S, Kahn M, Lawson-Lartego L, Liu J, Marshall Q, Mason-D'Croz D, Mathys A, Mathys C, Mazariegos-Anastassiou V, Miller A, Misra K, Mude AG, Shen J, Sibanda LM, Song C, Steiner R, Thornton P and Wood S (2020) S*ocio-technical innovation bundles for agri-food systems transformation. Report of the international expert panel on innovations to build sustainable, equitable, inclusive food value chains*, Cornell Atkinson Center for Sustainability and Springer Nature, Ithaca, NY, and London.

Bigio S, Zhang M and Zilberman E (2020) Transfers vs credit policy: Macroeconomic policy trade-offs during COVID-19, research working paper 27118 [online document], National Bureau of Economic Research.

Boucher S, Carter M, Flatnes JE, Lybbert T, Malacarne J, Marenya P and Paul LA (2021) Bundling genetic and financial technologies for more resilient and productive small-scale agriculture, research working paper 29234 [online document], National Bureau of Economic Research.

Brown CS, Ravallion M and van de Walle D (2020) Can the world's poor protect themselves from the new coronavirus? Research working paper 27200 National Bureau of Economic.

Caro T, Rowe Z, Berger J, Wholey P and 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 ER and Waibel H (2016) 'Contract farming and vulnerability to poverty among oil palm smallholders in Indonesia', *The Journal of Development Studies, 52(5):681–695.*

Chaudhuri S, Jalan J and Suryahadi A (2002) Assessing household vulnerability to poverty from cross-sectional data: a methodology and estimates from Indonesia. Discussion paper series 0102-52 [online document], Department of Economics, Columbia University.

Collier P and Dercon S (2014) 'African agriculture in 50 years: Smallholders in a rapidly changing world?' *World Development*, 63:92–101.

Deaton A (2018) *The analysis of household surveys (reissue edition with a new preface): A microeconometric approach to development policy, World Bank Group, Washington, DC.*

Deaton A and Zaidi S (2002) *Guidelines for constructing consumption aggregates for welfare analysis* [online document], World Bank Publications.

Grosh M and Glewwe P (2000) Designing household survey questionnaires for developing countries [online document], World Bank.

Hanlon J, Barrientos A and Hulme D (2012) *Just give money to the poor: The development revolution from the global south*, Kumarian Press.

Hayami Y (2002) 'Family farms and plantations in tropical development', Asian Development Review, 19(2):67–89.

Hegwood M, Langendorf RE and Burgess MG (2022) 'Why win–wins are rare in complex environmental management', *Nature Sustainability*, 5:674–680.

Hengl T, Mendes de Jesus J, Heuvelink GB, Ruiperez Gonzalez M, Kilibarda M, Blagotić A, Shangguan W, Wright MN, Geng X, Bauer-Marschallinger B, Guevara MA, Vargas R, MacMillan RA, Batjes NH, Leenaars JGB, Ribeiro E, Wheeler I, Mantel S, and Kempen B (2017) 'SoilGrids250m: Global gridded soil information based on machine learning', *PLoS one*, 12(2):e0169748.

Imai KS, Gaiha R and 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 RE and Pope RD (1978) 'Stochastic specification of production functions and economic implications', *Journal of Econometrics*, 7(1):67–86.

Just RE and Pope RD (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 [online document], Warwick Monash Economic Student Papers.

Kim Y and Kim (2022) 'Explainable heat-related mortality with random forest and SHapley Additive exPlanations (SHAP) models', *Sustainable Cities and Society*, 79:103677.

Mehrabi Z, Ellis EC and Ramankutty N (2018) 'The challenge of feeding the world while conserving half the planet', *Nature Sustainability*, 1:409–412

Molnar C (2023) 'Global Surrogate' in Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. Second Edition [online document].

Nolan E and 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 and Chiwaula LS (2012) 'Health and vulnerability to poverty in Ghana: Evidence from the Ghana living standards survey round 5', *Health Economics Review*, 2(1):11.

Pendrill F, Gardner TA, Meyfroidt P, Persson UM, Adams J, Azevedo T, Bastos Lima MG, Baumann M, Curtis PG, De Sy V, Garrett R, Godar J, Dow Goldman E, Hansen MC, Heilmayr R, Herold M, Kuemmerle T, Lathuillière MJ, Ribeiro V, Tyukavina A, Weisse MJ and West C (2022) 'Disentangling the numbers behind agriculture-driven tropical deforestation', *Science*, 377(6611):eabm9267.

Pimhidzai O, Fenton NC, Souksavath P and Sisoulath V (2014) Poverty profile in Lao PDR: poverty report for the Lao consumption and expenditure survey 2012–2013 [online document], The World Bank.

Pritchett L, Suryahadi A and Sumarto S (2000) Quantifying vulnerability to poverty: a proposed measure, applied to Indonesia [online document], The World Bank.

Sharaunga S, Mudhara M and Bogale A (2016) 'Effects 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, Cornell SE, Fetzer I, Bennett EM, Biggs R, Carpenter SR, De Vries W, De Wit CA, Folke C, Gerten D, Heinke J, Mace GM, Persson LM, Ramanathan V, Reyers B and Sörlin S (2015), 'Planetary boundaries: Guiding human development on a changing planet', *Science*, 347(6223):1259855

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

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*, Oxford University Press, New York.

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

Zhu L, Hughes AC, Zhao X, Zhou L, Ma K, Shen X, Liu M, Xu W and Watson JEM (2021), 'Regional scalable priorities for national biodiversity and carbon conservation planning in Asia', *Science Advances*, 7(35):eabe4261.

7 Appendices

credit: Shutterstock

7 Appendices

7.1 Appendix 1: Survey information

Table 7.1 Summary of household income and expenditure survey, Cambodia

Component	Description
Region	South-East 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)
Normalisation	Per capita

Table 7.2 Summary of household income and expenditure survey, Indonesia

Component	Description
Region	South-East Asia
Survey name	National Socioeconomic Survey (SUSENAS)
Rounds	2010 to 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	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
Normalisation	Per capita
Note	in 2015 the reference period for certain items (health) was extended

Table 7.3 Summary of household income and expenditure survey, Laos

Component	Description
Region	South-East 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
Normalisation	Per capita (household members)

Table 7.4 Summary of household income and expenditure survey, Myanmar

Component	Description
Region	South-East 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 and 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 (cars for example)
Normalisation	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 7.5 Summary of household income and expenditure survey, Timor-Leste

Component	Description
Region	South-East 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; Occupational and employment status of household members
Notes	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
Normalisation	Per capita

Table 7.6 Summary of household income and expenditure survey, the Philippines

Component	Description
Region	South-East 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 households 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
Normalisation	Per capita

Table 7.7 Summary of household income and expenditure survey, Vietnam

Component	Description
Region	South-East 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
Normalisation	

7.2 Appendix 2: Machine learning

Machine learning is an algorithmic approach to predicting outcomes (such as consumption) based on some variables (such as stocks of produced, natural, and human capital). There are techniques which are more complex than regression techniques and which may improve the accuracy of predictions when linearity does not hold.

Predictions of the outcome variable were not the focus of this analysis. However, we were interested in understanding how the different variables contributed to the predictions these models made.

Machine learning techniques vary in their degree of interpretability. There are several definitions of interpretability 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. This is because it is possible to predict the value of the dependent variable for any set of independent variable values. 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 possible to obtain some understanding of the model and the relationships between input and output variables. This may be via:

- supplementing with other models known as surrogate models
- visualisation of coefficient relationships
- development of variable importance scores
- 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
- cubist models.

Following are details on each of the models we used and their respective level of interpretability.

7.2.1 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 of each observation are the terminal or leaf nodes. Regression trees are useful for determining important splits in variables and overall importance of features in a tree.

7.2.2 Linear tree models

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

7.2.3 Random forests

Random forests are ensemble methods (methods that combine models) with many decision trees – in this case, regression trees. To employ the concept of bootstrap aggregation (bagging), we select random samples (bags) from the data for each tree.

Since 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. Using bootstrap aggregation and then taking an average improves model performance as variance of the model decreases without increasing bias. This is particularly important given the sensitivity of regression trees to training data.

We can then compute importance scores for variables 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
- averaged over the forest.

To calculate the after out-of-bag error, the values of the feature are removed from the training data and the out-of-bag error is calculated again on the 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 the decrease in accuracy from removing a variable, and vice versa.

7.2.4 Gradient boosted trees

Gradient boosted trees are another ensemble method. Similar to other boosting methods they are built step-wise. They combine multiple models to reduce variance without adding additional bias.

Gradient boosted trees are:

- set with a weak learner (prediction is the average outcome)
- supplemented with more trees until the predictive ability is at its best (at this point adding another tree does not reduce the error).

Similar to random forests, data are placed in random subsets as each tree is produced. Where data are poorly modelled they are prioritised in new trees – this is where gradient boosted trees differ. 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. It improves the likelihood of all relevant variables being included.

For each tree, the gain on each node can then be calculated for each variable. The contribution can also be summed across trees to gain a measure of variable importance.

7.2.5 Cubist models

Cubist models are rule-based models 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 have intermediate linear models at each step of the tree.

Cubist models partition data into subsets with characteristics similar to the target variable and covariates. Then they make a series of rules to define the partitions. Each of these rules can be based on one or more covariates. The result is a set of regression equations that are general in form but local to the subsets of data partitioned. This lessens the overall error.

In the cubist models developed here we also employ a scheme like boosting called committees. In committees, 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
- an ensemble approach combining the cubist prediction and nearest neighbour prediction is used.

The rules used can be directly observed. So, the interpretability of cubist trees is higher relative to random forests and gradient boosted trees. However, using supplementary committee and nearest neighbour approaches does lessen this interpretability.


ACIAR


Australian
Aid