STATISTICAL LEARNING APPLICATIONS IN DEVELOPMENT ECONOMICS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Linden E McBride

August 2018

STATISTICAL LEARNING APPLICATIONS IN DEVELOPMENT ECONOMICS

Linden E McBride, Ph.D.

Cornell University 2018

The focus of this dissertation is the application of statistical learning and computational thinking to stubborn problems in development economics and welfare dynamics including the problems of poverty targeting, the identification of heterogeneous welfare dynamics, and the assessment of the risks and returns of diverse asset portfolios. Approaching such problems with statistical learning has allowed me to overcome some of the commonly imposed constraints and weaken some of the commonly made assumptions that prevent us from learning more about the empirical problem. By using out-of-sample validation and algorithmic model building, the first chapter presents an important step forward in making poverty targeting more accurate and efficient. The second chapter considers the theory and empirics of poverty and welfare dynamics more generally; the findings include several important implications for the study of welfare dynamics in diverse asset environments. The final chapter finds evidence consistent with a pattern in which households with greater initial asset holdings also hold a riskier portfolios and enjoy greater returns to their assets; however the analysis is limited by a poor accounting of human capital assets. The chapter concludes that allowing for the heterogeneity of assets, including non-physical assets, that may play a role in the livelihoods of households in developing countries is important. This dissertation demonstrates some of the ways in which algorithmic approaches can assist us in learning from the data.

BIOGRAPHICAL SKETCH

Linden McBride received an MA in International Development from American University, an MA in British and American Literature from Marquette University, and a BA in Sociology and Anthropology from St Mary's College of Maryland. Prior to beginning her doctoral studies at Cornell, Linden McBride worked as a Senior Research Assistant at the International Food Policy Research Institute in Washington, DC and served as a small enterprise development Peace Corps volunteer in Burkina Faso.

This thesis is dedicated to Ebou and Rachele Kantiono.

ACKNOWLEDGEMENTS

I gratefully aknowledge invaluable training, guidance, and support from my advisor, Chris Barrett, and dissertation committee members, David Just and Carla Gomes. Members of the Barrett Research Group – both past and present – have provided valuable feedback and support on all of my research initiatives over the course of my time at Cornell; I am grateful to them and to Chris for providing such a welcoming and challenging environment in which to grow as a scholar. I am grateful to co-authors Austin Nichols, Julia Berazneva, Megan Sheahan, and Leah Bevis for great ideas and productive collaborations, as well as mentors Paul Dorosh, Don Stabile, Asif Dowla, Ho Nguyen, and Marc Bellemare for inspiration and guidance. I am incredibly grateful to my classmates, friends, and future co-authors Jakina Debnam, Yanan Li, Jenn Cisse, and Xiaoli Fan for late night brainstorming sessions, feedback, and support. My husband, Drew Gower, provided extensive technical and moral support for which I am extremely grateful. Finally, I am grateful to my daughter, August, for strictly enforcing work/life balance.

This research was supported by funding from the African Development Bank STAARS project, supported by the Korean government through the Korea-Africa Economic Cooperation (KOAFEC) Trust Fund and National Science Foundation Grant Number CCF-1522054, "CompSustNet: Expanding the Horizons of Computational Sustainability". Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the African Development Bank or the National Science Foundation. All errors are my own.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Learning from the data

In his 2001 paper titled, "Statistical Modeling: The Two Cultures," Breiman makes a case for the use of algorithmic modeling to solve interesting problems. He argues, "if our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools." While the intended audience of this paper is the statistics community, as a doctoral student in development economics, I drew a great deal of inspiration from this rallying cry as well.

Development economists depend on a variety of methods to solve problems including theoretical models and simulations, random control trials (RCTs), and analyses relying on observational data. Of course, these methods are not mutually exclusive – many analyses combine multiple methods for robust findings – however, in the past we have been overly reliant on theoretical models and at present we uphold RCTs as the gold standard. Meanwhile, there is a preponderance of observational data available for the study of development economics. Observational data do not easily yield identification for causal inference; observational data can be messy and may defy our theoretical models. However, there is a great deal that we can learn from such data.

My objective throughout my doctoral research has been to avoid "exclusive dependence" on any one method and instead adopt a diverse set of tools that would allow me to engage in the study of welfare dynamics and poverty using

observational data. In particular, I have heeded Breiman's (2001) call for an embrace of algorithmic modeling for problem solving.

The empirical study of welfare[1] dynamics and poverty in developing economies is confronted by a number of challenges. For example, the empirical search for dynamic welfare thresholds is often frustrated by the fact that observations are rarely found in the neighborhood of a threshold (Barrett *et al* 2016). Likewise, popular non-parametric estimators that allow for flexibility in the estimation of welfare dynamics can produce analyses that ignore the empirical heterogeneity that may have given rise to those dynamics (Naschold 2013). Moreover, tools produced to monitor welfare dynamics and to target beneficiaries for anti-poverty interventions are limited by the fact that the data used to parameterize the tool is necessarily from a different sample than that where the tool will be applied (McBride & Nichols 2016). Finally, complex and diverse asset environments are not easily accounted for in theory or methods.

Fortunately, increasing data availability, advances in computing power, and powerful methods in statistical and machine learning allow us to tackle or bypass some of these challenges. Such methods are increasingly being applied to other areas of economics, policy, and estimation such as predicting who will benefit from hip replacement surgery, predicting recidivism for bail setting, and the use of recursive partitioning for the estimation of heterogeneous treatment effects (Mullainathan & Spiess 2017, Athey 2017). However, development economics has been slow to embrace these approaches.

The focus of my research has been the application of statistical learning and computational thinking to stubborn problems in development economics and

[1]Welfare is used in the sense of well-being throughout this work.

welfare dynamics, including the problems of poverty targeting, the identification of heterogeneous welfare dynamics, and the assessment of the risks and returns of diverse asset portfolios. Approaching such problems with statistical learning has allowed me to overcome some of the commonly imposed constraints and weaken some of the commonly made assumptions that prevent us from learning more about the empirical problem.

Overall, my approach has been to seek data-driven insights while remaining grounded in economic theory. I apply this approach to three problems within the study of poverty and welfare dynamics: poverty targeting, identification of heterogeneous welfare dynamics, and assessment of the riskiness and returns of diverse asset portfolios in developing economies. Below I discuss the contributions of each of these papers in turn.

1.2 Poverty targeting

So as to maximize the impact of limited program budgets, anti-poverty programs such social safety nets, cash transfers, and food security interventions are often targeted towards the poorest households in a population. However, accurate identification of beneficiaries meeting the program criteria is non-trivial in the developing world where payroll stubs and tax returns are not readily available and where household level means assessment is costly and time consuming. Proxy means test (PMT) targeting is a short cut to beneficiary targeting in such settings.

The objective of a PMT targeting tool is to quickly and accurately identify households meeting particular targeting criteria using a model parameterized

with already available, often nationally representative, data. For PMT tools to serve their purpose, it is important that they perform well not only within the data set or sample in which they were parameterized but also, especially, within the targeted population. In chapter 2, "Retooling poverty targeting using out-of-sample validation and machine learning", written in collaboration with Austin Nichols, we present evidence that the prioritization of the out-of-sample performance of proxy means test targeting tools can substantially improve their accuracy.

Using the United States Agency for International Development (USAID) poverty assessment tool and data, we use two methods to improve the out-of-sample performance of PMT tools: 1) the selection of a PMT model based on its cross-validation performance and 2) the use of stochastic ensemble methods, which have cross-validation built in, to develop the tools. Cross-validation offers insight into how the tool will perform out of sample. Stochastic ensemble methods in general, and quantile regression forests in particular, offer data-driven model building (variable selection) and extremely flexible model parameterization. In the country level case studies we analyze, prioritization of the out-of-sample performance of these targeting tools via these two methods significantly improves their accuracy.

This work, published in the World Bank Economic Review in 2016, has been at the forefront of an increasing emphasis on the use of machine learning for more accurate poverty targeting at the household level.[2] The World Bank, Innovations for Poverty Action (IPA), and USAID, among others, are increasingly interested in using such approaches for improved poverty monitoring, target-

[2]Note that contributions as Jean *et al* (2016) show the promise of using machine learning at the geographic level for the purpose of poverty targeting.

ing, and impact assessment.

The approach taken in chapter 2 sticks close to the foundational concept of proxy means tests: the objective is to parameterize a set of proxies for the households' means (Grosh & Baker 1995). Going forward, poverty targeting may altogether abandon the link to "proxy means" and simply prioritize predictive power. Recent developments in poverty targeting from IPA, such as work by Kshirsagar *et al* (2017), take this approach.

By using out-of-sample validation and algorithmic model building, the work offered in this chapter represents an important step forward in making poverty targeting more accurate and efficient.

1.3 Heterogeneity in welfare dynamics

Chapter 3 steps back from the mechanics of poverty targeting and considers the theory and empirics of poverty and welfare dynamics more generally. Chapter 3, "Heterogeneous welfare dynamics and structural transformation in Tanzania," takes a livelihoods-based approach to the study of welfare dynamics in the potential presence of market failures using a long panel (1991-2010) dataset from Kagera, Tanzania.

As mentioned above, numerous challenges confront the study of welfare dynamics and the direct observation of dynamic welfare thresholds in an economy. Generally, studies of welfare dynamics that are focused on non-convexities coupled with multiple financial market failures – the necessary conditions for multiple welfare equilibria and poverty traps – either consider only simple single-

asset, two-livelihood economies or run simulations with two-technology models. The contribution of chapter 3 is to approach the data with a fully general model that accommodates a number of market failures and places no restrictions on the number of technologies or livelihoods available in the economy.

In addition, empirical approaches to estimating welfare dynamics often either collapse meaningful heterogeneity or impose researcher's assumptions in identifying heterogeneous subgroups. This paper allows welfare dynamics to differ by livelihood group(s), as defined over productive asset holdings and their allocations, thereby avoiding the collapse of diverse livelihoods into a single population mean and allowing for empirically meaningful heterogeneity. Estimation of the returns to assets within and across identified livelihood strategies offers insights into the economy's welfare dynamics and general development. This approach also results in insights on the marginal returns to factors across the agricultural and non-agricultural sectors of the Tanzanian economy as well as insights on the role of migration in increasing household welfare.

However, the task of identifying livelihood groups faces challenges, one of which is to avoid the arbitrary imposition of empirically unsupported assumptions on the number and content of livelihood groups within the data. Therefore the analysis in this chapter relies on cluster analysis over household asset holdings and their allocations so as to define a set of livelihood strategies. I then estimate livelihood-conditioned returns to assets and associated welfare dynamics. Migration is included in the estimation as another, possibly non-convex, technology, as there may be a prerequisite cash, human/social capital, or other asset, threshold to migration.

The cluster analysis finds that, between 1991 and 2004, a subset of house-

holds moves from the dominant, farm-based, livelihood to a livelihood that allocates more assets to off-farm wage and entrepreneurial activities. In estimating marginal returns to assets across livelihoods, I find significant differences in returns to business, labor, and human capital assets by livelihood strategy, suggesting that households would realize locally increasing returns if they could switch livelihoods. Locally increasing returns between livelihood strategies suggests the presence of the type of non-convexities that give rise to multiple welfare equilibria. When including migration as an additional technology that can interact with livelihoods, I find that livelihood shifts play a greater role than migration in the increase in household consumption based returns to their business, labor, and human capital assets.

Analysis of welfare dynamics across livelihoods suggests conditional convergence and uncovers heterogeneous welfare dynamics that would be masked in an analysis of population mean dynamics alone. Although beginning with a flexible framework and employing a data driven strategy, my analysis confirms many of the stylized facts of the structural transformation literature, in particular the emergence of two sectors, sector-differentiated returns to labor and other factors, and catch-up in the low return sector.

The findings in this chapter have several important implications for the study of welfare dynamics. First, the evolution from a single livelihood in 1991 to two livelihoods in 2004 suggests that there may be serious limitations to analyses that depend on demarcated asset environments, which is the approach taken in much of the welfare dynamics literature. For example, welfare dynamics estimated over land and/or livestock assets alone in this dataset would lead to extremely misleading results for the emergent livelihood group, as hold-

ings of land and livestock collapse to near zero for these households while their welfare increases significantly.

The analysis also suggests that estimation of welfare dynamics at population means, without allowing for heterogeneity to emerge, masks policy relevant findings. Whether subsets of households are facing poverty traps, conditional convergence, or eventual convergence, appropriate policies and interventions can be designed to meet their needs, so long as we're able to observe them.

Finally, as is suggested in chapter 3, use of heterogeneous treatment effects (Athey & Imbens 2016, Wager & Athey 2017) to better evaluate how anti-poverty programming affects households is long overdue in assessment of development interventions and is an objective for my future work.

1.4 Risks and returns in welfare dynamics

Building on the methods and insights from chapter 3, chapter 4 tackles the problem of assessing the relationship among risk, returns, and welfare in a setting where households may diversify their asset portfolios to mitigate consumption risk. The relationship among risk, returns, and welfare has important implications for the reproduction of inequality and persistent poverty and therefore is critical to understand for effective anti-poverty policy making. If a household with a low initial asset endowment is constrained to low return economic activities (or, similarily, if higher return activities come with greater risk and household risk preferences induce the household to chose the low risk, low return activities), then not only will that household remain poor, but the gap between those with a low endowment and those with a high endowment will only grow

overtime, reproducing and exacerbating inequality.

The problem of estimating the riskiness of diverse asset portfolios is similar to that of accommodating heterogeneity in the estimation of welfare dynamics. The literature generally does not account for off-farm efforts to mitigate agricultural risk, and a positive correlation among risk, returns, and initial welfare is largely taken for granted in developing country settings. This chapter again relies on unsupervised learning methods to avoid imposing researcher assumptions on the data and to avoid collapsing or assuming away the heterogeneity in asset portfolios.

In collaboration with Leah Bevis, in this chapter I estimate the relationship among the value of initial asset holdings, expected returns, risk, and downside risk among Tanzanian households using Living Standards Measurement Study-Integrated Survey on Agriculture data from 2008-2013. Cluster analysis assists us in identifying the set of asset portfolios available in the data. In estimating the portfolio-specific moments of the conditional consumption distribution via fixed effects regression of consumption on a quadratic function of the assets, we are able to estimate the conditional consumption, variance, and skewness of the households' portfolio specific consumption distributions.

In addition, making the assumption that households utility of consumption follows constant relative risk aversion preferences, we calculate the household level risk premium associated with each portfolio. Finally, we non-parametrically estimate the relationship among initial wealth, expected returns, and the risk premium of each of the identified asset portfolios.

Our analysis identifies three distinct asset portfolios within the data: two

rural/agricultural portfolios that are differentiated by the value (high and low) of their initial asset holdings, and one urban/business portfolio. Across the agricultural portfolios identified in our analysis, we find evidence consistent with a pattern in which households with greater initial asset holdings also hold a riskier portfolios and enjoy greater returns to their assets. However, we do not find clear within-portfolio relationships between initial asset holdings, risk, and returns. Importantly, our analysis is confronted by an important limitation that often plagues analyses focused on asset-based welfare dynamics: we are unable to properly account for human capital assets in our analysis. The importance of human capital in this chapter echos the findings in the second chapter that suggest that reliance on traditional assets such as land and livestock in such analyses will only tell part of the story.

Going forward, we need to better account for the heterogeneity of assets that may play a role in the livelihoods of households in developing countries, including non-physical assets. This dissertation has taken a few steps in this direction by demonstrating the ways in which algorithmic approaches can assist us in learning from the data itself.

CHAPTER 2

RETOOLING POVERTY TARGETING USING OUT-OF-SAMPLE VALIDATION AND MACHINE LEARNING*

∗This chapter was written in collaboration with Austin Nichols; a version of this paper was published in the World Bank Economic Review in 2016; the published paper can be found at doi.org/10.1093/wber/lhw056

Accurate targeting is one of the most important components of an effective and efficient food security or social safety net intervention (Barrett and Lentz 2013; Coady, Grosh, and Hoddinott 2004). To achieve accurate targeting, project implementers seek to minimize rates of leakage (benefits reaching those who dont need them) and undercoverage (benefits not reaching those who do need them). Full means tests for identification of project beneficiaries can include detailed expenditure and/or consumption surveys; while effective, such tests are also time consuming and expensive. Proxy means tests (PMTs), a shortcut to full means tests, were first developed for the targeting of social programs in Latin American countries during the 1980s. PMTs have become common tools for targeting and poverty assessment where full means tests are costly (Coady, Grosh, and Hoddinott 2004). Today they are used by USAID (United States Agency for International Development) microenterprise project implementing partners, the World Food Program, and the World Bank, among many others for the purpose of poverty assessment, beneficiary targeting, and program monitoring and evaluation in developing countries (PAT 2014; WBG 2011).

PMT tools are typically developed by assignment of weights, or parameters, to a number of easily verifiable household characteristics via either regression or principal components analysis (PCA) in an available, nationally representa-

tive data set. In the regression approach, household-level income/expenditures or poverty status are regressed on household characteristics with the objective of selecting and parameterizing a subset of those characteristics to explain a significant proportion of the variation in expenditures/income or poverty status. In the PCA approach, the parameters are generated by extracting from a set of variables an orthogonal linear combination of a subset of those variables that captures most of the common variation (Filmer and Pritchett 2001; Hastie, Tibshirani, and Friedman 2009). Although each approach has its advocates, those interested solely in targeting tend to rely on regression approaches, while PCA has become popular among those interested in generating asset indices that may or may not be used for targeting. Note that the problem of developing tools for poverty targeting can be a fundamentally different problem from that of generating asset indices ; this paper speaks only to the problem of developing targeting tools.[1]

The regression approach to PMT tool development requires practitioners to select from a large set of potential observables a subset of household characteristics that can account for a substantial amount of the variation in the dependent variable. In practice, this is usually done through stepwise regression and the best performing tool is selected as that which performs best in-sample; more recently, efforts to validate in-sample-generated tools via out-of-sample testing have also been introduced (Schreiner 2006). Once a PMT tool has been developed from a sample from a particular population, the development practitioner can apply the tool to the subpopulation selected for intervention to rank or clas-

[1]For example, we might be concerned about endogeneity but not concerned about out-of-sample performance when generating an asset index to estimate the relationship between school enrollment and wealth, as in Filmer and Pritchett (2001). We have no such endogeneity concern when generating targeting tools because we are not attempting causal inference; however, out-of-sample performance is a primary concern.

sify households according to PMT score. This process involves implementation of a brief household survey in the targeted subpopulation so as to assign values for each of the household characteristics identified during tool development. The observed household characteristics, x_{ij}, are then multiplied by the PMT tool weights, θ_j, for each characteristic j to generate a PMT score for household i, as shown in Equation 2.1:

$$PMT\,score_i = \Sigma_j x_{ij}\theta_j. \tag{2.1}$$

In many applications, the calculated PMT scores are used to rank households from poorest to wealthiest and the poorest households are selected as program beneficiaries.[2] In the case of the USAID poverty assessment tools that will be described below, the use is more conservative: the PMT scores are used to quantify the number of households above and below an identified poverty threshold so as to ensure proper allocation of USAID funds (PAT 2014). The methodological improvements we propose in this paper apply to both types of uses for PMT tools.

Overall, the objective of a PMT tool is to quickly and accurately identify households meeting particular criteria in a new setting (but under the same data-generating process) using a model parameterized with previously available data. Therefore, for PMT tools to serve their purpose, it is important that

[2]There are several long-standing debates as to whether targeting tools, PCA type asset indices, and/or the use of consumption or income data in the regression approach capture long run economic status, permanent income, current consumption levels, current welfare, nonfood spending, or something else altogether. Lee (2014) points out that much of the theoretical support for these various claims is dubious and offers a theoretically grounded approach to the development of asset indices to measure poverty. As much as possible, we remain agnostic on the particular type of well-being that PMT tools capture while noting that the methods we discuss and the way in which we discuss them (e.g., their interpretation as capturing household poverty status) are standard in the literature and in practice.

they perform well not only within the data set or sample in which they were parameterized but also, especially, within the new data set or sample. In other words, high out-of-sample prediction accuracy must be prioritized in the development of PMT tools. In the fields of machine learning and predictive analytics, stochastic ensemble methods have been shown to perform very well out-of-sample due to the bias- and variance-reducing features of such methods.

In this paper, we present evidence that the prioritization of the out-of-sample performance of PMT targeting tools can substantially improve their out-of-sample accuracy. We propose two methods for this prioritization: (1) selecting a tool based on its cross-validation performance and (2) using stochastic ensemble methods, which have cross-validation built in, to develop the tool. Stochastic ensemble methods offer the additional feature, over and above traditional methods combined with cross-validation, of selecting the variables with which to build the tool, an otherwise time-consuming process. We take a set of PMT tools that have been developed by the University of Maryland IRIS Center (IRIS: Institutional Reform and Informal Sector) for the purpose of USAID poverty assessment for demonstration of these methods; however, the methods applied in this paper should be considered for PMT and other poverty targeting tool development more broadly.

We next present the USAID poverty assessment tool development and accuracy evaluation criteria; we then introduce the stochastic ensemble algorithms, regression forests, and quantile regression forests, that we apply to the problem of developing more accurate out-of-sample targeting tools; an explanation of our data and methods follows. We close with results and conclusions.

2.1 The USAID Poverty Assessment Tool

The development of the USAID poverty assessment tool (PAT) dates from 2000, when the US Congress passed the Microenterprise for Self-Reliance and International Anti-Corruption Act, mandating that half of all USAID microenterprise funds benefit the very poor (PAT 2014). In the context of this legislation, the very poor are defined as those households living on less than the equivalent of a dollar per day or those households considered among the poorest 50 percent of households below the countrys own national poverty line (IRIS Center 2005). Subsequent legislation required USAID to develop and certify low-cost tools to enable its microenterprise project-implementing partners to assess the poverty status of microenterprise beneficiaries. USAID engaged the IRIS Center at the University of Maryland in 2003 to create the tools.[3] To date, the IRIS Center has developed, and USAID has certified, tools for 38 countries.[4]

Using existing Living Standards Measurement Study (LSMS) data as well as survey data collected by IRIS, the IRIS Center developed country-specific PAT tools following the general PMT development procedure: they first identified a subset of household characteristics (approximately 15) from the larger data set of 70125 available observables that accounted for the greatest variation in household level income via an R-squared maximization routine, SAS MAXR;[5]

[3]The implementing partners who are required to make use of the PAT include "all projects and partner organizations receiving at least USD 100,000 from USAID in a fiscal year for microenterprise activities in countries with a USAID-approved tool" (PAT 2014). In 2013, this entailed 71 partners receiving a total of 110 million dollars (USAID MMR).

[4]Albania, Azerbaijan, Bangladesh, Bolivia, Bosnia and Herzegovina, Cambodia, Colombia, East Timor, Ecuador, El Salvador, Ethiopia, Ghana, Guatemala, Haiti, India, Indonesia, Jamaica, Kazakhstan, Kenya, Kosovo, Liberia, Madagascar, Malawi, Mexico, Nepal, Nicaragua, Nigeria, Paraguay, Peru, The Philippines, Rwanda, Senegal, Serbia, Tanzania, Tajikistan, Uganda, Vietnam, and the West Bank.

[5]The MAXR procedure operates by selecting and rejecting variables one by one with the objective of maximizing the improvement in a models R2 (SAS 2009).

they then selected for the final tool the parameters identified by the statistical modelwhether ordinary least squares (OLS), quantile regression, logit, or probitthat produced the highest predictive accuracy in-sample. In some cases, but not all, out-of-sample validation tests were performed.

The predictive ability of the resulting PMT model was evaluated against a number of accuracy criteriatotal accuracy, poverty accuracy, undercoverage, leakage, and the balanced poverty accuracy criterioneach of which is defined below. These criteria allow for ex ante evaluation of the generated poverty assessment tools via systematic consideration of each possible outcome/error type as presented in the confusion matrix in Table 2.1: true positive (the true very poor, $p = 1$, are identified by the tool as very poor, $\hat{p} = 1$); false negative (the true very poor, $p = 1$, are identified by the tool as non very poor, $\hat{p} = 0$); false positive (the true non very poor, $p = 0$, are identified by the tool as very poor, $\hat{p} = 1$); true negative (and the true non very poor, $p = 0$, are identified by the tool as non very poor, $\hat{p} = 0$).

Table 2.1: Poverty Prediction Outcomes

	$p = 1$	$p = 0$
$\hat{p} = 1$	True positive (TP)	False positive (FP)
$\hat{p} = 0$	False negative (FN)	True negative (TN)

Source: Standard confusion matrix

The classification literature has developed many metrics based on confusion matrices, such as that presented in Table 2.1, for the assessment of classification accuracy; the IRIS Center draws on standard metrics from the literature and has also developed a new metric for their evaluation of the PAT. Following the IRIS

Center and relying on the categories given in Table 2.1, the accuracy criteria we use to assess PAT performance are defined as follows: total accuracy (TA) is the sum of the correctly predicted very poor and the correctly predicted non very poor as a percentage of the total sample, (TA=(TP+TN)/(TP+TN+FP+FN)). Poverty accuracy (PA) is the correctly predicted very poor as a percentage of the total true very poor, (PA=TP/(TP+FN)). The undercoverage rate is the ratio of true very poor incorrectly predicted as non very poor to total true very poor, (UC=FN/(TP+FN)), while the leakage rate is the ratio of true non very poor incorrectly identified as very poor to total true very poor, (LE=FP/(TP+FN)). Finally, the balanced poverty accuracy criterion (BPAC) is the correctly predicted very poor as a percentage of the true very poor minus the absolute difference between the undercoverage and leakage rates, (BPAC=TP/(TP+FN)-|FN/(TP+FN)-FP/(TP+FN)|). These accuracy criteria are summarized in Table 2.2.

Table 2.2: Targeting Accuracy Metrics

Total accuracy	TA=(TP+TN)/(TP+TN+FP+FN)
Poverty accuracy	PA=TP/(TP+FN)
Leakage	LE=FP/(TP+FN)
Undercoverage	UC=FN/(TP+FN)
Balanced poverty accuracy criterion	BPAC=TP/(TP+FN)-|FN/(TP+FN)-FP/(TP+FN)|

Source: Authors' summary based on IRIS Center 2005.

Total accuracy, or one minus mean squared error, is very familiar to economists as a metric for model assessment. However, there are several reasons why total accuracy might not be an adequate metric for assessing the ac-

curacy of a poverty tool. Consider an example wherein a population of 100 includes 10 poor households. A tool that simply classifies the entire population as nonpoor would have a total accuracy rate of 90 percent, which seems quite good. However, this tool would have failed to identify a single poor household. Therefore, metrics beyond total accuracy are necessary for assessment of poverty tool performance; these additional metrics include poverty accuracy (also known as precision in the classification and predictive analytics literature) and undercoverage (false negative) and leakage (false positive) rates. In the example just given, the poverty accuracy of the tool would be 0 percent, and the undercoverage rate would be 100 percent. These additional metrics offer a better picture of the tools performance than does total accuracy alone. The BPAC combines these three metricspoverty accuracy, undercoverage, and leakageby penalizing the poverty accuracy rate with the extent to which the leakage and undercoverage rates exceed one another. The BPAC is an innovation of the IRIS Center; it was created to balance the stipulations of the Congressional Mandate against the practical implications of the assessment tools (IRIS 2005). The other criteria are standard in PMT development. However, it should be noted that IRIS computes leakage in an unconventional manner.[6]

PAT model selection for each country was ultimately made by IRIS based on the BPAC results in-sample. While we follow the prioritization of the BPAC criteria in the analysis that follows, the methods we propose can just as easily be used to meet other prioritized accuracy criteria.

[6]Whereas leakage rates are commonly computed as FP/(TP+FP), IRIS computes leakage rates as FP/(TP+FN). This adjustment to the denominator in the calculation of leakage rates has two consequences: 1) it can lead to calculated leakage rates that are greater than one, producing a heavy penalty in the calculation of BPAC where leakage occurs (it is not clear that IRIS intended for this outcome); 2) it keeps constant the denominator across poverty accuracy, undercoverage, and leakage rates, allowing IRIS to easily perform the addition and subtraction necessary for the BPAC calculation. We assume this was IRISs purpose in modifying the denominator.

2.2 Stochastic ensemble methods: Regression forests and quantile regression forests

Classification and regression trees are a class of supervised learning methods that produce predictive models via stratification of a feature (in the case of poverty tool development, a feature is a variable or characteristic) space into a number of regions following a decision rule (Hastie, Tibshirani, and Friedman 2009). A canonical and intuitive example of a classification tree is that of predicting, based on a number of features such as age, gender, and class, who survived the sinking of the Titanic.[7] While both classification and regression trees can be used to make predictions regarding the poverty status of households based on observable household characteristics, this paper focuses on regression and, in particular, quantile regression forests due to the advantages the latter offers in terms of making predictions about households concentrated at the lower end of the income distribution.

Regression trees operate via a recursive binary splitting algorithm as follows (Hastie, Tibshirani, and Friedman 2009): for N observations of response variable, y_i, and a vector of characteristics, x_{ij}, where $i = 1, 2, N$ is the number of observations and $j = 1, 2, J$ is the number of features, consider the splitting variable, x_j, and the split point, where $x_{ij} = s$, that define the half planes R_1 and R_2, as indicated in Equation 2.2:

$$R_1(j, s) = \{x_{ij} | x_{ij} \leq s\} \quad \text{and} \quad R_2(j, s) = \{x_{ij} | x_{ij} \leq s\} \tag{2.2}$$

[7] See Varian (2014) for an example. Many examples and data are also available at The Comprehensive R Archive Network at http://cran.r-project.org.

The algorithm selects x_j and s to solve the minimization problem,

$$min_{js}\left[min_{\frac{1}{n}\Sigma_i(y_i|x_i\in R_1(j,s))}\Sigma_{x_i\in R_1(j,s)}(y_i-\frac{1}{n}\Sigma_i(y_i|x_i\in R_1(j,s)))^2+\right.$$
$$\left.\min_{\frac{1}{n}\Sigma_i(y_i|x_i\in R_2(j,s))}\Sigma_{x_i\in R_2(j,s)}(y_i-\frac{1}{n}\Sigma_i(y_i|x_i\in R_2(j,s)))^2\right]$$

In words, the regression tree algorithm chooses the variable, x_j (the splitting variable), and the value of that variable, s (the split point), which minimizes the summed squared distance between the mean response variable and the actual response variables for the observations found in each of the resulting regions. In this manner, the algorithm effectively weights the response variables by the predictive value of the observations within each region (Lin and Jeon 2006). Once the optimal split in Equation 2.3 is identified, the algorithm proceeds within the new partitions.

One way to think about a regression tree is as an OLS regression for which one knows in advance all of the split variables and split points across which to partition, and then conditionally partition, the feature space, which therefore defines appropriate binary variables and interaction terms to capture these partitions. Such an OLS would return the same results as a regression tree built over the same data. However, such split variables and split points are not known in advance; therefore, what the regression tree algorithm offers over and above an OLS is a heuristic method for the selection of those variables, split points, and conditional splitsthe binary variables and their interactionswith which to build the model so as to minimize prediction error. To do this using OLS would require a stepwise regression that iterates and then conditionally iterates through

each split point of each variablea computationally intensive process.

The recursive binary splitting process of the regression tree can continue until a stopping criterion is reached; however, larger trees may overfit the data. In the case that we want to bootstrap over this algorithma good idea, as the algorithm may make different splitting decisions in different subsets of the datait becomes apparent that a bias for variance trade-off is made as we allow the trees to grow large.[8] A collection of larger trees will have high variance but low bias while a collection of smaller trees will have low variance but high bias.

Fortunately, in this setting, the bias-variance trade-off can be somewhat overcome via a process called bootstrap aggregation, or bagging. Bagging involves bootstrapping a number of approximately unbiased and identically distributed regression trees and then averaging across them so as to reduce the variance of the predictor. However, bagging cannot address the persistent variance that arises due to the fact that the trees themselves are correlated, as they were generated over the same feature space (Hastie, Tibshirani, and Friedman 2009). Consider, for example, a set of B identically distributed but correlated regression trees, each with variance σ^2. If ρ represents the pairwise correlation between the trees, then the variance of the average of these trees is $\rho\sigma^2+(1-\rho)/B\sigma^2$. As B grows large, the term $(1-\rho)/B\sigma^2$ will approach zero, reducing the overall variance (Hastie, Tibshirani, and Friedman 2009). However, the first term, $\rho\sigma^2$, persists (Hastie, Tibshirani, and Friedman 2009).

Reducing this persistent variance component of the bagged predictor is the innovation of random forests. Introduced by Breiman (2001), regression forests

[8] A variety of options for pruning trees exist to address these issues in a regression tree framework (Hastie, Tibshirani, and Friedman 2009). We dont discuss these here but move on instead to random forests, which address the problem without pruning.

improve the variance reduction feature of bagged regression trees by decorrelating the trees, and thereby reducing via a random selection of the features (variables) over which the algorithm may split. The number of random features available to the algorithm at any split is typically limited to one-third of the total number of features (Hastie, Tibshirani, and Friedman 2009); this is a tuning parameter of the algorithm.

Critically, in a random forest algorithm, the mean squared error of the prediction is estimated in the out of bag sample (OOB), the (on average) third of the training data set on which any given tree has not been built (Breiman 2001), in a manner similar to k-fold cross-validation. This OOB sample offers an unbiased estimate of the models performance out-of-sample.

The random forest training algorithm produces a collection of B trees, denoted as $\{T(x;\Theta_b)\}_1^B$, where Θ_b indicates the b^{th} tree. The regression forest predictor is then the bagged prediction

$$\widehat{f(x_i)} = \frac{1}{B}\Sigma_{b=1}^{B}\left[T(x_i;\Theta_b)\right]. \tag{2.4}$$

It has been shown that regression forests offer consistent and approximately unbiased estimates of the conditional mean of a response variable (Breiman 2004; Hastie, Tibshirani, and Friedman 2009). However, as elaborated by Koenker (2005), among others, the conditional mean tells only part of the story of the conditional distribution of y given X. Therefore, we also apply quantile regression forests, as developed by Meinshausen (2006), to our PMT tool development. Meinshausen (2006) draws on insights from Lin and Jeon (2006), who show that random forest predictors can be thought of as weighted means of the

response variable, y_i, as shown in Equation 2.5:

$$\widehat{f(x_i)} = \frac{1}{B}\Sigma_{b=1}^{B} T(x_i;\Theta_b) = \Sigma_{i=1}^{N} \frac{\Sigma_{b=1}^{B} w_i(x_i;\Theta_b)}{B} y_i. \tag{2.5}$$

In Equation 2.5, $w_i(x_i;\Theta)$ represents the weight vector obtained by averaging over the observed values in a given region $R_l, (l = 1 \ldots L)$. Application of the weight vector to the response variable is simply another way of considering the conditional averaging of the response variable, as represented in Equation 2.3 above and shown in Equation 2.6:

$$w_i(x_i;\Theta)y_i = \frac{1}{n}\Sigma_i(y_i|x_i \in R_l(j,s)). \tag{2.6}$$

With this insight, Meinshausen (2006) produces quantile regression forests, as a generalization of regression forests in which not only the conditional mean, but the entire conditional distribution of the response variable is estimated (Equation 2.7):

$$\widehat{f(x_i)} = \frac{1}{B}\Sigma_{b=1}^{B} T(x_i;\Theta_b) = \Sigma_{i=1}^{N} \frac{\Sigma_{b=1}^{B} w_i(x_i;\Theta_b)}{B} \mathbb{1}\left\{y_i \leq y\right\}. \tag{2.7}$$

Meinshausen (2006) provides a proof for the consistency of this method and demonstrates the gains in predictive performance of quantile regression forests over linear quantile regression. These gains are due to the fact that quantile regression forests retain all the bias-minimizing and variance-reducing components of regression forests in that they bootstrap aggregate across a great number of decorrelated trees; quantile regression forests additionally offer the ability

to make predictions across the conditional distribution. A quantile approach is particularly useful for the purposes of PMT tool development due to the fact that the very poor are often concentrated at one end of the conditional income distribution, far from the conditional mean.

The advantages that stochastic ensemble methods, such as the regression forest and quantile regression forest algorithms, offer over traditional PMT development tools include the selection of the variables that offer the greatest predictive accuracy without the need to resort to stepwise regression and/or running multiple model specificationsrather, the algorithms build the modeland built-in cross-validation via the out-of-bag error estimates.

Therefore, using regression forest and quantile regression forest algorithms, we expect to realize improvements in the out-of-sample targeting accuracy of the PAT. We note, however, that this method requires the critical assumption that the data-generating process remains unchanged between tool development and tool application. That is, the algorithm can perform well out of sample but not out of population. This limitation plagues any sample-based estimation routine.

2.3 Empirical method and data

We produce a set of country-specific examples from the survey data that was used by the IRIS Center to construct their PATs. We replicate the PAT development process by extracting the same variables that IRIS extracted from the same data sets and then generating identical estimation models. We are limited in our replication process to the use of LSMS data sets that are publicly available.

We have additionally constrained ourselves to the LSMS data sets for which income or expenditure aggregates are also publicly available due to the challenges of precisely replicating an income or expenditure aggregate that IRIS may have generated.

From the publicly available data sets meeting these criteria, we selected three nearly arbitrarily: the 2005 Bolivia Encuesta de Hogares (EH), the 2001 Timor Leste Living Standards Survey (TLSS), and the 2004-2005 Malawi Second Integrated Household Survey (IHS2). These data sets present a reasonable representation of the settings in which PATs have been developed. Each data set differs in number of observations, poverty level, and IRIS-selected household characteristics. The data are summarized in Table 2.3, where we can see that the number of household level observations ranges from 1,800 in East Timor to 11,280 in Malawi. Likewise, the USAID-defined poverty rates range considerably, from 24.2 percent in Bolivia to 64.8 percent in Malawi.

Table 2.3: LSMS Surveys and Variables Used in PAT Development and Replicated by Authors

County	Data	Obs.	Poverty rate (%)
Bolivia	2005 Encuesta de Hogares	4,086	24.03
Malawi	2004-2005 Second Integrated Household Survey	11,280	64.78
East Timor	2001 Timor Leste Living Standards Survey	1,800	44.73

Source: Authors summary based on the data indicated as well as reports from IRIS Center 2007, 2009, and 2012.

We provide the IRIS reported in-sample accuracy estimates for each country-level data set in each row 1 of Appendix Table A.1. These are the estimates on which the IRIS model selection was made. We provide the IRIS-reported out-

of-sample accuracy assessment results for each country in rows 24 of Table A.1. We replicate the IRIS in-sample models and report the replication estimates in each row 5 of Appendix Table A.1. Within-country comparisons of our replication estimates (Table A.1, row 5), with the estimates reported by IRIS (Table A.1, row 1), serve as a check on how well we have replicated the PAT tool development process. In the case of Bolivia, our replication estimates do not perform as well as those of IRIS; however, it should be noted that IRIS built the Bolivia PAT tool on a randomly selected subset of the data. We cannot replicate precisely the same random draw and so report the full sample estimates. The full sample replication does not perform as well as the half sample performance reported by IRIS, but that half sample is unusual in its high performance, and not representative of the thousand half sample splits we explored or that IRIS reported for their calculation of out-of-sample performance (see rows 2 through 4 of Appendix Table A.1 for Bolivia). For this reason, we are not concerned about spuriously overestimating the performance of our methods relative to those of IRIS and therefore retain this data set in our analysis. In the case of East Timor and Malawi, our replication estimates are very close to those reported by IRIS, and we are likewise not concerned about unfair comparisons of our methods with those of IRIS.

Our empirical approach is to randomly draw, with replacement, two samples of size $N/2$ from each country-level data set, producing a training sample and a testing sample. Over this split of the data, we first reproduce IRISs methods, training their preferred model in the training data and then testing it on 1,000 bootstrap samples of the testing data. However, instead of basing tool selection on in-sample performance as IRIS does, we perform k-fold cross-validation in the training sample and select as our preferred model the

one that produces the best BPAC in cross-validation. For this exercise, we use k-fold cross-validation; in particular, we produce 500 iterations of three-fold cross-validation, which entails training the model on two-thirds of the training data set and assessing performance in the remaining third of the training data set on which the model was not trained. We take this approach because it most closely approximates the out-of-bag error produced using the stochastic ensemble methods.

Following the method for out-of-sample testing used by the IRIS center, we test the classification accuracy of the cross-validation-selected tool using 1,000 bootstrapped samples of the testing sample. The out-of-sample performance of this tool in the testing sample is presented for each country in figures 13, as well as in Appendix Table A.1, rows 6 through 8. We refer to this approach of using cross-validation to select the best-performing model in the training sample as the "cross-validation" approach throughout remaining sections to distinguish it from both IRISs approach and from the stochastic ensemble method approach (note that stochastic ensemble methods also use cross-validation; however, it is referred to as out-of-bag error in that setting).

We next turn to the stochastic ensemble methods. Over the same split of the data as used for the cross-validation approach, the random forest and quantile regression forest models are built in the training sample where, for any given (x_i, y_i), an average of two-thirds of the training data are used to build bagged regression trees and the remaining third is reserved for out-of-bag, and therefore unbiased, running estimates of the prediction error over a forest of 500 trees.[9] We run the regression forest and quantile regression forest algorithms in R us-

[9]Five hundred trees is the default setting in the randomForest package in R. From casual observation, the OOB error has largely stabilized by the time the forest has reached 200300 trees; this observation is consistent with the literature (Hastie, Tibshirani, and Friedman 2009).

ing packages developed by Liaw and Wiener (2002) and Meinshausen (2016), respectively. We select our preferred model as that with the lowest BPAC error in the OOB sample. This model is then taken to the testing sample to assess classification accuracy. The performance of this tool in the testing sample is presented for each country in figures 13, as well as in Appendix Table A.1, rows 9 through 11.

We statistically compare the mean of the IRIS-reported bootstrapped accuracy estimates with those produced using both of our approaches to tool developmentthe cross-validation approach and the stochastic ensemble approachusing Tukey Kramer tests, selected to account for the family-wise error rate. The results are reported in Table 2.4.

Finally, so as to assess the robustness of our results to the poverty thresholds in each country, we report in Appendix Table A.2 the performance of our methods as compared with those of IRIS under two new poverty lines: one that is half the original poverty line and a second that is twice the original poverty line. We cannot observe actual IRIS tool performance metrics under these new poverty lines, but we estimate the best possible results IRIS could have gotten using their methods and preferred tools by adapting those tools to obtain the greatest BPAC under the new poverty lines. In practice, this means selection of the quantile that offers the best in-sample BPAC under the new poverty lines in Bolivia and Malawi. In the case of East Timor, we include a quantile regression approach along with IRISs preferred approach under the original poverty line, the probit model, because the probit performs poorly at the lower poverty line. This means we are comparing our cross-validation and ensemble method approaches to the best possible outcomes of the approach employed by IRIS.

2.4 Results

Results of the cross-validation (CV) and stochastic ensemble (SE) approaches to PMT tool development are displayed graphically in Figures 1, 2, and 3 and numerically in Appendix table A1. In both formats, we compare the out-of-sample bootstrap accuracy estimates of the IRIS-produced tools (rows 24 in the Table A.1) with those produced by each of our approaches. The confidence bars in each figure display the nonparametric bootstrap confidence intervals, where the lower bound is the 2.5th percentile and upper bound is the 97.5th percentile bootstrap estimate. Standard errors are reported in Table A.1. In addition, Tukey Kramer tests of the differences in the out-of-sample bootstrap means are reported in Table 2.4.

While cross-validation improves on the total accuracy of the IRIS-generated tool only in the case of Bolivia and the stochastic ensemble methods do not improve on the total accuracy at all (Figure 2.1, first graph), gains in poverty accuracy are observed using cross-validation across all countries and using stochastic ensemble methods in both East Timor and Malawi (Figure 2.1, second graph). Recall from the discussion above that total accuracy has serious limitations as a metric for assessing the performance of a poverty-targeting tool.

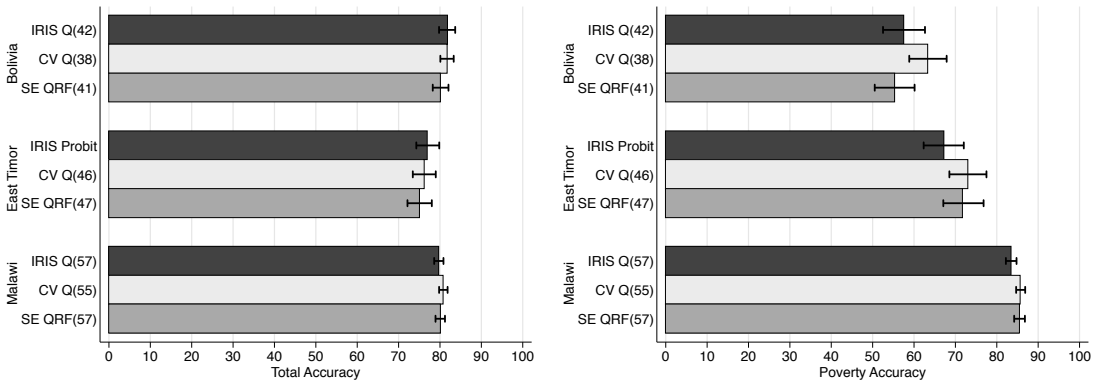
Table 2.4: Tukey-Kramer Tests of Equality of Bootstrap Poverty Accuracy and BPAC Means across Estimates

		Poverty accuracy		BPAC	
	Estimation	Difference	TK test statistic	Difference	TK test statistic
Bolivia	CV vs IRIS	5.79*	37.55	8.61*	28.20
	SE vs IRIS	-2.25*	-14.07	0.85	2.38
	CV vs SE	8.04*	54.14	7.76*	29.04
East Timor	CV vs IRIS	3.69*	23.89	2.78*	11.87
	SE vs IRIS	2.43*	15.43	1.29*	5.39
	CV vs SE	1.26*	8.40	1.49*	7.68
Malawi	CV vs IRIS	2.25*	59.06	2.19*	50.03
	SE vs IRIS	2.06*	49.11	1.43*	30.85
	CV vs SE	0.19	4.90	0.76*	17.51

Note: CV = cross-validation estimates; IRIS = IRIS reported estimates; SE = stochastic ensemble estimates.
∗ Indicates difference is significant at 1% significance level.

Source: Authors estimates using data and procedures detailed in the text.

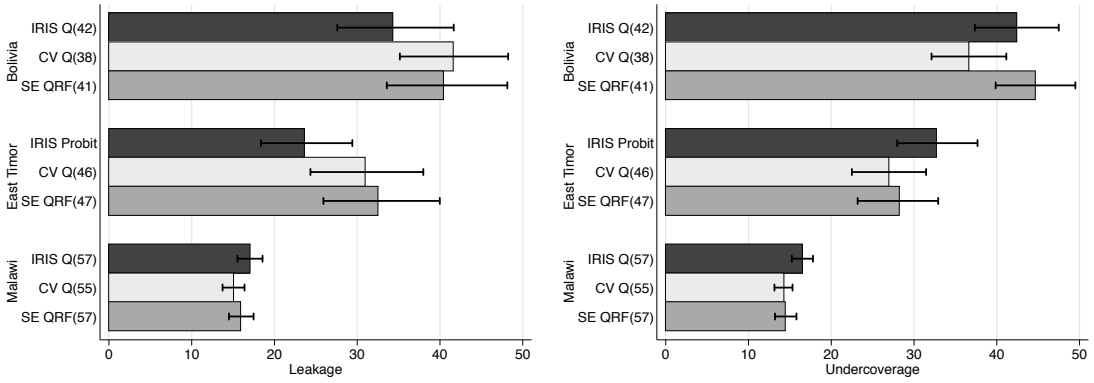
Figure 2.1: Total and Poverty Accuracy



(a) Total Accuracy (b) Poverty Accuracy

From Figure 2.2 (first graph), we can see that these gains in poverty accuracy are not without trade-offs: the leakage rates for the cross-validation and stochastic ensemble approaches are significantly greater than those reported for the IRIS-generated tools in both Bolivia and East Timor, meaning that these tools err on the side of classifying nonpoor households as poor. Given that leakage rates are heavily penalized by the IRIS accuracy metrics, these increases are not very surprising. Meanwhile, the cross-validation approach performs much better than IRIS's in terms of undercoverage rates; the undercoverage rate is decreased across all countries (Figure 2.2, second graph). The stochastic ensemble approach likewise outperforms IRISs in both East Timor and Malawi.

Figure 2.2: Leakage and Undercoverage

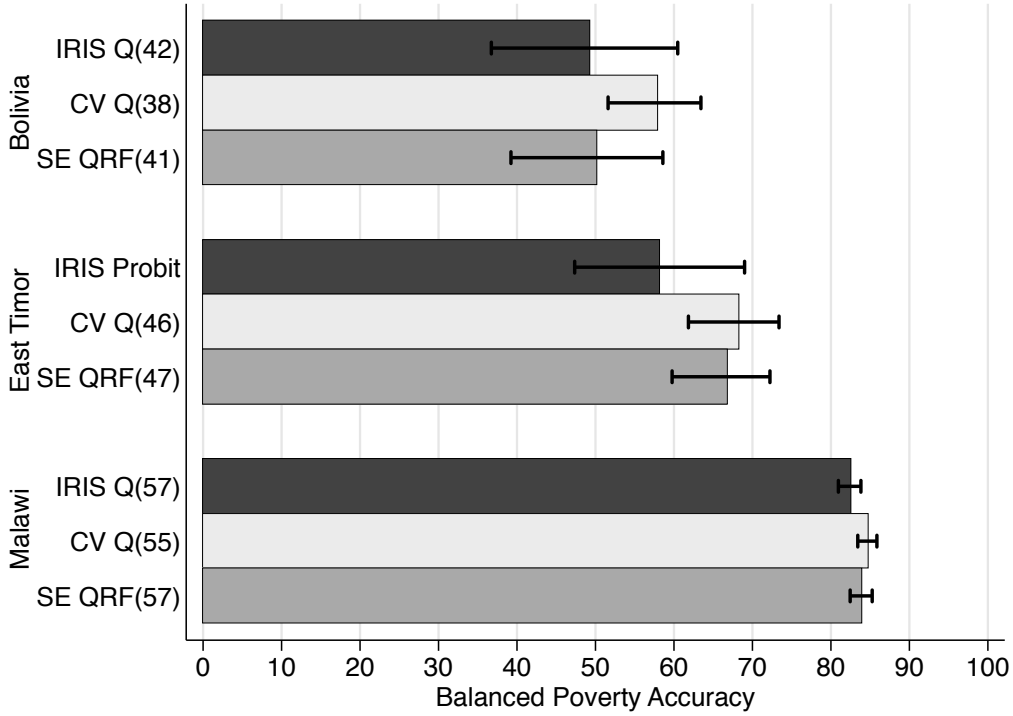


(a) Leakage (b) Undercoverage

The critical question, then, is how these trade-offs net out in terms of USAIDs key accuracy metric, the BPAC. Figure 2.3 demonstrates that the accuracy of the cross-validation approach outperforms that of the IRIS-generated tool in each country. Improvements range from 2.7 percent in Malawi to 17.5 percent in Bolivia. The performance of the stochastic ensemble approach closely follows that of the cross-validation approach in both East Timor and Malawi; although the

cross-validation results are statistically significantly different from the stochastic ensemble results, the magnitude of those differences is trivial in the case of Malawi and quite small in the case of East Timor (Table 2.4).

Figure 2.3: BPAC



In addition to gains in average BPAC, we also see large gains in the lower bound (2.5th percentile) performance using cross-validation and stochastic ensemble methods. The cross-validation (stochastic ensemble) approach improves the lower bound BPAC accuracy in Boliva by 38 (7) percent, in East Timor by 11 (8) percent, and in Malawi by 3 (2) percent.

Although the gains in poverty accuracy and BPAC in Malawi using the cross-validation approach are not as impressive as those in Bolivia and East Timor, note that the tool is able to outperform the already relatively accurate IRIS tool

for Malawi in terms of these metrics while also reducing both the leakage and undercoverage rates.

The relatively strong performance of the cross-validation approach compared with the stochastic ensemble approach is due to the fact that the cross-validation approach benefits from IRISs time and effort in selecting from a large set of possible variables a subset that explains much of the variation in the dependent variable. Because we have limited our analysis to the same subset of variables as selected by IRIS for their preferred models, the relative strengths of the stochastic ensemble methods in terms of variable selection are not well displayed through this analysis. Therefore, it remains an open question (that we plan to address in a later paper) as to whether our stochastic ensemble approach would outperform the combination of IRISs parametric model with cross-validation had we begun with the full set of 70125 variables instead of the selected subset. Our analysis does suggest, however, that the proxy means test tool developer who prefers to skip the time-consuming and computationally intensive process of stepwise regression followed by the comparison of multiple model specifications would do at least nearly as well in terms of out-of-sample performance as the tool developer who does take the time to perform these analyses and then combine them with cross-validation.

Finally, the robustness results for the assessment of tool performance under new poverty lines are reported in Appendix Table A.2. From a comparison of rows 2, 6, and 9 for each country, we can see that the cross-validation and stochastic ensemble approaches perform about the same as the IRIS approach under the new poverty lines. Overall, however, across all results, including the robustness results, we find that the cross-validation and stochastic ensemble

approaches do no worse than, and in many cases substantially outperform, the traditional approach to PMT tool development.

2.5 Conclusion

We have proposed methods for the improvement of a particular type of poverty-targeting tool: proxy means test targeting. In the country-level case studies analyzed here, prioritization of the out-of-sample performance of these targeting tools during tool development either through selecting a model based on its cross-validation performance or using a method such as stochastic ensemble methods that both selects variables and performs cross-validation along the way can significantly improve the out-of-sample performance of these tools. In particular, we find that application of cross-validation and stochastic ensemble methods to the problem of developing a poverty-targeting tool produces a gain in poverty accuracy, a reduction in undercoverage rates, and an overall improvement in BPAC in comparison to traditional methods.

Our analysis takes as given the IRIS-selected PAT variables so as to demonstrate the power of machine learning methods in this setting; however, beginning with a larger set of variables over which the stochastic ensemble methods may build a targeting model may produce even greater gains in targeting accuracy for this approach than observed here. Therefore, the gains in accuracy we have reported are likely conservative. Moreover, applying a stochastic ensemble approach over a larger set of variables would obviate the time-consuming tasks of both stepwise regression for variable selection and the process of running and comparing the performance of multiple statistical models, as was done by the

IRIS center. Overall, our findings suggest that further exploration of machine learning methods for PMT tool development is merited.

CHAPTER 3

HETEROGENEOUS WELFARE DYNAMICS AND STRUCTURAL TRANSFORMATION IN TANZANIA

3.1 Introduction

Welfare dynamicsi.e. the evolution of welfare over time in terms of income, expenditures, assets, or another measure that captures the economic well-being of an individual or householdcan offer insight into the inequality between, and growth prospects of, the poorest sectors of an economy. In addition, multiple equilibria welfare dynamics arising from non-convex technologies and multiple financial market failures are posited as one explanation for the puzzling and persistent gap in productivity between agricultural and non-agricultural sectors as well as the consumption gap between rural and urban households in economies across the globe.

Whereas much of the literature on welfare dynamics relies on two-production-technology theoretical models and simulations and whereas much of the literature on the agricultural productivity gap categorizes households into agricultural and non-agricultural households based on a binary decision rule, this paper begins with a flexible theoretical model and allows the number and type of livelihood strategies, unknown *a priori* to the researcher, to be determined by the data. Such an approach allows for the possibility that household income generating activities may be incremental and therefore fail to fit neatly into agricultural/non-agricultural-sector dichotomies. For example, McCullough (2016) finds that households in Tanzania 2010/2011 diversify into non-agricultural activities without leaving agriculture. In addition, there

may be multiple, sufficiently different, income-generating activities within each of the agricultural and/or non-agricultural sectors to constitute an additional livelihood, the returns to which merit further investigation. Despite the possibility of heterogeneous welfare dynamics within and between sectors, popular approaches to the estimation of welfare dynamics and to the identification of the causes of the productivity gap are limited to population means and ad-hoc dichotomies.

In this paper, I examine welfare dynamics in a setting where the livelihood strategy choice set is complex and evolves over time, and where returns to assets are potentially conditioned by livelihood strategies and by geography via migration. By livelihood strategies, I mean the Barrett et al. (2000) definition of livelihoods as the opportunity set afforded an individual or household by their asset endowment and their chosen allocation of those assets to generate a stream of benefits (p.2). This definition of livelihoods focuses on mapping assets and their allocations to welfare and will serve as the basis for the theoretical model developed below. My approach is to empirically identify livelihood strategies using *k*-medoids cluster analysis, allowing the number of clusters to be determined by the gap statistic method (Tibshirani et al. 2001). I then assess marginal returns to assets by livelihood and by livelihood and migration status. Locally increasing returns by livelihood or migration status would offer additional, micro-level, insights to the empirical findings on the productivity and consumption gaps between sectors and rural/urban environments observed by Gollin et al (2014) and Young (2013). I also examine the welfare dynamics for each of the identified livelihood groups. This approach also allows me to observe, in an entirely data driven way, any structural shifts taking place in the economy through differentiated returns to the livelihoods that emerge. The

analysis uses three waves of the Kagera (Tanzania) Health and Development Survey (KHDS), 1991 to 2010.

I find that, between 1991 and 2004, a subset of households moves from the single, farm-based, livelihood of Kagera, Tanzania to a livelihood that allocates more assets to off-farm wage and entrepreneurial activities. In other words, the cluster analysis splits households between agricultural and non-agricultural livelihoods, into the classic dual economy assumed elsewhere (Timmer 1988, Gollin et al 2014). I find evidence for differences in returns to business, labor, and human capital assets by livelihood strategy and by migration, supporting the findings on the production and consumption gap of Gollin et al (2014) and Young (2013), but without imposing a binary division in the data. In addition, I find evidence for heterogeneous welfare dynamics, such as would be masked in an analysis of population mean welfare dynamics alone; however, the equilibria appear to converge over time, suggesting a catch-up in returns in the agricultural sector. These findings offer another observation in the debate as to whether livelihood shifts or geography (migration) drives the increase in returns. I find that, in this setting, livelihood shifts play a greater role in increasing returns than does migration. Finally, there is no evidence of a multiple equilibria poverty trap in this setting.

3.2 Background and literature review

This paper draws on and speaks to a number of literatures, including the literature on welfare dynamics and the theory of poverty traps, the literature on structural transformation and the agricultural productivity gap, and the ongo-

ing debate over the role of geography in both welfare and productivity.

The theory of poverty traps suggests that we should see multiple equilibria welfare dynamics emerge in the presence of multiple market failures and non-convex production technologies (Galor & Ziera 1993, Barrett 2005, Barrett et al 2016). Generally, studies of welfare dynamics that are focused on non-convexities coupled with multiple financial market failures either run simulations with two-technology models or study empirical data on simple, two-technology economies such as livestock based economies in rural Kenya, Ethiopia, and Zimbabwe (Lybbert et al. 2004, Barrett et al. 2006, Santos & Barrett 2016, Hoddinott 2006). In such settings, two technologies are available to households: 1) a sufficiently large herd size to sustain transhumance, and 2) a small herd size that constrains households to sedentary living and a poorer, cultivation-based livelihood. The combined outer envelope of these productive technologies is non-convex, suggesting that households would experience increasing returns to their livestock holdings if they could switch from the low-return technology to the high return technology. In the face of market failures, such as thin credit and insurance markets, this non-convexity means that initial conditions determine long run outcomes and that shocks may have devastating permanent consequences (Barrett & Carter 2013).

While multiple equilibria poverty traps have been empirically observed in such rural nomadic economies, observation outside of such settings is rare. As Kraay & McKenzie (2014) argue in their review of the evidence on poverty traps, multiple equilibria welfare dynamics should not emerge where multiple production technologies are available and where it is relatively easy to move from one technology to another. Even in the face of market failures, if there exist suf-

ficiently many technologies, the outer envelope of the productive technology set may be convex. Such a scenario might exist in settings where livelihoods include various combinations of cultivation, wage labor, and small household enterprises such that the shift from one technology to another is incremental, e.g., raising additional livestock or investing in seeds for an additional agricultural commodity. Kraay & McKenzie (2014) support this point by showing that the distribution of start-up costs across a range of microenterprises in Sri Lanka is not only relatively continuous but also heavily right-skewed. With a few exceptions (Adato et al 2006, Carter et al. 2007, Naschold 2012, Kwak & Smith 2013), estimation of welfare dynamics in complex economies fails to find multiple equilibria welfare dynamics.

In an economy where multiple livelihood strategies are available, population mean welfare dynamics may disguise underlying heterogeneity (Adato et al 2006); it is not enough to consider mean dynamics. In analyses of welfare dynamics in economies with complex asset environments, various parametric (Adato et al 2006) and non-parametric (Naschold 2012) methods are used to generate an asset index. Because assets are collapsed into a single index using these approaches, heterogeneity in welfare dynamics based on particular initial assets, or combinations of assets, is generally not observed. Moreover, the welfare dynamics that are observed in these settings are sensitive to the method used to construct the asset index (Michelson et al. 2013). In this paper, I allow welfare dynamics to differ by livelihood groups, as defined over productive asset holdings and their allocations, thereby avoiding this collapse and allowing for empirically meaningful heterogeneity in my estimated welfare dynamics.

Very few papers consider heterogeneity in welfare dynamics; in part this is

because there are many ways to slice a data set or an analysis into heterogeneous groups, but few are theoretically or empirically meaningful. Rather, approaches entail either examining population mean dynamics (e.g., Adato et al. 2006), theoretically specifying differences in advance, such as high or low technology and high or low ability, and then observing dynamics in the two dimensional space they create–this is the approach taken by Ikegami et al. (2016) and Santos & Barrett (2016)–or examining heterogeneity in observable individual, household, or geographical characteristics (e.g., Naschold 2012, Giesbert & Schindler 2012, Kwak & Smith 2013).

Assessment of heterogeneity in welfare dynamics by looking at differences along observable characteristics has the drawback that it may simply impose the researchers' assumptions on the data without yielding empirical insights. For example, heterogeneity in dynamic welfare equilibria is examined by Naschold (2012) in terms of differences in caste, education, and landholdings in India and by Giesbert & Schindler (2012) in terms of differences in immigration status and education in Mozambique. However, the equilibrium values for each of the researcher identified subgroups have overlapping confidence intervals. Alternatively, Kwak & Smith (2013) examine both geographic and income heterogeneity in welfare dynamics in Ethiopia, finding that welfare dynamics differ if one is in the 25th versus the 75th quantile of the income distribution and that the Enset growing region of Ethiopia faces stagnation as compared with others. The few approaches that consider heterogeneity in welfare dynamics emerging from initial heterogeneous conditions in asset holdings do so through simulation. Both Dercon (1998) and Zimmerman & Carter (2003) find heterogeneous portfolio strategies emerging from heterogeneity in initial wealth/asset holdings based on dynamic stochastic models of asset accumulation that account for risk and

market failures.

The contribution of this paper is to approach the data with a fully general model that accommodates a number of market failures and places no assumptions or restrictions on the number of technologies/livelihoods available in the economy. Under this approach, we can be confident that any differences in livelihoods and/or equilibria that emerge in the economy under study are due to the data and not to theoretical or ad-hoc assumptions.

While this paper did not set out to identify differences in marginal returns to assets between agricultural and non-agricultural households, the data driven strategy produced just such a divide. Consequently, the findings speak to a large literature that attempts to explain production and consumption gaps between agricultural and non-agriculutral households within (and across) countries. The agricultural productivity gap – the empirical observation that returns to labor are greater outside of agriculture than within agriculture (Gollin et al 2014) and the associated finding that cost of living adjusted consumption is greater in urban than in rural areas (Young 2013) – presents a compelling problem as it suggests great opportunity for arbitrage as well as a means through which to address inequality and spur growth, by correcting the missallocation of factors across sectors and locations (Young 2013, Lakagos & Waugh 2013, Gollin et al 2014, McMillan & Rodrik 2011).

Gollin et al (2014) attempt to identify the source of the agricultural productivity gap by considering all the usual suspects, e.g., systematic measurement errors, differences in working hours, and differences in human capital across sectors. Depsite adjusting for all these factors, Gollin et al (2014) find that the productivity gap remains large. They conclude that their findings are consistent

with a story of self-selection wherein those with sufficient skill switch sectors. A strong case for the selection view has been made by others as well: Herrendorf & Schoellman (2018) find that "agricultural workers have lower innate ability" than do those in non-agricultural sectors, Lakagos & Waugh (2013) find that those in agriculture have both a comparative and absolute advantage in that sector (Lakagos & Waugh 2013), and Young's (2013) findings are consistent with the efficient allocation of human capital as those with unobserved skill (correlated with observed education) relocate to the urban environment.

Young (2013), Lakagos & Waugh (2013), Gollin et al (2014), and Herrendorf & Schoellman (2018) find that the productivity and consumption gaps are due to efficient sorting (selection) of labor on innate ability, comparative advantage, or unobserved skill rather than the consquences of barriers to mobility, market failures, or poverty traps. In fact, Herrendorf & Schoellman (2018) find that the barriers to the movement of labor from one livelihood to another are very small and Lakagos & Waugh 's (2013) model suggests that wage differences between the agricultural and non-agricultural sectors can exist even in the absence of barriers to labor mobility.

If innate ability, comparative advantage, or unobserved skill are randonly distributed, then the absence of barriers arguments would be more compelling. However, if there are path dependencies to the distribution of ability (which, e.g., Lagakos & Waugh (2013) proxy for with schooling attainment and which Young (2013) finds is highly correlated with the unobserved skill on which labor geographically sorts) – and there is overwhelming evidence that human capital development is linked to parental resources – then the possibility that multiple equilibria welfare dynamics are playing a role cannot be dismissed.

For example, in an analysis of livelihoods and welfare dynamics in Western Tanzania in the 1990s, Dercon (1998) observed occupational choices that could not be explained by comparative advantage; rather, due to market failure, risk, and the lumpy investments required to access a greater asset accumulation dynamic pathway, those with low initial endowments remain in low-risk, low-return livelihood activities, meaning that initial poverty is self-perpetuating.

With some exceptions, most of the data used for analysis of the productivity gap rely on macro-level data. Gollin et al (2014) and Young (2013) present the first approaches using micro-data, with Gollin (2014) relying on LSMS data and Young (2013) on DHS data. McCullough (2017) also presents a departure from the aggregate data, macro approach and arrives at novel conclusions regarding the source of the agricultural productivity gap. Using household level LSMS-ISA data from Tanzania 2010/11, McCullough (2017) finds that the productivity gap is smaller than reported using macro-data and that at least half of the gap is due to fewer labor hours supplied in the agricultural sector (rather than lower productivity per hour-worker in the agricultural sector).

Therefore, to this literature, I make a few contributions. As mentioned above in relation to heterogeneous welfare dynamics, in contrast to Gollin et al (2014) and others, I allow the data to sort itself into meaningful livelihood groups based on household asset holdings and their allocations, which additionally allows for the possibility that there may be fewer or more meaningful sectors in the economy than the agricultral and non-agricultral sectors (though this doesn't turn out to be the case). In contrast to Lakagos & Waugh (2013) and Herrendorf & Schoellman (2018), I rely on household level survey data. In contrast to Young (2013) who considers only migration, I consider both livelihood

and migration, allowing me to assess the relative contributions of each to the higher consumption outcomes that are observed. Finally, I estimate welfare dynamics within and between livelihoods.

In addition to Young's (2013) finding of geographic sorting based on skill, a number of other scholars have estimated the returns to migration and found that migration is a promising route out of poverty (Clemens, Montenegro & Pritchett 2008, Narayan & Petesch 2007, Bryan, Chowdhury, & Mobarak 2014, Christiaensen et al 2013). However migration is also tied up with livelihood shifts, self-selection, welfare dynamics, and poverty traps. There may be a prerequisite cash, human/social capital, or other asset, threshold to migration that cannot be overcome due to local market failures, giving rise to bifurcating welfare dynamics characterized by high returns for those who migrate and low returns for those who remain behind. In other words, one might consider migration just another, possibly non-convex, technology.

Ravallion & Woden (1999) ask, are there poor areas, or only poor people? Using data from Bangladesh in the early 1990s, they find that poor areas are not a consequence of the concentration of households with observable attributes that foster poverty; rather they find geographically determined differences in returns. Likewise, Jalan & Ravallion (2002) and Kwak & Smith (2013) find that geographic characteristics can affect the productivity of households productive capital. And in their review, Kraay & McKenzie (2014) find that the evidence most consistent with the theory of poverty traps is that of geographic poverty traps.

However, examination of geographically determined welfare poses several challenges; in particular, selection into migration or geographic location is en-

dogenous. Using an RCT set-up, Bryan et al (2014) overcome this challenge. They find positive and significant local average treatment effects for seasonal internal migration in Bangladesh, raising the question of why more households don't migrate. Their findings are consistent with poverty trap-like dynamics in which very poor households must overcome a cash-on-hand threshold. They also find that migration is an "experience good", meaning that any learning must be individual; in this sense there is also an experience threshold to overcome.

In addition, Christiaensen et al (2013) find that the majority of households from Kagera, Tanzania exiting poverty in the period 1991-2010 do so by migrating to secondary towns (towns with a population of 500,000 or less) and not to urban centers. However, those who did migrate to urban centers experienced faster consumption growth. In other words, there is heterogeneity in returns in terms of migration destination. I include migration to any destination (not just urban areas) in my estimation as a technology that can interact with livelihoods. I also consider both livelihood and location shifts and the extent to which each contributes to observed gains in consumption. However, in contrast to Bryan et al (2014), I have no plausible identification.

Finally, the region of study in this paper, Kagera, Tanzania, has been extensively studied due to the unique longitudinal panel data set collected there. Key analyses include De Weerdt (2010), Beegle et al (2011), and Christiaensen et al (2013). Overall these analyses capture important transitions between 1991 and 2010 in Kagera in particular and Tanzania in general as households grow, split, diversify, and migrate; each analysis finds significant welfare returns to livelihood diversification, migration, and living in less remote areas (either initially,

or though migration).

Using both quantitative and qualitative analyses, De Weerdt (2010) identifies two pathways out of poverty in Kagera, Tanzania between 1991 and 2004: agriculture and business/trade. Initial conditions in 1991/94in particular, initial stocks of land and human capital as well as location factors such as the degree of connectedness of the place of residencedetermine outcomes in 2004. Overall, he finds that those individuals who diversified their livelihood activities (crops, non-farm earnings) had better outcomes in 2004 than those who remained in traditional farming.

Beegle et al. (2011) focus on the role of migration in improved welfare for individuals from Kagera. Like De Weerdt (2010), Beegle et al. (2011) find that there are greater returns to diversification than to traditional farming but that those who have migrated have greater gains in consumption no matter their livelihood activity. While De Weerdt (2010) identifies the value of connectedness in initial location, Beegle et al. (2011) find that the connectedness of the location to which an individual migrates is also important, as it has a significant positive effect on consumption regardless of livelihood activity.

Christiaensen et al. (2013) take a closer look at the diversification and migration patterns suggested by De Weerdt (2010) and Beegle et al. (2011). Christiaensen et al. (2013) examine the transitions among farming and non-farming activities in small towns (rural areas and secondary cities), and industry and service labor in cities between 1991 and 2010, finding that the majority of those who escaped poverty did so not by moving to cities but by either diversifying into non-farm activities or migrating to small towns, or both. These findings suggest that it is not necessary to migrate to the city to realize returns to diver-

sification, migration, and connectedness.

3.3 Theoretical model

To incorporate the flexible understanding of a livelihood as a function that maps assets and their allocations to a stream of benefits (Barrett et al. 2000) into a model that allows for a variety of household specific market failures (deJanvry, Fafchamps, & Sadoulet 1991, deJanvry & Sadoulet 2005), my approach is to combine the Barrett (2008) model of household market participation decisions with a dynamic model of asset accumulation building on Ikegami et al (2016), Carter & Ikegami (2009), and Buera (2009). I extend these models to include K livelihood strategies, each of which can contain any combination of productive technologies.

Assume that household h at time t has asset stock A^d_{ht}, where each asset is indexed by $d = 1, ...D$; these assets might include labor, land, livestock, other physical capital (such as business and farm assets), and human capital (such as education and health). A set of livelihood strategies, $L_k, k = 1, K$ are available to the household (Expression 1). Each livelihood strategy is a function of a set of production technologies, $f^j(A^d_{ht})$, where $j = 1, J$ indexes the commodity output.

$$\left\{L_1(f^j(A^d_{ht})), L_2(f^j(A^d_{ht})), ...L_K(f^j(A^d_{ht}))\right\} \tag{3.1}$$

To illustrate how technologies might combine to form livelihood strategies

that produce commodities, consider two example livelihood strategies, L_1 and L_2. L_1 might include both maize farming and wage labor,

$$L_1 = (f^{maize}(A_{ht}^{land}, A_{ht}^{labor}), f^{wagelabor}(A_{ht}^{labor}, A_{ht}^{education})) \tag{3.2}$$

while L_2 might include maize farming and milk production

$$L_2 = (f^{maize}(A_{ht}^{land}, A_{ht}^{labor}), f^{milk}(A_{ht}^{livestock}, A_{ht}^{equipment})). \tag{3.3}$$

There exist fixed costs, FC_{L_k}, and transactions costs to employing a given combination of technologies (i.e., a given livelihood strategy). While the fixed costs faced by a household depend only on the livelihood strategy employed by the household, transactions costs faced by a household, $TC_{ht}^{j}(\mathbf{Z}_{ht}, \mathbf{A}_{ht}, \mathbf{E}_t, \mathbf{f}^j)$, are a function of household characteristics, $\mathbf{Z}_{ht}$, household asset stocks, $\mathbf{A}_{ht}$, characteristics of the local environment, $\mathbf{E}_t$, and the vector of productive technologies employed, $\mathbf{f}^j$. Along the outer envelope of optimal livelihood strategies, greater fixed costs are associated with higher return livelihoods such that $FC_{L_k} < FC_{L_{k+1}} < FC_{L_K}$, as any option with high fixed costs but low returns would be strictly dominated.[1]

The household can either be a net seller, M^{js}, or a net buyer, M^{jb}, of a given commodity, where $M^{js}, M^{jb} \in \{0, 1\}$; a household can also be autarkic with re-

[1] With this simplifying assumption, I am assuming households select their optimal livelihood strategies conditional on associated fixed costs.

spect to a commodity, in which case $M^{js} = M^{jb} = 0$. A household cannot be both a net seller and net buyer; i.e., there is no case where $M^{js} = M^{jb} = 1$.

The household faces market price, p_t^j, for each commodity it buys and sells; however, the household specific price, $p_{ht}^{j^*}$, is modulated by transactions costs as well as the households status relative to the market,

$$\begin{aligned} p_{ht}^{j^*} &= p_t^{jm} + (M_{ht}^{jb} - M_{ht}^{js}) TC_{ht}^j(\mathbf{Z}_{ht}, \mathbf{A}_{ht}, \mathbf{E}_t, \mathbf{f}^j), \ \textit{where } M_{ht}^{jb} \neq M_{ht}^{js} \\ p_{ht}^{j^*} &= p_{ht}^{ja}, \ \textit{where } M_{ht}^{jb} = M_{ht}^{js} = 0 \end{aligned} \tag{3.4}$$

Barrett (2008) points out that market participation decisions are analytically similar to technology choice decisions; a market exchange that tranforms physical goods and services into money metric net revenue has the same properties it is a quasi-concave and monotone mapping from the goods and services sold into net revenues as a production technology, allowing one to nest market participation decisions within the choice of production technologies. One can think of the transactions costs to technology adoption as the costs generating shadow prices that influence market participation decisions; therefore, just as multiple technologies can be employed in a single livelihood, so can we include participation (or non-participation) in multiple markets, such as selling maize in the market and producing milk for home consumption only, as show in L_3. Similarly, we can think of the decision to migrate or not as either a labor market participation decision or a technology adoption decision.

$$L_3 = (f^{maize_{market}}(A_{ht}^{land}, A_{ht}^{labor}), f^{milk_{autarkic}}(A_{ht}^{livestock}, A_{ht}^{equipment})) \tag{3.5}$$

The household earns income, y_{ht}, from the production and sale of commodities, having selected[2] its optimal livelihood strategy,

$$y_{ht} = (\Sigma_{j=1}^{J_{L_k}} p_{ht}^{j^*} f^j(\mathbf{A}_{ht}) | max\left\{L_1(f^j(A_{ht}^d)), L_2(f^j(A_{ht}^d)), ...L_K(f^j(A_{ht}^d)) | FC_{L_k}\right\}) \tag{3.6}$$

For all commodities not traded in the market due to market participation decisions emerging from the transactions costs faced by the household, i.e., for all $j \notin M$, consumption is constrained by household production (assuming away carryover stocks from one period to the next),

$$c_{ht}^j \leq f^j(\mathbf{A}_{ht}) \tag{3.7}$$

The household maximizes utility over consumption of a vector of agricultural, small enterprise, or wage-labor produced commodities, $\mathbf{c}^j$, as well as other tradables, $\mathbf{x}$, that the household cannot produce. The household is subject to budget constraints. Let $\mathbf{p}^x$ represent the price of commodities the household cannot produce. Then the household budget is,

[2]This model abstracts from whether the livelihood strategy selection is due to inherent ability or risk preferences.

$$\mathbf{p}_t^{x\prime}\mathbf{x}_{ht} + \Sigma_{j=1}^{J_{L_k}} p_{ht}^{j^*} c_{ht}^{j} + I^d(y_{ht}) \leq y_{ht} \tag{3.8}$$

where $I^d(y_{ht})$ is a function that maps income into assets via investment. The asset accumulation law of motion is

$$A_{ht+1}^d \leq \delta_t A_{ht}^d + I^d(y_{ht}) \tag{3.9}$$

where $\delta_t > 0$ can be either greater than one (to capture interest, the fact that livestock beget more livestock, etc) or between zero and one (to capture depreciation).

Finally, let $\mathbf{A}_{ht} \geq -B(\mathbf{A}_{ht})$ where B is the net borrowing constraint as a function of household assets, meaning that financial market failures may be household specific. Households with adequate asset holdings might be deemed creditworthy; that is, with a sizable positive entry in one element of the A vector (e.g., land holdings), the household will be able to borrow (i.e., have significant negative net holdings of) another asset (e.g., cash) as it is able to offer some assets as collateral.

Altogether, the households dynamic welfare maximization problem can be

represented as follows:

$$
\begin{gathered}
max_{\mathbf{c}_{ht},\mathbf{x}_{ht},I^d} \Sigma_{t=1}^{\infty} \beta^{t-1} u(\mathbf{c}_{ht}\mathbf{x}_{ht}) \\
sub\,ject\,to: \\
\mathbf{p}_t^{x\prime}\mathbf{x}_{ht} + \Sigma_{j=1}^{J_{L_k}} p_{ht}^{j^*} c_{ht}^{j} + I^d(y_{ht}) \leq y_{ht} \\
y_{ht} = (\Sigma_{j=1}^{J_{L_k}} p_{ht}^{j^*} f^j(\mathbf{A}_{ht}) | max \left\{ L_1(f^j(A_{ht}^d), L_2(f^j(A_{ht}^d)), ...L_K(f^j(A_{ht}^d)) | FC_{L_k} \right\}) \\
c_{ht}^{j} \leq f^j(\mathbf{A}_{ht}),\ j \notin M \\
A_{ht+1}^d \leq \delta_t A_{ht}^d + I^d(y_{ht}) \\
\mathbf{A}_{ht} \geq -B(\mathbf{A}_{ht})
\end{gathered}
\tag{3.10}
$$

This model allows for, but does not assume, multiple market failures such as borrowing constraints and non-separability of household production and consumption decisions. For example, where a household can borrow, it will optimally choose a livelihood with a marginal return equal to the marginal rate of substitution between consumption today and consumption in the next period; but if it cannot borrow ($\mathbf{A}_{ht} \geq 0$), the standard Euler equation becomes kinked (Deaton 1991) and the household dissaves. Where production and consumption decisions are non-separable, household shadow prices create a wedge between sales and purchase prices leading to poor or non-transmission of market prices and other inefficiencies (Barrett 2008). In addition, this model is not limited to two technologies or two livelihoods; in fact it imposes no constraints on the technology choice set. In relaxing the assumption of complete and competitive markets and in imposing no constraints on the technology choice set, this is a fully general model.

The key structural arguments of the livelihood conditioned returns to assets are estimable in reduced form. In particular, returns to assets will be estimated via Taylor series expansion of the reduced form expression relating welfare to assets in Equation 3.11,

$$e_{ht} = f^j(A^d_{ht}) + \epsilon_{ht} \tag{3.11}$$

where e_{ht} represents consumption expenditures, the best available representation of welfare in the KHDS data, measured as $e_{ht} = \Sigma_{j=1}^{J_{L_k}} c^j_{ht}(p^j_t + TC^j_{ht}(\mathbf{Z}_{ht}, \mathbf{A}_{ht}, \mathbf{E}_t, \mathbf{f}^j))$.

3.4 Data

The analysis uses three waves of the KHDS. The first wave began in 1991 (and continued through 1994), the second tracked and revisited households in 2004, the third in 2010. The survey instrument changes between 1991, 2004, and 2010 such that by 2004 it is no longer possible to observe land (acres) allocated to different crops and by 2010 it is no longer possible to observe labor (hours) allocated to different occupations. Therefore, the cluster analysis and returns to assets estimations are performed using only the 1991 and 2004 data sets. The 2010 data are included in the estimation of welfare dynamics.

The KHDS data are interesting for several reasons: they present a long panel with very low attrition rates92 percent of baseline households were tracked through to 2010and they cover a period when Tanzania is undergoing structural transformation (Christiaensen, De Weerdt, & Todo 2013). The initial 1991 sur-

vey was implemented for the purpose of studying the effects of the HIV/AIDS epidemic on welfare, and households were purposively sampled to that end. The sample is not intended to be representative of the general population of Tanzania or of Kagera. For further details about the region and the data, see De Weerdt (2010), Beegle et al. (2011), De Weerdt & Hirvonen (2016).

Table 3.1 suggests that financial market failures are a possible constraint in this setting: of those individuals starting businesses in 2010, [3] 50% relied on own savings for start-up capital, 15% sold assets or crops, and 18.6% relied on friends/relatives; only 5.6% used formal or informal institutions. Table 3.1 also suggests that there is not a great deal of diversification of sourcing for start-up capital, as 86% of businesses reported not drawing on a second source for funding.

Households from and within Kagera have enjoyed growth in consumption over the course of the longitudinal study. Cumulative densities of per capita consumption in 1991, 2004, and 2010, using data in 2010 TZS value that has been deflated using a regional price index, are presented in Figure 3.1. The horizontal line in the figure represents the national poverty line. From 1991 to 2004, most of the shift in consumption takes place above the poverty line, suggesting that those below the poverty line may be trapped in a low welfare equilibrium; however, between 2004 and 2010 we see movement along the full distribution and a much larger shift overall.

Overall, the sample is upwardly mobile with 59 percent of the 1991 poor transitioning out of poverty by 2004 and 60 percent of the 2004 poor transition-

[3]Data on sources of start-up capital were not collected in earlier waves of the KHDS. Given the attention on microfinance in the 2000s, it is reasonable to assume that credit availability to households in Kagera in earlier periods was no better than, and possibly worse than, that in 2010.

Table 3.1: Sources of start up capital for household-owned enterprises, KHDS 2010

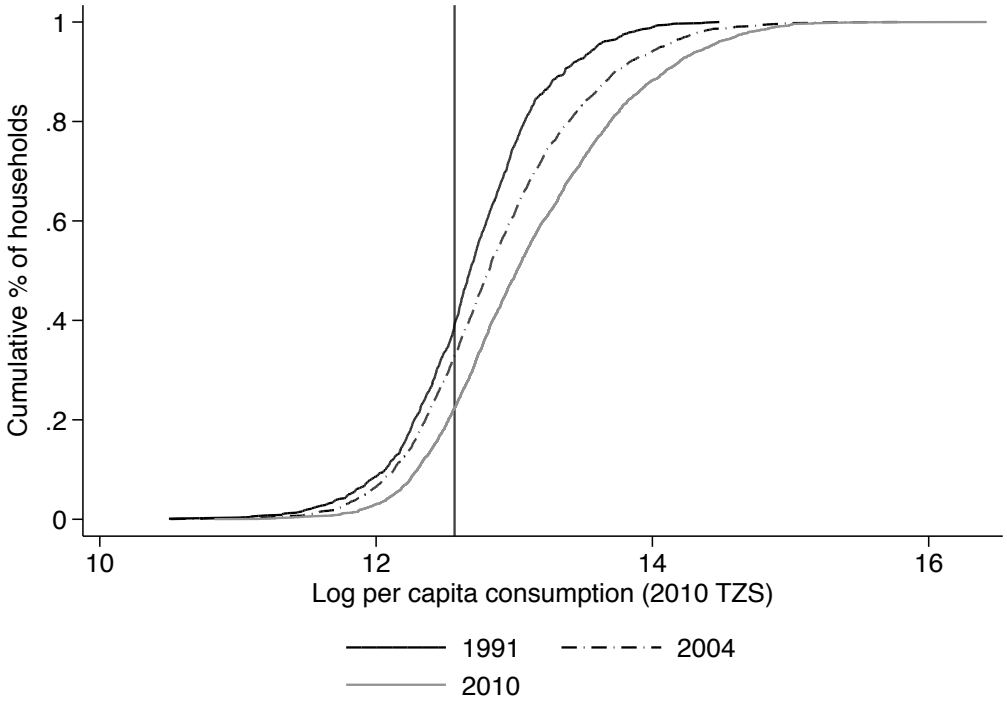
	First most important		Second most important	
	N	%	N	%
Savings	669	50.19	75	6.20
Bank Loan	12	0.90	9	0.74
Informal Insurance Group Loan	49	3.68	10	0.83
Loan From Relatives	42	3.15	13	1.08
Loan From Friends	64	4.80	17	1.41
Gift From Relatives	124	9.30	20	1.65
Gift From Friends	18	1.35	4	0.33
Business Partner	10	0.75	2	0.17
Microfinance Institution	14	1.05	15	1.24
Sold assets or crops	200	15.00	0	0.00
Other (specify)	7	0.53	3	0.25
No Start Up Capital Needed	124	9.30	1,041	86.10
Total	1,333	100	1,209	100

ing out of poverty by 2010 (Table 3.2); however, 30 percent of the 1991 poor remain poor in 2010.

Table 3.2: Poverty transition matrix (%)

		2004		2010	
		Poor	Nonpoor	Poor	Nonpoor
1991	Poor	40.68	59.32	30.16	69.84
1991	Nonpoor	25.46	74.54	17.24	82.76
2004	Poor			39.12	60.88
2004	Nonpoor			17.47	82.53

Figure 3.1: Cumulative consumption 1991, 2004, 2010



Note: The horizontal line is the national poverty line. Consumption data are in real 2010 TZS

3.5 Empirical approach

My empirical approach is as follows: 1) define a set of livelihood strategies based on household asset holdings and land and labor allocations using *k*-medoids cluster analysis, 2) estimate returns to assets conditional on livelihood choice using a second order approximation of a function relating consumption to assets via fixed effects estimation, and 3) estimate welfare dynamics within and between identified livelihood groups.

3.5.1 Identifying livelihoods

The task of identifying livelihood groups in a data driven manner poses several challenges. The first challenge is to use a method that avoids arbitrary imposition of empirically-unsupported assumptions on the number and content of groups. Otherwise, it may be easy to find a sufficient number of livelihoods to make the outer envelope of the livelihood set convex or an insufficient number to make it non-convex. For this task, I use *k*-medoids cluster analysis (Kaufman & Rousseeuw 1990) and rely on the gap statistic method (Tibshirani, Walther, & Hastie 2001) to identify the optimal *k* in the data. The method of *k*-medoids cluster analysis is more robust to outliers than *k*-means because within-cluster dissimilarity is calculated via Manhattan distance as opposed to sum of squares. *K*-medoids cluster analysis operates by identifying the *k* medoid observations, or representative objects, that, once the other observations in the data set are assigned to closest representatives, best minimize dissimilarities in the resulting clusters through an iterative algorithm (Kaufman & Rousseeuw 1990).

While many methods for the selection of k are available in the literature, most are undefined for k=1; whereas the gap statistic method allows the data to identify a single cluster. Therefore, I rely on the gap statistic as an unbiased method for the identification of the appropriate number of livelihood clusters. The gap statistic identifies the optimal k as that for which the log of the within-cluster dissimilarity measure is furthest (i.e., has the greatest gap) from the expected log of the within-cluster dissimilarity measure for a null reference distribution (Tibshirani et al. 2001). The gap statistic was developed by Tibshirani et al. (2001) as an objective alternative to the commonly used elbow method heuristic.

The cluster analysis procedure involves first normalizing each dataset, using the gap statistic method to identify the optimal number of clusters for each dataset, and then assigning households to their clusters. The gap statistic procedure (Tibshirani et al. 2001) entails iterating through the generation of k=1,...K-medoids clusters (I select K=15), and calculating the within-cluster dissimilarity measure for each k, W_k. The same procedure is applied to B bootstrap samples of the data (drawn uniformly from the support for each variable used in the cluster analysis so as to create a null reference distribution), producing W_{kb}^r. The gap statistic for each k is then the distance between the true within-cluster dissimilarity and the average within-cluster dissimilarity for the bootstrapped samples,

$$gap(k) = \frac{1}{B}\Sigma_b log(W_{kb}^r) - log(W_k) \tag{3.12}$$

The optimal k is selected where the gap of k is greater than that of $k+1$ minus the standard deviation, s_k, of $k+1$,

$$gap(k) \geq gap(k+1) - s_{k+1} \tag{3.13}$$

where the standard deviation for each k is calculated as the product of the standard deviation of the bootstrap and the simulation error,

$$s_k = \left[\frac{1}{B}\Sigma_b(log(W_{kb}^r) - \frac{1}{B}\Sigma_b log(W_{kb}^r))^2\right]^{\frac{1}{2}} \sqrt{1+\frac{1}{B}} \tag{3.14}$$

The first term of Equation 3.14 is the standard deviation of the B bootstrapped W_{kb}^r; the second term accounts for the simulation error. In implementing this approach, I follow the Tibshirani et al. (2001) option of using principle components rotation for the generation of the uniform distribution of the null reference set, as this proved robust to both k=1 and elongated clusters in Tibshirani et al. (2001). I select the number of bootstraps as B=500.

With the appropriate k, denoted k^*, determined by the gap statistic, the k-medoids clustering algorithm, partitioning around medoids (PAM), proceeds as follows: it first selects in stepwise fashion an initial set of medoids, up to k^*, that minimize dissimilarity in the resulting clusters; it then iteratively replaces these medoids with observations one by one, stopping when the dissimilarity measure cannot be further minimized. Formally, the program minimizes the

objective function in Equation 4.21, by iteratively choosing cluster medoids i_k (Hastie et al. 2009),

$$min_{C,\{i_k\}_1^{k^*}} \Sigma_1^{k^*} \Sigma_{C(i)=k} d_{ii_k} \tag{3.15}$$

where d_{ii_k} is the distance between the cluster medoid and the other members of cluster $C(i)$.

I implement the analysis in R using the cluster package by Maechler et al. (2017) and select tuning parameters as specified above. The resulting clusters are described below. Note that although this approach to identifying livelihoods is data driven and not mechanically correlated with the measure of welfare in this analysis (household consumption), it does not guarantee that cluster assignment is orthogonal to welfare.

The second challenge in identifying livelihood groupings in a data driven manner is deciding on the appropriate set of variables to include in the analysis. The number of clusters may be affected by the level of (dis)aggregegation in the data; for example, should variables such as number of pigs and number of cows be aggregated to tropical livestock units [4] (TLUs) or left as individual variables? The literature [5] offers little guidance in these decisions.

[4]Tropical livestock units allow researchers to aggregate various livestock into a single, internationally comparable, cattle equivalency.

[5]There is one paper: using cluster analysis to identify livelihood strategies among rural Kenyans, Brown et al. (2006) aggregate the available data into eleven different activities including the production of annual food crops, perennial cash crops, coffee, tea, perennial forage crops, improved and unimproved dairy cattle, non-dairy cattle, small ruminants and pigs, and skilled and unskilled wage employment, reflecting a mix of productive assets and activities as well as outputs; from these eleven activities they identify five different livelihoods using k-means cluster analysis, having selected $k = 5$.

So as to produce a set of livelihood strategies based on land and labor allocation and asset holdings in line with the theoretical model described above, I perform the cluster analysis over variables indicating household land and labor allocation and productive assets only. In addition, so as to minimize researcher influence in the final number of clusters and their contents, I keep the data as granular as possible. This means, for example, that if the survey instrument asks about the number of pigs and the number of cows owned by the household, I use number of pigs and number of cows as separate variables in the analysis, as opposed to aggregating livestock into TLUs. However, the available data set comes with some limitations; for example, labor allocated to the production of different types of crops or livestock cannot be observed. Note that the keeping the data as granular as possible not only keeps the research enterprise honest, it also provides the clustering algorithm with more information over which to parse the data.

3.5.2 Estimating returns to assets by livelihood

To estimate returns to assets by livelihood, I estimate a second order Taylor series expansion of a function relating welfare (log per capita consumption expenditures), e, to the productive asset variables, A_d, available in the data, with an interaction term for livelihood strategy. I estimate individual, location, and time fixed effects using the 1991 and 2004 waves of the KHDS to address unobserved time invariant heterogeneity as well as annual trends that may be correlated with welfare and the employment of particular assets or choice of livelihood strategy. Identifying variation comes from changes in productive asset holdings and livelihood strategies at the household level. A vector of time-varying indi-

vidual and household characteristics, $\boldsymbol{X}_{ht}$, including age and squared age, marital status, farm inputs, and the number of businesses the household operates is included to control for time varying observables. Note that location fixed effects will not capture unobserved location-specific heterogeneity for those who moved out of the Kagera region by 2004, due to the fact that non-Kagera locations are not observed in the first wave.

The productive assets used in the estimation[6] include labor hours per week, land area, the log value of business assets, total TLU, and years of education. In Equation 4.25, i indexes individual, t indexes time, l indexes location, and d indexes the productive assets included in the estimation.

$$e_{itl} = \Sigma_d^5 \beta_d A_{itdl} + \frac{1}{2}\Sigma_d^5 \beta_{dd} A_{itdl}^2 + \Sigma_d^5 \Sigma_j^5 \beta_{dj} A_{itdl} A_{itjl} + \boldsymbol{\beta}_x \boldsymbol{X}_{it} + \Sigma_d^5 \gamma_d A_{itdl} L_{it} + \frac{1}{2}\Sigma_d^5 \gamma_{dd} A_{itdl}^2 L_{it} + \Sigma_d^5 \Sigma_j^5 \gamma_{djl} A_{itdl} A_{itjl} L_{it} + \boldsymbol{\gamma}_x \boldsymbol{X}_{it} L_{it} + w_l + \alpha_i + \psi_t + \epsilon_{itl} \tag{3.16}$$

With the resulting coefficient estimates, I trace out the marginal returns by livelihood strategy for each asset over its support,

$$m(A_r) = \hat{\beta}_r + \hat{\beta_{rr}} A_r + \Sigma_s^4 \hat{\beta_{rs}} \bar{A}_s + \hat{\gamma}_r + \hat{\gamma_{rr}} A_r L + \Sigma_s^4 \hat{\gamma_{rs}} \bar{A}_s L \tag{3.17}$$

where r indexes the support of the asset of interest and $\bar{A}$ indicates that a

[6] In contrast with the cluster analysis approach, I aggregate assets into meaningful categories here.

variable is being held at its mean. Standard errors are produced using the delta method.

If the data are consistent with a multiple equilibria welfare dynamics scenario, we should observe the marginal returns to assets differing by livelihood strategy such that the livelihoods requiring greater fixed costs offer higher returns to the same asset holdings, producing locally increasing returns in any shift from a low return livelihood to a higher return livelihood (i.e., generating local non-convexities in the outer envelope of the livelihood choice set).

3.5.3 Livelihood group welfare dynamics

I examine livelihood group welfare dynamics between 1991 and 2004 and between 2004 and 2010 to observe whether initial welfare status determines long run dynamics both within and across livelihoods. I use the flexible fractional polynomial estimator (Royston & Altman 1994, StataCorp 2009) to graph these dynamics, regressing logged consumption in the later period on that of the earlier period.

3.6 Results

3.6.1 Identifying livelihoods

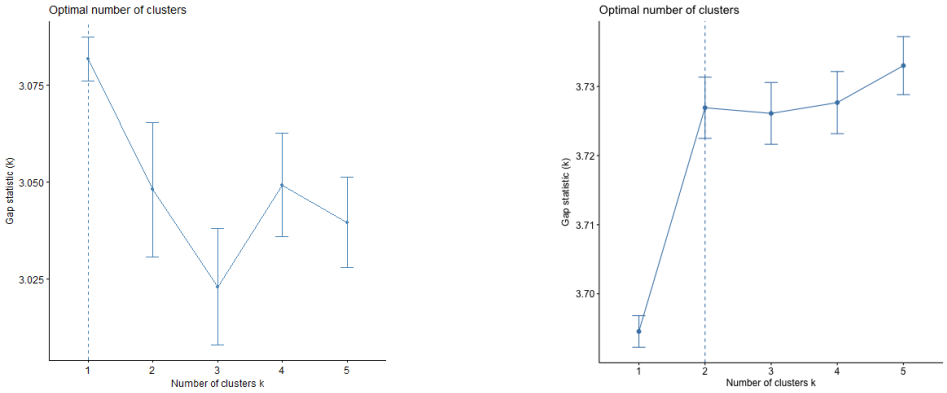
The cluster analysis is performed over a set of 99 (94) variables for the 1991 (2004) data set; these variables capture labor allocation across various wage,

small enterprise, and farm activities; land allocation across cash, staple, and sustenance crops as available in each data set; stocks of land, livestock, farm, financial (including unearned income), and business assets; and expenditures on hired labor and farm inputs. They also capture human capital assets in terms of education and health. Using the gap statistic method, a single livelihood was identified in the 1991 data and two livelihood clusters were identified in the 2004 data. From Figure 3.2, we can see that there is a single cluster (the full data set) in the 1991 data and that there are two well-defined clusters in the 2004 data, though greater than two, less well-defined, clusters or subclusters might also be identified.[7] Summary statistics for, and a plot of, the 2004 clusters are available in Appendix Table B.1 and Figure B.1. The cluster plot in Figure B.1, presenting the projection of the data on to its first two principle components, suggests that the clusters are well separated. Descriptions of each of the identified livelihood strategies follow.

The two livelihood clusters that emerge in the 2004 data might be best referred to by their most salient characteristics: the 2,216 households in cluster one have, on average, larger household sizes, larger land holdings, greater livestock assets, and allocate more land to every crop (excepting rice) than do those in cluster two (Appendix Table B.1). Therefore Ill refer to cluster one as the farm-based livelihood strategy. The 558 households in cluster two have higher education, allocate more labor to wage labor (excepting farm wage labor) and non-farm self-employment, and hold greater non-farm business assets; therefore, Ill refer to cluster two as the wage labor/entrepreneur livelihood strategy.

[7] As a robustness check on the stability of the clusters, I rely on the boostrapped Jaccard coefficient approach described in Hennig (2007). The Jaccard coefficient offers a measure of the similarity of cluster membership across bootstrapped clusterings of the data. The approach identified two clusters with Jaccard coefficients of 0.987 and 0.949 across 100 bootstrap samples of the data, indicating that the identified clusters are a highly stable structure in the data.

Figure 3.2: Optimal number of clusters 1991 ($N = 915, v = 99$) and 2004 ($N = 2774, v = 94$) using the gap statistic



(a) 1991 (b) 2004

Compared with households in the farm-based livelihood group, the wage labor/entrepreneur households allocate more hours per week to wage labor in skilled, professional, or services industries; they also allocate more labor to self-employment as merchants, in transportation, in services and other skilled industries (Appendix Table B.1). While they have many fewer livestock, land, and other farm assets than the farmers, the wage labor/entrepreneurs have much larger business asset holdings: the total value of their business buildings is 2.5 times greater than that of the farmers and the value of their business vehicle and equipment assets is approximately twice as large. However, there is no difference in the total number of businesses operated by household members between the two livelihood strategiesin both livelihoods, households own, on average, half of a business. Meanwhile, the farm-based households allocate more labor to farm and livestock activities. They hold on average 3.6 acres of farmland as compared with the 0.14 acres of the wage labor/entrepreneur group. They also own more sheep/goats, cattle, pigs, and other livestock.

In terms of unearned income and financial assets, the wage la-

bor/entrepreneurs have no pension, no dowry, and do not play the lottery, perhaps reflecting the fact that these households have younger [8] heads of household (32 years old as compared with 44 years old in the farmer group) and are less likely to be married (51 percent as compared with 79 percent in the farmer group). On the other hand, the wage labor/entrepreneurs receive much greater income from interest on savings (7.4 times greater), sale of durables (4.9 times greater), and receive larger remittances (1.9 times greater) than do the farmers.

The average household size in the wage labor/entrepreneur group is 3 compared with that of 5 for the farmer group. While they have fewer laborers per household, the wage labor/entrepreneur households have higher human capital in terms of education and health. Households in the wage labor/entrepreneur livelihood group have a higher share of household members who have completed secondary school (18 percent of the household compared with 4 percent in the farmer group), advanced (3 percent compared with 0 in the farmer group), and university (1 percent compared with 0) degrees. They also enjoy slightly higher health: on average, 53 percent of household members reported being free of illness or injury over the past 4 weeks as compared with 48 percent of household members in the farmer group.

Although not included as variables in the cluster analysis, consumption levels, poverty status, and moved or migrated statuses differ by livelihood. The wage labor households have 2.5 times higher consumption than the farm households and are much less likely to be poor (9 percent compared with 51 percent). A greater share of the wage labor/entrepreneur household has moved from the original homestead (50 percent compared with 21 percent) and the household

[8] Neither age nor marital status variables were used for the clustering; however, it is instructive to compare these demographic data across clusters.

is more likely to have migrated out of their original sampling cluster (84 percent compared with 43 percent). This suggests that the wage laborers and entrepreneurs are able to earn a higher return on their labor and or entrepreneurial activities because of migration, education, both, or an omitted variable correlated with both consumption and livelihood. Unobservable individual heterogeneity, such as inherent ability, will be addressed below via fixed effects estimation of the returns to assets.

Looking back to 1991 [9] asset holdings based on households 2004 identified livelihood strategies, differences between those households that eventually enter the wage labor/entrepreneur livelihood and those that do not are not great, as we might expect given that cluster analysis was not able to parse the 1991 data. However, we do see the following: the 139 households in 1994 that grow into the 558 wage labor/entrepreneur households by 2004 had slightly higher consumption (1.2 times greater), were slightly less poor (42 percent compared with 49 percent), had slightly higher shares of primary (65 percent compared with 61 percent) and secondary (5 percent as compared with 3 percent) educated households members, and slightly greater health (93 percent compared with 91 percent). Although statistically significant, these differences are all very small in magnitude. We also see slight differences in the number of businesses owned (greater in wage labor/entrepreneur group), the amount of time allocated to farm and fish wage labor (smaller in wage labor/entrepreneur group) and factory wage labor (greater in wage labor/entrepreneur group), and allocation of land area to certain crops.

The only large-in-magnitude differences are the value of business build-

[9]The 2004 data are weighted by their 1991 quantities so as to not spuriously find significant differences. Table not presented but available on request.

ing assets (3.2 times greater in wage labor/entrepreneur group), land holdings (0.55 acres smaller in wage labor/entrepreneur group), and financial assetsthe wage labor/entrepreneurs have three times as great value from ROSCA participation and two times greater value of other non-labor income. The wage labor/entrepreneurs also have 1.6 fewer per capita on farm labor hours and 0.3 fewer per capita herding hours per week than do the farm households. Note that own farm labor hours are the only labor activity to which households allocate significant amounts of time in 1991 whereas in 2004 allocated labor hours are more diversified, especially in the wage labor/entrepreneur group.

Overall, the evolution of small initial differences in asset holdings in 1991 into larger differences 13 years later suggests bifurcating welfare dynamics. However, although the cluster analysis identifies only two livelihood strategies in the data, and although they can be described within the generic farm and off-farm categories, the composition of the two livelihood strategies identified in the data show within-livelihood diversification. In fact, the diversification within livelihoods observed in this Kagera-specific sample has been observed in Tanzania more broadly: in a study of occupational choice using nationally representative data from 2010-2011 Tanzania, McCoullough (2016) finds that, in response to productivity gains in both sectors, households will diversify into self- and wage-employment without leaving farming. Therefore, comparison of the identified livelihoods, and consideration of the assets and allocations of which they are composed, suggests incremental and surmountable shifts within livelihoods. The question remains as to whether shifts between livelihoods are also incremental and surmountable.

3.6.2 Heterogenous and locally increasing returns

Marginal returns in consumption to each asset by livelihood strategy are shown in Figures 3.3 through 3.7 where the marginal returns are estimated at unit increments along the support of each asset, holding all other assets at their means. Below each marginal return figure is a kernel density plot showing the data density dissagregated by livelihood. The assets that offer statistically discernable returns by livelihood strategy are business assets (Figure 3.3), labor assets (Figure 3.4), and human capital assets (Figure 3.7).

Marginal returns to business assets (Figure 3.3) are increasing for individuals in farm households while they are indistinguishable from a flat line (constant returns) for the wage labor/entrepreneur group. However, the returns are higher for the wage labor/entrepreneurs except at the tail end of the asset distribution where returns for the two groups appear to converge. Individuals in the wage labor/entrepreneur livelihood enjoy greater returns to each hour of labor (Figure 3.4) than do the farmers.

While it appears that those in the wage labor/entrepreneur livelihood experience increasing returns to their land holdings (Figure 3.5), these estimates are based on extremely sparse data, as reflected by the density plot below the figure. Where the data are most dense, there is no distinguishable difference in returns to land holdings by livelihood strategy. Marginal returns to livestock holdings by livelihood strategy (Figure 3.6) are also statistically indistinguishable from one another. Returns to human capital assets in terms of years of education are greater in the wage labor/entrepreneur labor group (Figure 3.7); returns are slightly increasing for both livelihoods across the distribution. The data are dense at seven years of education, indicating the completion of primary

school.

Figure 3.3: Marginal returns to business assets by livelihood strategy

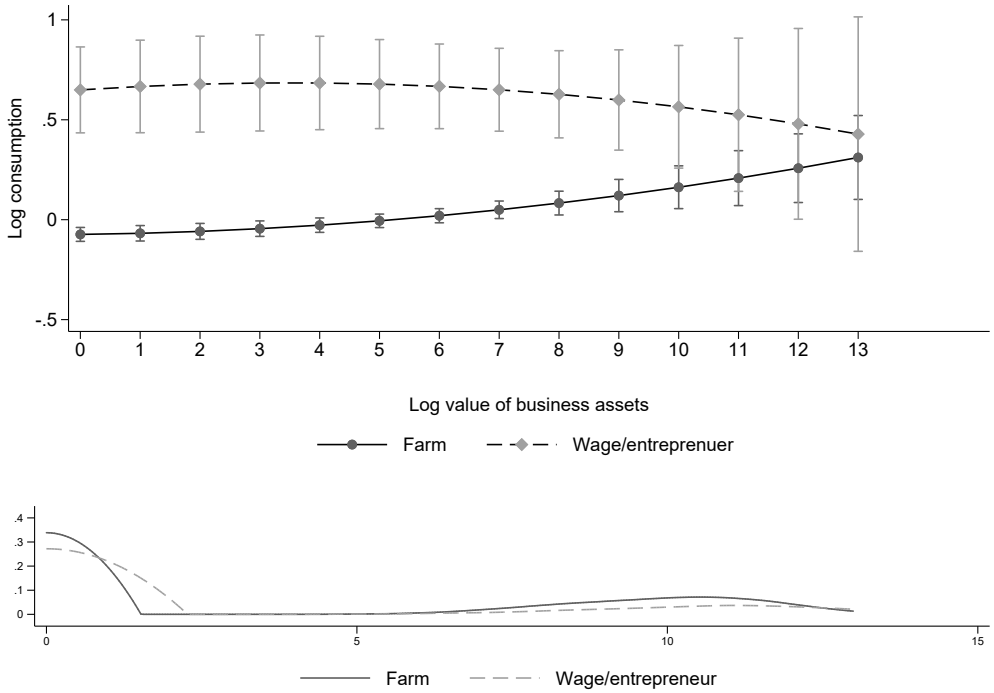


Figure 3.4: Marginal returns to labor by livelihood strategy

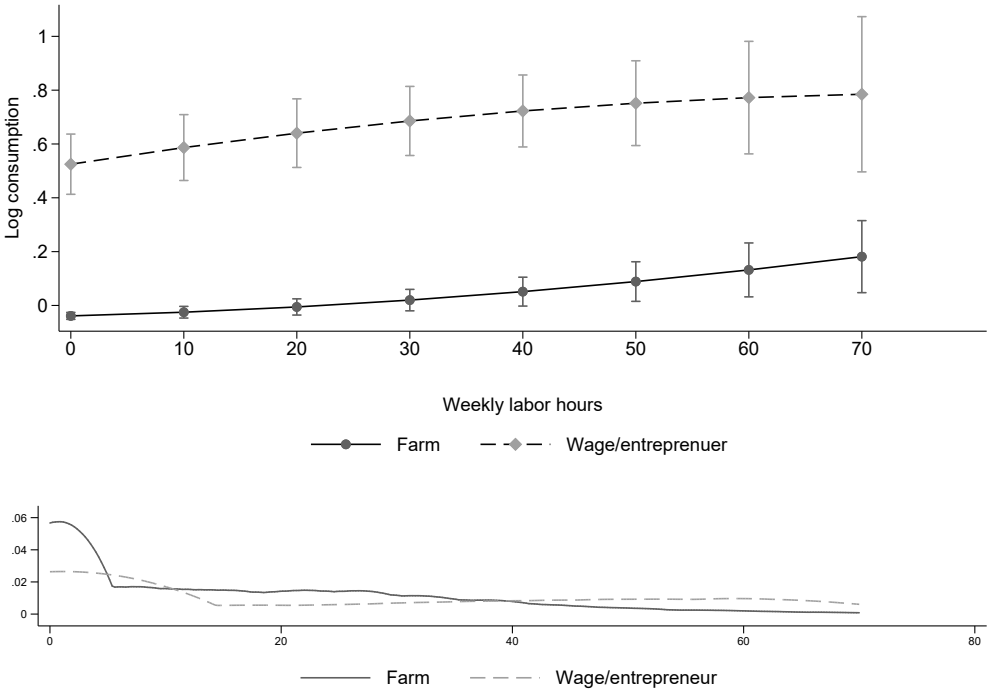


Figure 3.5: Marginal returns to land holdings by livelihood strategy

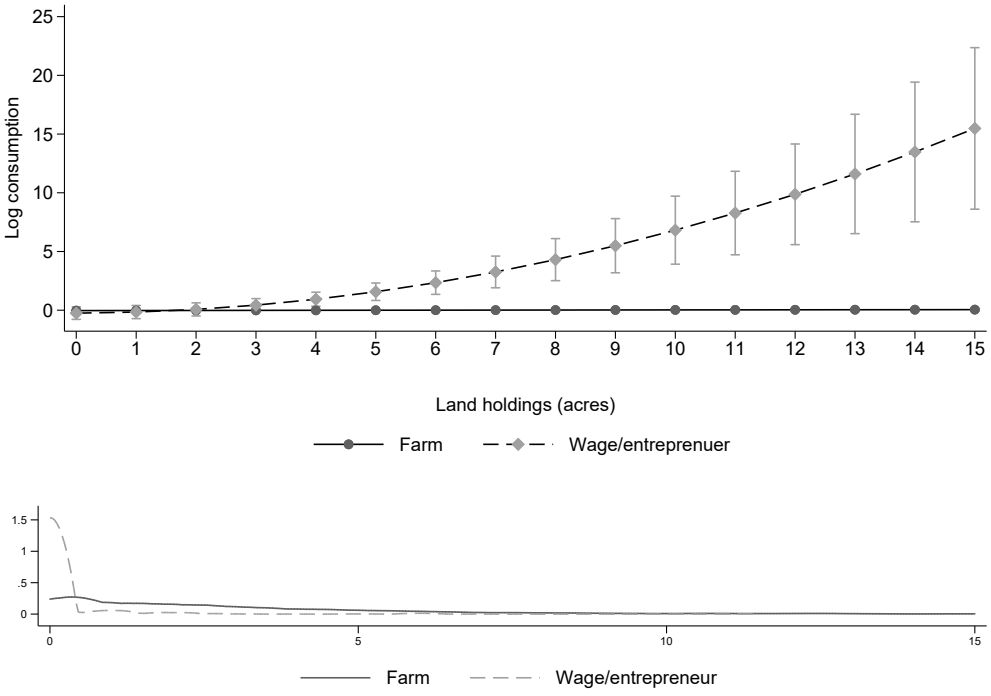


Figure 3.6: Marginal returns to livestock holdings by livelihood strategy

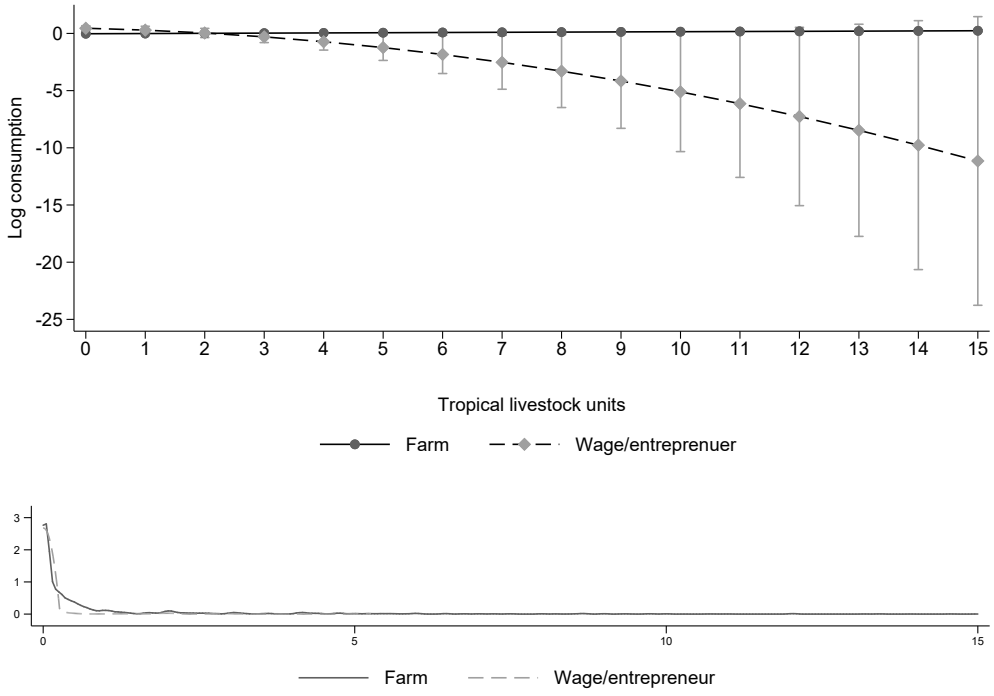
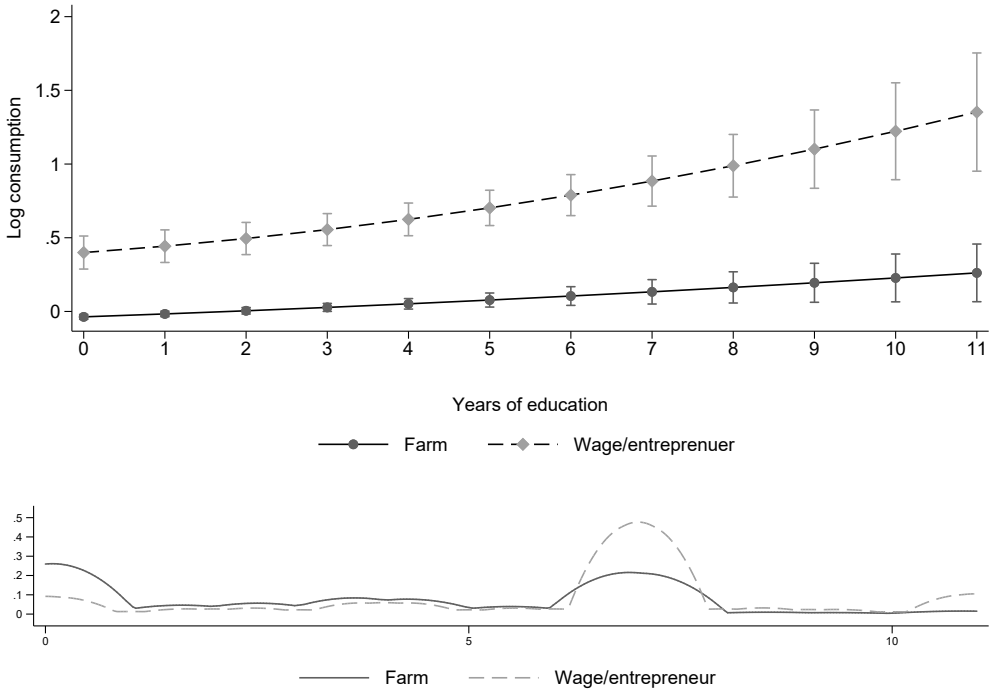


Figure 3.7: Marginal returns to education by livelihood strategy



Overall, the estimated marginal returns to assets by livelihood strategy suggest that, holding all else constant, if households could move from the farm to the wage labor/entrepreneur livelihood, they would experience greater returns to their business, labor, and human capital assets. However, we know from both the livelihood summary statistics as well as other research (Beegle et al. 2011, Christiaensen et al. 2013, and De Weerdt & Hirvonen 2016) that a great deal of migration is also occuring between 1991 and 2004 and that migration is correlated with the change from a farm to an off-farm livelihood. Therefore, the role of migration as an additional technology in this setting must also be considered. To do so, I treat migration as a technology that can interact with the identified livelihoods, estimating Equation 4.25 with three livelihoods instead of the original two: Remain & Farm, Move & Farm, and Move

& Wage/Entrepreneur. There is an insufficient number of observations of Remain & Wage/Entrepreneur to produce estimates for this group. The results are presented in Figures 3.8 through 3.10.

The returns to assets for those who move and switch livelihoods (Move & Wage/Entrepreneur) are greater than those who remain in farming, regardless of whether or not they have moved. Comparing the estimated returns to assets by livelihood (Figures 3.3 through 3.7) with those interacted with migration (Figures 3.8 through 3.10) suggests that most of the differences in returns are driven by livelihood status and not by migration alone. However, migration plays an important role.

Altogether, these findings support those of Beegle et al. (2011), Christiaensen et al. (2013), and De Weerdt & Hirvonen (2016) in showing that migration has played an important role in the increasing welfares of the Kagera households, regardless of livelihood strategy, and in showing that the combined strategy of migration plus adoption of an off-farm livelihood offers the highest returns. To return to the poor people or poor places question, these results show that while returns are not determined uniquely by geography, it clearly plays a role. In addition, these findings add nuance to those of Young (2013) who saw differentiated returns due to regional (rural/urban) demand for skill but did not consider livelihoods.

As with Young (2013), Gollin et al (2014), Herrendorf & Schoellman (2018), and Lakagos & Waugh (2013), my findings are consistent with a selection story in that those with higher education are in the off-farm livelihood. The long panel data as well as the spell length between panel waves means that I may be observing the return to households' livelihood choices following a failed liveli-

hood switch or a failed migration attempt from which the household has since returned (to initial livelihood or location). In addition, note that household composition is changing over the duration of the panel as households marry and move. Household fixed effects allow me to control for ”dynasty” effects such as a family having a greater initial endowment of intelligence, skill, or other unobservable resources in the first wave; however, they do not provide me with sufficient identification to claim that the same, e.g., labor assets would realize higher returns were the household to switch livelihood or location (or both).

Figure 3.8: Marginal returns to business assets by migration status

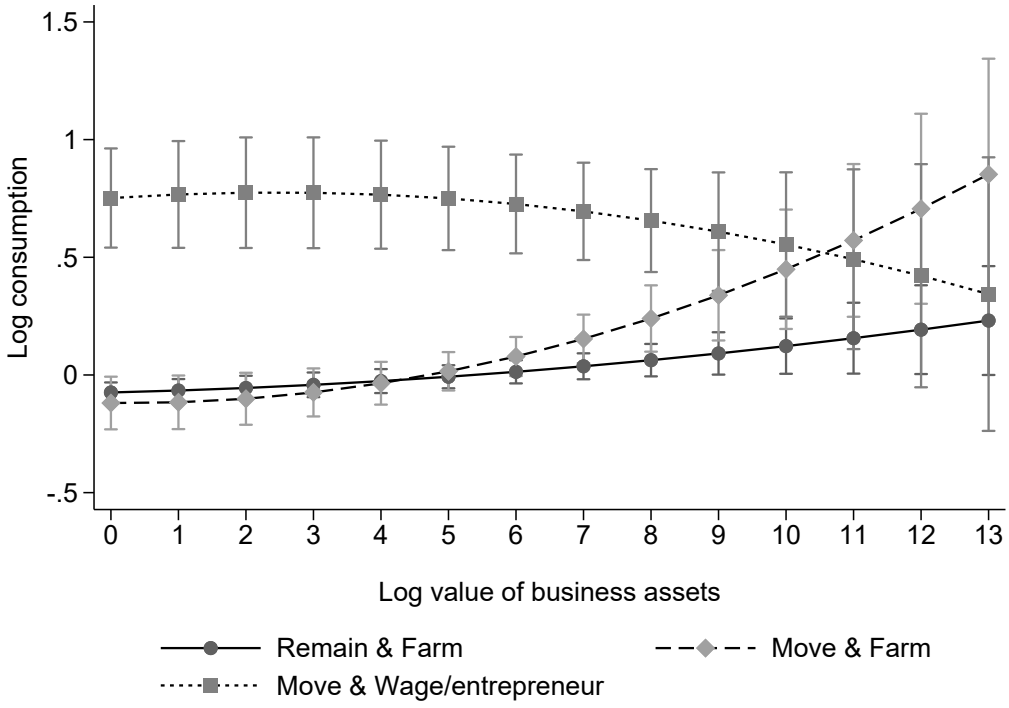


Figure 3.9: Marginal returns to labor by migration status

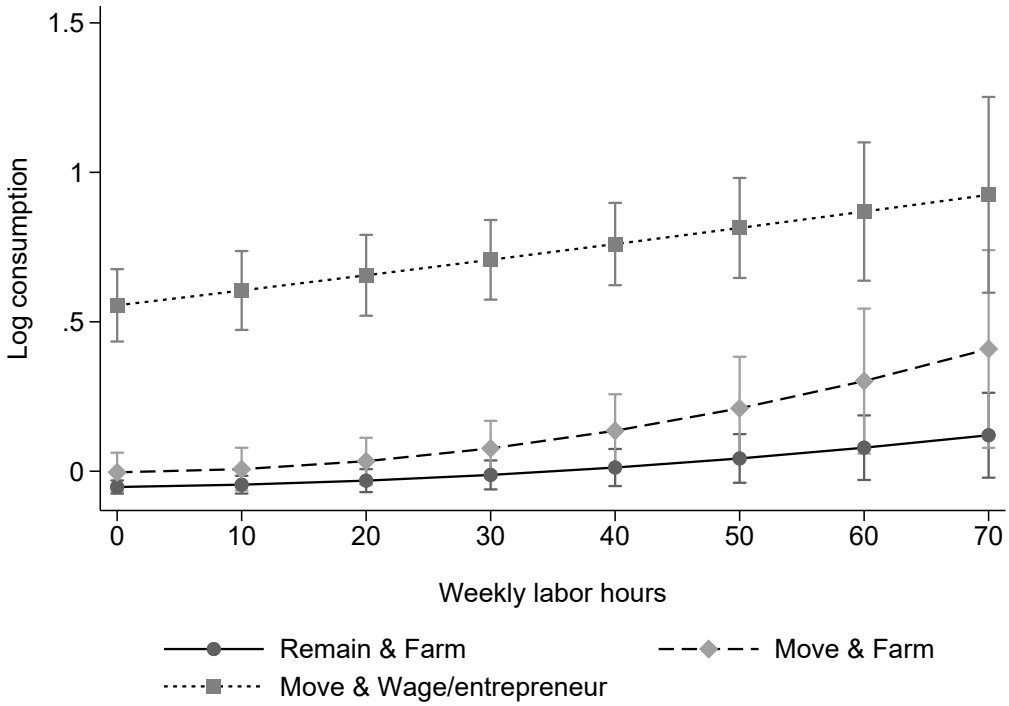
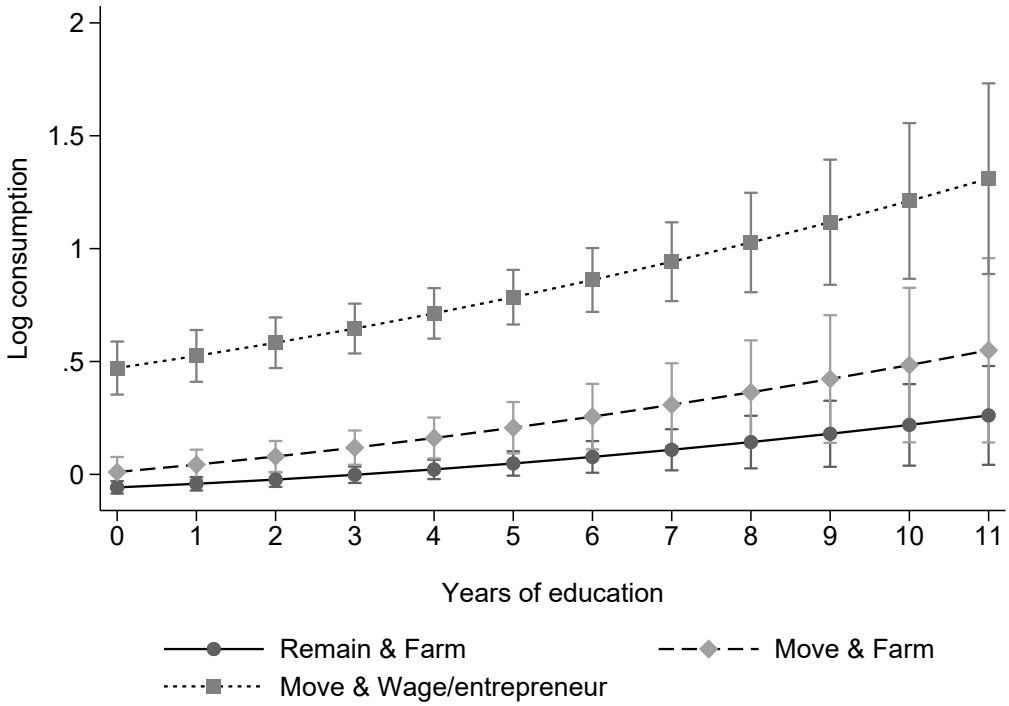


Figure 3.10: Marginal returns to education by migration status



Although we see results consistent with locally increasing returns between livelihoods, these are a necessary but not sufficient condition for multiple equilibria welfare dynamics. Therefore, Ill next estimate welfare dynamics by livelihood strategy.

3.6.3 Livelihood group welfare dynamics

Although the average household in the data is on a non-poor consumption dynamic path (Figure 3.11; horizontal and vertical lines indicate the poverty line), those households adopting the wage labor/entrepreneur strategy in 2004 enjoy a higher equilibrium in 2004 (Figure 3.11a) and 2010 (Figure 3.11b) than those

who do not. Whether considering mean population dynamics or livelihood specific dynamics, neither single nor multiple equilibria poverty traps emerge in this setting.

In comparing the 1991 to 2004 livelihood specific welfare dynamics (Figure 3.12a) with those of 2004 to 2010 (Figure 3.12b), we see conditional convergence give way to convergence. The structural transformation literature suggests that this convergence is due to increasing returns to factors in the low return sector, freeing up resources for other sectors (Timmer 1988, 2002; Gollin 2014). Unfortunately, due to data limitations, it is not possible to assess whether the relative welfare increase by 2010 of those households in the farm livelihood group in 2004 is due to increasing returns, livelihood transitions, or other causes.

Figure 3.11: Mean consumption dynamics (a) 1991 to 2004 (b) 2004 to 2010

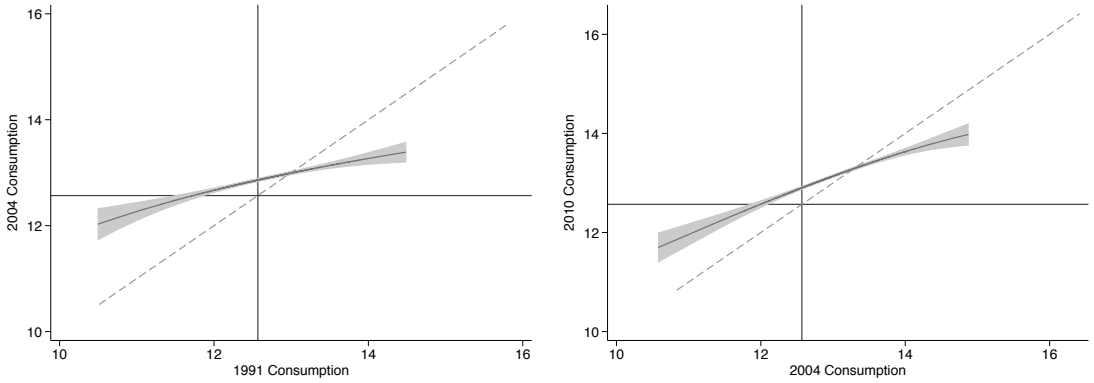
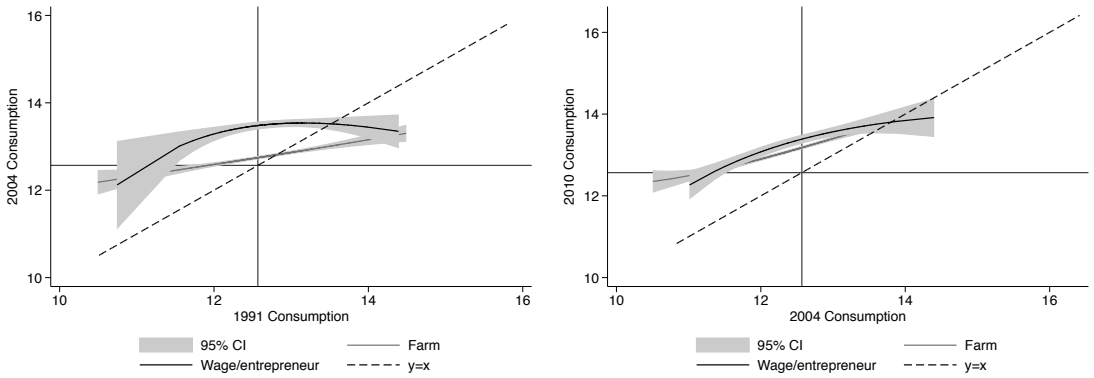


Figure 3.12: Consumption dynamics by 2004 livelihood strategy (a) 1991 to 2004 (b) 2004 to 2010



3.7 Conclusion

Using a flexible, theoretically grounded, data driven approach to the identification of livelihood strategies based on assets and their allocations, I observe the emergence of an off-farm livelihood between 1991 and 2004. Estimated returns to assets differ by livelihood, suggesting locally increasing returns in the move from one livelihood strategy to another; however, as with the rest of the literature (Young 2013, Gollin et al 2014, Herrendorf & Schoellman 2018, Lakagos & Waugh 2013), the greater returns are likely due to selection. While I am not able to identify the causes of the consumption gap, my findings add additional support to the literature on this phenomena in that my findings emerged from the data with minimal assumptions. Additionally, my findings suggest that livelihood change plays a greater role in increasing consumption than does geographic change. The asset content of each of the identified livelihood strategies is diverse, suggesting mobility. According to the observed welfare dynamics, neither livelihood group is trapped in poverty. However, when heterogeneity

in livelihood strategies is allowed for in the estimation of welfare dynamics, conditional convergence is observed. By 2010 the farm livelihood group has caught up to the wage/entrepreneur group, suggesting convergence in welfare. Despite beginning with a flexible framework and employing a data driven strategy, findings support many of the stylized facts of the structural transformation literature such as the emergence of two sectors, sector-differentiated returns to labor and other factors, and catch up in the low return sector.

This exercise – the estimation of welfare dynamics over heterogeneous livelihoods that have been identified in a data driven manner – and its findings (farm and off-farm livelihoods, locally increasing returns, conditional convergence, and convergence) have several important implications. First, the evolution from a single livelihood in 1991 to two livelihoods in 2004 suggests that there exist serious limitations to the estimation of welfare dynamics over a single asset or just those assets that are observed to play a large role in household livelihoods at baseline, as is done in much of the welfare dynamics literature. For example, if one were to estimate returns to only land and livestock assets between 1991 and 2004, it would appear as though the wage labor/entrepreneur group was earning much higher returns on much smaller asset holdings than the farm group, when in fact they are relying on returns to other productive assets such as human capital and business investments. Likewise, welfare dynamics estimated over land and livestock assets alone would be extremely misleading for the wage labor/entrepreneur group, as holdings collapse to near zero for these households; we might spuriously conclude that the wage labor/entrepreneur group is trapped in poverty when in fact they've switched to a (more lucrative) livelihood that relies on a different set of assets.[10] The analysis also suggests

[10] Additional limitations of asset based welfare analysis in the Kagera data have been demonstrated by De Weerdt (2010), who used quantitative and qualitative evidence to explore why

that estimation of welfare dynamics at population means, without allowing for heterogeneity to emerge, masks policy relevant findings. Whether subsets of households are facing poverty traps, conditional convergence, or eventual convergence, so long as we're able to observe their plights and prospects appropriate policies and interventions can be designed to meet their needs.

How can we reconcile the observed differences in returns to assets between livelihoods in the (likely) presence of market failures – i.e., the two conditions that give rise to poverty traps – with a failure to observe multiple welfare equilibria in this setting? We have seen that new livelihoods can emerge over time, meaning that even if the livelihood choice set is non-convex, it is not fixed. Moreover, the content of each livelihood strategy is diverse, suggesting incremental movement within, and possibly between, livelihoods. We also see convergence in returns to assets once migration is accounted for in the estimation. As an additional technology, migration increases returns to a livelihood because individuals are moving to more connected locations in terms of roads, markets, and other infrastructure, as observed by De Weerdt (2010), Beegle et al. (2011), and Christiaensen et al. (2013). In addition, market failures are household specific and a matter of degree; as a household moves to a more connected area, that household may also be less constrained by market failures.

and how individuals deviated from their asset-based growth path trajectories. Through focus group discussions, De Weerdt (2010) finds that those whose asset growth between 1991 and 2004 exceedes their predicted asset growth are more likely to have diversified their farming activities (food crops, cash crops, and livestock), expanded their land holdings, and diversified into non-farm activities (national and international food trade, small shop ownership). Those whose asset growth underperformed relative to their predicted growth were more likely to have experienced major illness or death in the family. He ascribes the failure of his predictive model to: a failure to account for occupational choices (i.e. diversification decisions), shocks (i.e. death and illness, price shocks, weather shocks), unobservables (social capital in terms of networks and trust, experience in trade, and exposure to life outside their village), and model specification error (omitted interactions between village remoteness and initial conditions), several of which he is able to identify through qualitative analysis. While his comments are focused on the Kagera data, De Weerdt's (2010) insights on the limitations of asset based welfare analysis apply to such analyses in general.

The absence of a multiple equilibria welfare dynamics in this setting – where heterogeneity of welfare and conditional convergence are observed – has implications for and raises important questions about appropriate anti-poverty intervention points. It is generally challenging to distinguish cases of conditional convergence from a poverty trap (Ghatak 2015, Barrett & Carter 2013), and convergence may be so slow as to make the promise of convergence practically meaningless, as eventual attainment of a high equilibrium is little consolation to households facing long run poverty and inequality. There is a long-standing debate in the academic (and public) anti-poverty programming discourse as to whether intervention stifles local growth and innovation, leaving households, regions, and nations dependent upon the benevolence of donors (Easterly 2006) or is absolutely necessary to assist households in reaching higher, long-run growth paths (Sachs 2005). A productive way forward may be to assess the heterogeneous treatment effects of anti-poverty programs using innovative methods developed by Athey and co-authors (Athey & Imbens 2016, Wager & Athey 2017); this is an objective of my future work.

CHAPTER 4

RISK, RETURNS, AND WELFARE

∗*This chapter was written in collaboration with Leah Bevis*

4.1 Introduction

A positive correlation between the riskiness and returns of households' asset portfolios and their initial asset endowments is often taken for granted in development economics, especially in settings where financial market failures are likely (Barrett *et al.* 2016). The relationship among risk, returns, and welfare has important implications for the reproduction of inequality and persistent poverty and therefore is critical to understand for effective anti-poverty policy making. If a household with a low initial asset endowment is constrained to low return economic activities (or, similarily, if higher return activities come with greater risk and household risk preferences induce the household to chose the low risk, low return activities), then not only will that household remain poor, but the gap between those with a low endowment and those with a high endowment will only grow overtime, meaning growing inequality.

The relationship among risk, returns, and welfare is particularly salient in the case of the Sub-Saharan African household for which multiple household specific market failures, including insufficient access to credit and insurance markets, can compound and for which proximity to asset-based poverty thresholds (e.g. sufficient land holdings to make investments in inputs and mechanization economically viable, sufficient herd size to sustain transhumance) might produce risk seeking or risk avoiding behaviors as households either

seek to move above or remain above an asset threshold (Barrett *et al.* 2016, Barrett *et al.* 2006). Moreover, the relationship has important implications for popular interventions such as index insurance, microfinance, and social safety nets – interventions designed to correct the sort of market failures that can induce households to choose low-risk low-return activities despite the existence of more lucrative options.

While there is a lot of theoretical support for the existence of a positive correlation among among risk, returns, and welfare (Eswaran & Kotwal 1990, Deaton 1991, Zimmerman & Carter 2003), empirical evidence is thin. Most available empirical evidence considers agricultural (Di Falco & Chavas 2006, 2009) or livestock portfolios only, or a combination of the two (e.g. Rosensweig & Binswanger 1993, Dercon 1996) and cannot account for efforts to mitigate agricultural risk via off-farm diversification. In addition, most theoretical and empirical treatments of the relationship among initial welfare, risk, and returns stop at the mean-variance approach to portfolio choice (Meyer 1987, Markowitz 1952, Just & Pope 1979) and therefore either fail to account for, or assume away, the possible influence of downside risk (skewness), implicitly placing strong assumptions on risk preferences (Chavas 2004).

Do initial [1] asset holdings determine the riskiness (both upside and downside risk) and expected returns of a household's asset portfolio? We investigate this question in the increasingly economically diversified setting of Tanzania using unsupervised learning methods to identify the set of productive asset portfolios available in the Tanzanian economy and Antle's (1983) moments approach to estimate the first three conditional moments (mean, variance, and skewness) of the returns to those asset portfolios. We then non-parametrically

[1] Here we mean "initial" in the sense of initial conditions in a first order Markov process.

estimate the relationship among the conditional moments as well as the relationship between the value of initial asset holdings and the estimated moments for each portfolio.

In addition, assuming constant relative risk aversion (CRRA), we estimate the risk premium associated with each portfolio. The risk premium is defined as the amount of money one would be willing to pay to insure against risk or, similarly, the amount one would be willing to pay to hold the source of random variation at its mean (DiFalco & Chavas 2009).

In subsequent sections we situate our research question in the theoretical and emprical literature on the relationship among risk, returns, and welfare, develop a theoretical model and derive the risk premium, and explain the methods, data, and results.

4.2 Background

Current thinking about the relationship among risk, returns, and welfare in settings of multiple financial market failures – in particular the idea that, absent credit and insurance markets, initial endowments may determine households' access to high risk, high return assets–draws on a long theoretical and empirical literature. Eswaran & Kotwal (1990) provide a theoretical model demonstrating that, even where risk preferences are identical, differences in risk behavior can emerge from differences in access to credit. Deaton (1991) shows that in the face of income risk and borrowing constraints, assets serve as a buffer stock and are used to smooth consumption.

Building on Deaton's (1991) theoretical work, Dercon (1996) shows empirically that rural Tanzanian households that have built up a large buffer stock of liquid assets, such as livestock, are more likely to undertake high risk activities. Rosenzweig & Binswanger (1993) find that farming households in regions of India with greater rainfall variation choose asset portfolios that are less sensitive to rainfall risk and therefore offer lower returns. Dercon (1998) shows through simulation and empirical estimates that rural Tanzanian households that have greater initial endowments are more likely to enter into the high-risk, high-return activity of cattle raising as compared with relatively lower risk, lower return activities such as growing crops or wage labor. Carter (1997) finds a high correlation between low initial land endowments and effective risk exposure in Burkina Faso, suggesting both that households with lower endownments will be less inclined to adopt new technologies and that asset inequality will grow. Zimmerman & Carter (2003) build on Dercon (1996, 1998) by allowing for divisible assets (i.e. non-livestock assets such as grain and land) and endogenizing asset price risk in the portfolio selection choice. Using parameters from rural households in Burkina Faso, Zimmerman & Carter (2003) show, via simulation, that the correlation between initial wealth and return on investments is a consequence of asset-based risk coping in an environment where insurance and savings mechanisms are not available. They also demonstrate that this relationship produces self-perpetuating inequality and poverty.

Both theoretical and emprical analyses of welfare dynamics suggest that the relationship between risk and welfare may not be strictly monotonic. Barrett *et al.* (2006), Carter & Barrett (2006), Lybbert & Barrett (2011), Carter & Lybbert (2012), and Lybbert *etal* (2013) show that just below a dynamic asset-based welfare threshold (also known as the "Micawber threshold," the dynamic threshold

at which high and low welfare equilibria bifurcate) agents should be more risk-seeking on the chance that a positive random draw will push them above the threshold, while those just above the threshold should be more risk averse (and more inclined to asset smooth than to consumption smooth) on the chance that a negative random draw will push them below the threshold.

While the theoretical literature has established a relationship among risk, returns, and initial asset-based welfare (Eswaran & Kotwal 1990, Deaton 1991, Zimmerman & Carter 2003), the empirical literature has been constrained to analyses of agricultural and livestock portfolios (Di Falco & Chavas 2006, 2009, Rosensweig & Binswanger 1993, Dercon 1996) and therefore cannot account for the role of off-farm diversification in risk management. In addition, extension of the analysis beyond the mean-variance approach to portfolio choice (Meyer 1987, Markowitz 1952, Just & Pope 1979) to consideration of downside risk (skewness) is important, as variance alone does not allow one to distinguish between upside and downside risk (Chavas 2004, Di Falco & Chavas 2006, 2009). As an example, in an analysis of crop biodiversity, Di Falco & Chavas (2009) find that biodiversity and soil fertility increase risk but decrease downside risk. Had their analysis only considered the second moment of the yield distribution, it would have found that biodiversity and soil fertility were welfare decreasing for a risk-averse farmer when in fact these inputs reduce the farmer's exposure to downside risk.

The contribution of this paper is to move beyond the agricultural inputs/livestock assets models as well as beyond the mean-variance approach to asset portfolio selection to empirically examine the extent to which the mean, variance, and skewness of returns to a diverse array of asset holdings is cor-

related with initial asset endowments, and therefore the extent to which the heterogeneity of initial endowments may perpetuate poverty and inequality.

4.3 Theory

Without an ability to control for all market constraints, such as a lack of access to credit or insurance, it is not possible to disentangle risk preferences from observed risk behavior. Indeed, this is the finding in Eswaran & Kotwal (1990); even where preferences are identical, household-specific capital and credit market failures may heterogeneously impact risk premia and therefore the risk behavior of households. Likewise, Lybbert *etal* (2013) find that observed risk behavior may be a response to wealth dynamics that are observable to the household but not to the econometrician, resulting in the econometric misattribution of observed risk behavior to risk preferences. Therefore, the focus of this paper will be on observed risk behavior in terms of the riskiness of households' selected asset portfolios. Initially, we will make no assumptions about household risk preferences. To parameterize the risk premium, we will assume constant relative risk aversion (CRRA), which has the nice feature that it allows for decreasing absolute risk aversion (DARA).

Following Rosenzweig & Binswanger (1993), Chavas (2004), and Di Falco & Chavas (2006, 2009), let a household have asset stock vector $\boldsymbol{x}_t$ at time t, $\boldsymbol{x}_t \geq 0$ (let bolding indicate vectors throughout). The household enjoys the returns, $\boldsymbol{g}(\boldsymbol{x}_t, \boldsymbol{z}_t, \boldsymbol{v}_t)$, as a function of the household's assets, other productive inputs, $\boldsymbol{z}_t$, $\boldsymbol{z}_t \geq 0$, and a vector of random variables, $\boldsymbol{v}_t$, that includes such sources of random variation as rainfall, sales prices unknown to the household at the time of

the productive allocation of its assets, and other sources of variation in returns. One should think of $\boldsymbol{g(.)}$ as the set of livelihoods or technologies available in the economy with which households can derive a flow of profits from their stock of productive assets. The household also realizes expenses associated with purchased inputs at price, $\boldsymbol{p_{zt}}$. The net profit function vector is,

$$\boldsymbol{\Pi_t = g(x_t, z_t, v_t) - p'_{zt} z_t} \tag{4.1}$$

Note that we are assuming away exogenous unearned income, as our focus here is on returns to asset holdings and investments. We are also assuming away implicit rental payments on owned assets.

The household's ability to access credit each period is a function, $\boldsymbol{\eta_t(x_t)}$ of its asset holdings in each period. In the context of this model, where it is assumed that savings are in terms of productive assets only, one can think of $\boldsymbol{\eta_t(x_t)}$ as a liquidity function, as it might involve collateralizing assets so as to access credit, reselling assets, and dissaving. The household can also invest in additional assets, $\boldsymbol{I_t}$ (denominated in the same units as the assets vector, $\boldsymbol{x_t}$). Therefore the household's budget constraint is,

$$\boldsymbol{C_t \leq \Pi_t + \eta_t(x_t) - I_t} \tag{4.2}$$

where $\boldsymbol{C_t}$ represents the household's consumption in period t, $\boldsymbol{C_t \geq 0}$.

The household realizes costs (or gains) each period due to the depreciation (or appreciation)[2] of its asset holdings, $\boldsymbol{\delta \in (-1, 1)}$, where negative values cap-

[2]For example, assets such as a tractor may depreciate in value while assets such as livestock

ture asset appreciation and postive values capture asset depreciation. Therefore, the household's asset stock evolves following a deterministic law of motion,

$$\boldsymbol{x}_{t+1} = (\mathbf{1} - \boldsymbol{\delta})\boldsymbol{x}_t + \boldsymbol{I}_t \tag{4.3}$$

Let the household's consumption risk preferences be represented by a von Neumann-Morgenstern utility function, $U(\boldsymbol{C}_t)$. Let the utility function be additively separable over time with $\rho \in [0, 1]$ representing the household's discount factor. We use a consumption-based (as opposed to an income-based or terminal wealth-based) model of economic behavior under risk in aknowledgement that consumption and production decisions are generally non-separable in the setting under study. The household's objective is to maximize the expected utility of consumption subject to the given budget, asset accumulation, and non-negativity constraints,

$$Max_{\boldsymbol{I}_t, \boldsymbol{C}_t, \boldsymbol{x}_{t+1}, \boldsymbol{z}_t} \Sigma_{t=0}^{T} \rho^t \Big\{ EU(\boldsymbol{C}_t) \Big\} \tag{4.4}$$

s.t.

$$\boldsymbol{C}_t \leq \boldsymbol{g}(\boldsymbol{x}_t, \boldsymbol{z}_t, \boldsymbol{v}_t) - \boldsymbol{p}'_{zt}\boldsymbol{z}_t + \boldsymbol{\eta}_t(\boldsymbol{x}_t) - \boldsymbol{I}_t \tag{4.5}$$

$$\boldsymbol{x}_{t+1} = (1 - \boldsymbol{\delta})\boldsymbol{x}_t + \boldsymbol{I}_t \tag{4.6}$$

$$\boldsymbol{x}_{t+1}, \boldsymbol{z}_t, \boldsymbol{C}_t \geq 0 \tag{4.7}$$

$$\boldsymbol{x}_t \geq 0, \quad given \tag{4.8}$$

may beget more livestock and therefore increase in value.

For simplicity, let C_t be a scalar capturing the monetary value of all consumed goods, whether produced or purchased and let $lim_{c\to 0}U'(C) = \infty$, meaning $C_t = 0$ will never be optimal (and we therefore do not need to address the non-negativity constraint on consumption; it will not bind). Substituting the budget constraint into the asset law of motion, the remaining choice variables are C_t, $\boldsymbol{x}_{t+1}$, and $\boldsymbol{z}_t$. The Lagrangean expression of the problem is,

$$\begin{aligned} L = \Sigma_{t=0}^{T}\rho^t\Big\{EU(C_t) + \lambda_t[(\mathbf{1}-\boldsymbol{\delta})'\boldsymbol{x_t} + \\ \boldsymbol{g}(\boldsymbol{x_t}, \boldsymbol{z_t}, \boldsymbol{v_t}) - \boldsymbol{p}'_{zt}\boldsymbol{z_t} + \boldsymbol{\eta_t}(\boldsymbol{x_t}) - C_t - \boldsymbol{x}_{t+1}] + \gamma_{t+1}x_{t+1} + \zeta_t z_t\Big\} \end{aligned} \tag{4.9}$$

The first order conditions (letting subscripts indicate derivatives) are,

$$\frac{\partial L}{\partial C_t} : EU_{C_t} - \lambda_t = 0 \tag{4.10}$$

$$\begin{aligned} \frac{\partial L}{\partial \boldsymbol{x_{t+1}}} : -\rho^t\lambda_t + \rho^{t+1}\lambda_{t+1}[(1-\delta) + \boldsymbol{g}_{x_{t+1}}(\boldsymbol{x_{t+1}}, \boldsymbol{z_{t+1}}, \boldsymbol{v_{t+1}}) \\ +\eta_{t+1,x_{t+1}}(x_{t+1})] + \rho^t\gamma_{t+1} = 0 \end{aligned} \tag{4.11}$$

$$\frac{\partial L}{\partial \boldsymbol{z_t}} : \rho^t\lambda_t[\boldsymbol{g}_{z_t}(\boldsymbol{x_t}, \boldsymbol{z_t}, \boldsymbol{v_t}) - \boldsymbol{p}_{zt}] + \rho^t\zeta_t = 0 \tag{4.12}$$

With K-T conditions,

$$
\begin{aligned}
\gamma_{t+1}\boldsymbol{x}_{t+1} &= 0 \\
\gamma_{t+1} &\geq 0 \\
\boldsymbol{x}_{t+1} &\geq 0 \\
\lambda_t &\geq 0 \\
\zeta_t \boldsymbol{z_t} &= 0 \\
\boldsymbol{z_t} &\geq 0 \\
\zeta_t &\geq 0
\end{aligned}
\tag{4.13}
$$

Simplifying and combining Equations 4.10 and 4.11 offers the Euler equation, relating consumption in time t to that in time $t + 1$,

$$
EU_{C_t} = \rho EU_{C_{t+1}}[(1-\delta) + \boldsymbol{g}_{x_{t+1}}(\boldsymbol{x_{t+1}}, \boldsymbol{z_{t+1}}, \boldsymbol{v_{t+1}}) + \eta_{t+1,x_{t+1}}(x_{t+1})] + \gamma_{t+1} \tag{4.14}
$$

Equation 4.14 tells us that, where the Lagrange multiplier for the liquidity constraint γ_{t+1} is non binding ($\gamma_{t+1} = 0$), the marginal utility of current consumption over the first three moments of the consumption distribution must equal the discounted marginal utility of the first three moments of the next period's consumption distribution, moderated by the marginal return of next period's assets. However if $\gamma_{t+1} > 0$, the marginal utility will be greater at time t than at $t + 1$ (Deaton 1991, Dercon 1996, 1998), meaning that the household will consume rather than invest, dissaving all its assets. Because $\boldsymbol{g}(\boldsymbol{x_t}, \boldsymbol{z_t}, \boldsymbol{v_t})$ is a function of household asset holdings, inputs, and exogenous variation, if the liquidity

constraint binds and there are any barriers to entry to a particular livelihood or technology, e.g. a minimum asset threshold for participation in a given livelihood, then we would expect to find not only different optimal asset accumulation paths based on initial asset holdings but also different conditional consumption distributions among households.

Following Antle (1987), Chavas (2004), and Di Falco & Chavas (2006, 2009), a local approximation of the household's risk premium can be derived from the household's utility function over the first three moments of the distribution of the utility of consumption. Inclusion of the third moment, skewness, represents a departure from the mean-variance portfolio choice literature, which assumes either that the household has no preferences over the third moment or that the third moment is zero (Meyer 1987, Chavas 2004). Inclusion of the third moment relaxes this assumption, allowing downside risk to play a role in determining the household's utility of consumption. Let the utility function, $U(C)$ be thrice continuously differentiable. And let $E(C)$, $V(C)$, and $S(C)$ indicate the mean, variance, and skewness of the utility of consumption distribution, respectively.

To derive the risk premium, set the utility function equal to the certainty equivalent,

$$U(C) = U(E(C) - R) \tag{4.15}$$

and take a third order Taylor Series expansion of $U(C)$ with respect to C, letting U^d indicate the d^{th} derivative of the utility function,

$$U(C) = U(E(C)) + U^1E(C - E(C)) + \frac{1}{2}U^2E(C - E(C)^2 + \frac{1}{6}U^3E(C - E(C))^3 \tag{4.16}$$

Take the expectation,

$$EU(C) = EU(E(C)) + U^1(C - E(C)) + \frac{1}{2}U^2(C - E(C))^2 + \frac{1}{6}U^3(C - E(C))^3 \quad (4.17)$$

Because $E(C - E(C)) = 0$, $E(C - E(C))^2 = V(C)$, and $E(C - E(C))^3 = S(C)$, we can simplify to,

$$EU(C) = EU(E(C)) + \frac{1}{2}U^2V(C) + \frac{1}{6}U^3S(C) \quad (4.18)$$

Take a first order Taylor series expansion of $U(E(C) - R)$ with respect to R

$$U(E(C) - R) = U(E(C)) - U^1R \quad (4.19)$$

Setting the two results equal (i.e., substituting them back into the risk premium definition in Equation 4.15) and simplifying, we have an expression for the household's risk premium,

$$R = -\frac{1}{2}\frac{U^2}{U^1}V(C) - \frac{1}{6}\frac{U^3}{U^1}S(C) \quad (4.20)$$

In Equation 4.20, $-\frac{U^2}{U^1}$ is the Arrow-Pratt (AP) coefficient of risk aversion; it captures the proportionality between the risk premium and the variance of risk in the neighborhood of the riskless case, *i.e.*, where R=0 (Chavas 2004). Likewise, $-\frac{U^3}{U^1}$ is the coefficient of downside (DS) risk aversion; it captures the proportionality between the risk premium and the skewness of risk in the neighborhood of the riskless case (Menezes *etal* 1980).

With a few assumptions, the variance and skewness of the consumption distribution, $V(C)$ and $S(C)$, can be estimated from the data (Antle 1983); this pro-

cedure will be detailed in the methods section. Identifying and/or estimating appropriate values for AP and DS is more challenging. A structural approach is offered in Antle (1978). To avoid making structural assumptions that would allow for estimation of these parameters but assume away many of the constraints motivating this analysis (e.g. non-separability), we instead follow Di Falco & Chavas (2006) and assume a constant relative risk aversion (CRRA) utility function of the form $U(C) = -C^{1-r}$, where r is the relative risk aversion parameter, to parameterize AP and DS in Equation 4.20. Details are provided in the methods section.

4.4 Data

The analysis draws on three waves of Living Standards Measurement Study-Integrated Surveys on Agriculture (LSMS-ISA) data from Tanzania: 2008-09, 2010-11, and 2012-13. The survey follows an original 3,280 households from 2008 to 2013; the overall household attrition rate for the panel is 4.84 percent (NBS 2014).

The Tanzanian setting is particularly compelling for an empirical analysis of the relationship among risk, returns, and initial asset holdings. As documented by Christiaensen *etal* (2013), De Weerdt (2010), Beegle *etal* (2011), and McBride (2018), the Tanzanian economy has been undergoing structural transformation, entailing both migration and the diversification of livelihoods, since the late 1990s.

Summary statistics of the data are presented in Table 4.1. All asset variables are presented in terms of their Tanzanian Shilling (TSh) equivalence, i.e., the

value one would get from selling the asset. All monetary values are spatially deflated within-wave using a Fisher Index generated for this purpose by the Tanzania National Statistics Office. The 2008-09 and 2010-11 data are then inflated to 2012-13 values so that all data are in real 2012-13 terms. Total household expenditures and all asset values have been transformed via inverse hyperbolic sine transformation (abbreviated as asinh throughout) so as to approximate a natural log transformation without modifying or sacrificing zero-valued assets.

As seen in Table 4.1, the sample[3] is majority rural, with the share of rural rising significantly by 2012-13 as a consequence of households breaking off, due to marriage or migration, from the original. Generally, household heads have completed some primary school education. Household expenditures and plot size fall over the course of the panel while the value of livestock holdings generally fall between 2008-09 and 2010-11, likely due to severe droughts that swept the region in 2009 and 2011, but then recover by 2012-13. The value of business assets (captial, stock, goods) rises over the course of the panel while the value of household, fishing, and farming assets follows no discernable trend.

[3]The summary statistics are reported without survey weights.

Table 4.1: Tanzania LSMS-ISA summary statistics by year

	2008-09		2010-11		2012-13	
	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev
Rural	0.63	0.48	0.64	0.48	0.84	0.36
Household size	5.48	3.22	5.50	3.40	4.99	3.08
Adult equivalence	4.51	2.66	4.52	2.77	4.08	2.50
Head of household female	0.25	0.43	0.25	0.43	0.26	0.44
Head age	46.73	15.32	46.62	15.75	46.06	16.06
Head married	0.72	0.45	0.72	0.45	0.69	0.46
Head migrated to this area	0.45	0.50	0.49	0.50	0.58	0.49
Head has completed ≤ primary	0.43	0.50	0.42	0.49	0.40	0.49
Head has completed primary	0.39	0.49	0.47	0.50	0.48	0.50
Head has completed secondary	0.10	0.30	0.10	0.30	0.11	0.32
Head has completed university	0.01	0.09	0.01	0.08	0.01	0.09
Total household expenditures	15.76	0.76	15.72	0.76	15.61	0.76
Value of plot holdings	9.37	7.17	9.04	7.41	8.68	7.42
Value of bull holdings	1.24	4.03	0.97	3.61	1.17	3.91
Value of cow holdings	2.06	5.08	1.51	4.45	1.75	4.72
Value of steer holdings	0.87	3.51	0.71	3.16	0.84	3.43
Value of heifer holdings	0.88	3.41	0.57	2.81	0.91	3.48
Value of male calf holdings	1.03	3.61	0.55	2.73	0.98	3.51
Value of female calf holdings	1.06	3.68	0.67	3.12	1.03	3.63
Value of goat holdings	2.65	5.24	1.85	4.61	2.21	4.95
Value of sheep holdings	0.83	3.14	0.60	2.73	0.66	2.85
Value of pig holdings	0.71	2.95	0.38	2.20	0.53	2.61
Value of chicken holdings	5.22	5.94	3.64	5.58	4.47	6.08
Value of other livestock holdings	1.80	3.95	1.19	3.51	0.00	0.00
Value of hand hoe farm asset	7.01	4.62	6.88	4.68	6.36	4.69
Value of sprayer farm asset	0.55	2.46	0.46	2.25	0.46	2.22

Value of ox plough farm asset	0.84	3.20	0.88	3.50	0.86	3.45
Value of seed planter farm asset	0.01	0.39	0.85	3.18	0.81	3.07
Value of ox cart farm asset	0.31	2.05	0.00	0.02	0.00	0.27
Value of tractor farm asset	0.02	0.65	0.28	1.96	0.25	1.83
Value of tractor plow farm asset	0.03	0.69	0.05	0.91	0.02	0.60
Value of tractor harrow farm asset	0.00	0.03	0.02	0.60	0.02	0.50
Value of shellerthresher farm asset	0.03	0.54	0.00	0.02	0.01	0.41
Value of watering can farm asset	0.13	1.11	0.11	1.05	0.07	0.87
Value of farm buildings	0.55	2.48	0.57	2.55	0.39	2.09
Value of geri cansdrums farm asset	0.88	2.93	0.30	1.80	0.20	1.46
Value of fishing nets	0.27	1.88	0.16	1.42	0.05	0.79
Value of fishing lines	0.18	1.27	0.12	1.12	0.00	0.21
Value of fishing boats	0.28	1.96	0.16	1.47	0.11	1.19
Value of fishing motors	0.07	1.01	0.02	0.52	0.01	0.41
Value of business capital	3.44	5.87	3.73	6.02	4.77	6.08
Value of business stock	0.51	2.44	0.53	2.44	0.95	3.11
Value of business goods	1.98	4.68	2.04	4.73	2.83	5.27
Value of land line phone	0.30	1.86	0.20	1.54	0.06	0.85
Value of mobile phone	5.85	6.13	7.85	5.92	8.33	5.30
Value of fridgefreezer	1.68	4.44	1.85	4.62	1.64	4.32
Value of sewing machine	1.49	4.05	1.38	3.91	1.17	3.60
Value of computer	0.50	2.66	0.50	2.62	0.52	2.63
Value of gaselectric stove	0.93	3.34	0.89	3.26	0.76	2.95
Value of stove	4.47	4.98	5.26	4.98	3.76	4.58
Value of vehicle	0.63	3.24	0.62	3.18	0.54	2.96
Value of motorcycle	0.46	2.60	0.70	3.17	0.87	3.47
Value of bicycle	4.90	6.00	5.34	6.08	4.62	5.80

All monetary values have been transformed via the inverse hyperbolic sine function (asinh) and are reported in real, spatially delfated, 2012-13 Tanzanian Shilling (TSh). Summary statistics are reported without survey weights.

4.5 Methods

We estimate the relationship among initial asset holdings, expected returns, risk, and downside risk among Tanzanian households as follows. First, we use cluster analysis to identify the set of asset portfolios available in the data. We then estimate the portfolio-specific moments of the conditional consumption distribution via regression of consumption on a quadratic function of the assets, with household, time, and portfolio fixed effects as well as controls for time varying characteristics. These regressions additionally allow us to estimate the contribution of each asset to the mean, variance, and skewness of the conditional consumption distribution within each portfolio.

Next we estimate the relationship among a household's initial asset holdings and the riskiness and returns of its asset portfolio using fractional polynomials and a generalized additive model. In addition, making the assumption that the household utility of consumption is CRRA, we calculate the household level risk premium associated with each portfolio and estimate the relationship among initial asset holdings, returns, and the risk premium for each portfolio using a generalized additive model.

In the following subsections, we detail the methods for each of part of this analysis.

4.5.1 Cluster analysis

Unsupervised learning methods such as cluster analysis allow for the classification or grouping of data based on similarity or some other objective function. Here we use *k*-medoids cluster analysis, implemented via the partitioning around medoids (PAM) algorithm developed by Kaufman & Rousseeuw (1990),

to identify the set of asset portfolios available in the data. PAM has several advantages over other approaches to clustering, such as k-means; in particular, because within-cluster dissimilarity is calculated using Manhattan distance (or L^1 distance), PAM is more robust to outliers than is k-means, which minimizes the sum of squared distance.

The PAM algorithm operates by stepwise selection of an initial set of k^* medoids, i_k, from the set of observations, $i, i = 1, ...N$, so as to minimize d_{ii_k}, the distance between cluster medoid and the other members of cluster $C(i)$. The algorithim then iteratively replaces the initial medoids so as to minimize the sum of within cluster dissimilarities, as shown in the objective function in Equation 4.21 (Hastie et al. 2009),

$$min_{C,\{i_k\}_1^{k^*}} \Sigma_1^{k^*} \Sigma_{C(i)=k} d_{ii_k} \tag{4.21}$$

To select the appropriate k^*, we rely on the silhouette method (Rousseeuw 1987, Kaufman & Rousseeuw 1990). The silhouette method entails calculating the silhouette width, $s(i)$, for each observation, i, over a set of possible ks, where $k = 1, ...K$. The k with the largest average silhouette width, $\frac{1}{N}\Sigma_{i=1}^{N} s(i)$ offers the best number of clusters, given the clustering method, for compact and well separated groupings (Rousseeuw 1987), and is therefore selected as k^*.

The silhouette width is the ratio of the difference between the average within-cluster dissimilarity and the minimum average between-cluster dissimilarity to whichever dissimilarity (within or between) measure is greater. The silhouette width for a single observation is calculated as

$$s(i) = \frac{b(i) - a(i)}{max[a(i), b(i)]} \tag{4.22}$$

Where $a(i)$ is the average dissimilarity between observation i and all the other members of the cluster to which i has been assigned, and $b(i) = min(d(i, C))$, where $d(i, C)$ is the average dissimilarity between i and the members of another cluster, C, to which i has not been assigned (Rousseeuw 1987).

The intuition behind the silhouette method for identification of the appropriate number of clusters is that it offers a summary of how well each observation has been clustered by finding the distance between the fit of the observation's own cluster and that of the observation's second best choice cluster (i.e., the smallest average observation-to-other-cluster distance, $min(d(i, C))$, is the second best choice cluster for observation i) and then considering the distance between these fits as a share of whichever offers the poorer fit – the cluster's own fit or the fit of the next best cluster.

The silhouette width, $s(i)$, grows smaller as the own cluster fit grows worse and larger as it grows better. Likwise, the average silhouette width, $\frac{1}{N}\Sigma_{i=1}^{N} s(i)$, will range between -1 and 1, with values close to 1 indicating well defined and separated clusters and those close to -1 indicating misclustered observations.

So as to identify the set of asset portfolios in the data, the cluster analysis is conducted using only household productive assets data. Note that we are focused on physical assets only, treating human capital assets, such as education, as fixed factors of production in this analysis. This exclusion of human capital assets from the analysis is due the length of the panel data with which our empirical analysis is conducted: the (at most) five year time horizon is too short for payoffs from human capital investments to be realized. Assuming adequate data, inclusion of human capital assets would be a valuable extension of the

present work.

A total of 41 assets are included in the estimation. These assets include land, livestock, physical infrastructure such as homes and other buildings, financial assets such as savings and investments, and transportation, communication, and enterpreunuerial assets such as vehicles, cell phones, and sewing machines. The summary statistics for these assets are shown in Table 4.1.

4.5.2 Conditional moments

To identify the conditional moments of the consumption distribution, we follow Antle (1983, 1987). The uncertainty of the (consumption-based) returns to the household's productive assets can be characterized via econometric estimation of the moments of the relationship between consumption and asset holdings. Let the relationship $\boldsymbol{C_i} \leq \boldsymbol{g}(\boldsymbol{x_i}, \boldsymbol{z_i}, \boldsymbol{v_i}) - \boldsymbol{p_z'} \boldsymbol{z_i} + \boldsymbol{\eta_i}(\boldsymbol{x_i}) - \boldsymbol{I_i}$ be estimated by the reduced form expression, $C_i = f_1(\boldsymbol{x_i}, \boldsymbol{\beta_1}) + u_{i1}$, where $f_1(\boldsymbol{x_i}, \boldsymbol{\beta_1}) = \hat{E}[C_i | f(\boldsymbol{x_i})]$ and $E(u_{1i}) = 0$.

The advantage of this reduced form approximation of the budget constraint is that it is explicitly non-separable and it implicitly captures the fact that access to credit, as a function of asset holdings, can smooth consumption. The disadvantage is that it does not control for time varying inputs, $\boldsymbol{z_i}$, which are endogenous to the relationship between consumption and asset holdings but are only partially observed in our data.

Let μ_{mi} represent the m^{th} moment of household i's conditional consumption distribution; *e.g.*, $\mu_{1i} = \hat{E}[C_i | f(\boldsymbol{x_i})]$. Then μ_{i2} and μ_{i3} can be estimated as

$$\mu_{2i} = (\hat{u_{1i}})^2 = f_2(\boldsymbol{x_i}, \boldsymbol{\beta_2}) + u_{i2}, E(u_{i2}) = 0 \tag{4.23}$$

$$\mu_{3i} = (\hat{u_{1i}})^3 = f_3(\boldsymbol{x_i}, \boldsymbol{\beta_3}) + u_{i3}, E(u_{i3}) = 0 \tag{4.24}$$

We exploit the panel nature of the dataset to control for household and time fixed effects when estimating the conditional moments. To do so, we estimate a quadratic function of the productive assets with interaction terms for the cluster-identified portfolios, as specified in Equation 4.25, which presents the regression equation for the estimation of the first conditional moment, μ_{1i}. For estimation of subsequent moments, the regressand in Equation 4.25 is replaced with $\hat{u_{1i}}^m$, where $m = 2, 3$. Due to the way in which the LSMS-ISA data were collected (consumption data are collected via recall while data on asset holdings are collected via an account of present asset holdings), so as to ensure that we are conditioning consumption on the stock of household asset holdings, we use next period's consumption as the dependent variable (equivalently we could use last period's assets).

$$\begin{aligned} C_{it+1} = \Sigma_a^A \beta_a X_{ita} + \frac{1}{2}\Sigma_a^A \beta_{aa} X_{ita}^2 + \boldsymbol{\beta}_h \boldsymbol{h}_{it} \\ \Sigma_k^{K-1}\Sigma_a^A \beta_{aP_k} X_{ita} P_k + \frac{1}{2}\Sigma_k^{K-1}\Sigma_a^A \beta_{aaP_k} X_{ita}^2 P_k + \Sigma_k^{K-1} \boldsymbol{\beta}_{hP_k} \boldsymbol{h}_{it} P_k \\ + \alpha_i + \psi_t + \epsilon_{it} \end{aligned} \tag{4.25}$$

In Equation 4.25, i indexes household, t indexes time, a indexes assets with $a = 1, ...A$, k indexes portfolios with $k = 1, ...K$, and $\boldsymbol{h}_{it}$ is a vector of time varying household characteristics. As above, the assets included in the analysis are the 41 land, livestock, infrastructure, financial, transportation, communication, and enterpreunuerial assets observed in the data. Equation 4.25 is estimated with heteroskedasticity and cluster robust standard errors.

The estimated conditional moments allow us to generate cumulative distribution functions of the conditional consumption distribution for each portfolio so as to assess the relative stochastic dominance of each portolio. We fit the es-

timated moments to log normal and gamma distributions, where the shape and scale parameters are ν and σ, and κ and θ, respectively (Equation 4.26). We use a simple optimization approach to identify the distribution-specific shape and scale parameters that best describe the estimated moments, minimizing the objective function in 4.26 subject to the distribution-specific moments constraints. We use calculated shape and scale parameters (calculated from the first two estimated moments) as starting points to initiate the optimization.

$$MinQ \begin{Bmatrix} Lognormal: & \nu, & \sigma \\ Gamma: & \kappa, & \theta \end{Bmatrix} = (\hat{\mu_1} - \mu_1)^2 + (\hat{\mu_2} - \mu_2)^2 + (\hat{\mu_3} - \mu_3)^2 \tag{4.26}$$

s.t the following distribution-specific moment constraints,

$$\begin{cases} \mu_m & Lognormal & Gamma \\ \mu_1 = & exp(\nu + \sigma^2/2) & \kappa\theta \\ \mu_2 = & exp(\sigma^2 - 1)exp(2\nu + \sigma^2) & \kappa\theta^2 \\ \mu_3 = & exp(\sigma^2) + 2)\sqrt{(exp(\sigma^2) - 1)} & 2/\sqrt{\kappa} \end{cases}$$

Drawing from the distribution produced by the shape and scale parameters identified by the optimization routine, we then produce CDFs for each portfolio within each distribution. Finally, the relationships among the estimated conditional moments are estimated via fractional polynomials.

4.5.3 Risk premium

The estimated conditional second and third moments are used to calculate the portfolio specific risk premia, as derived in Equation 4.20, as $\mu_{2i} = V(C_i)$ and

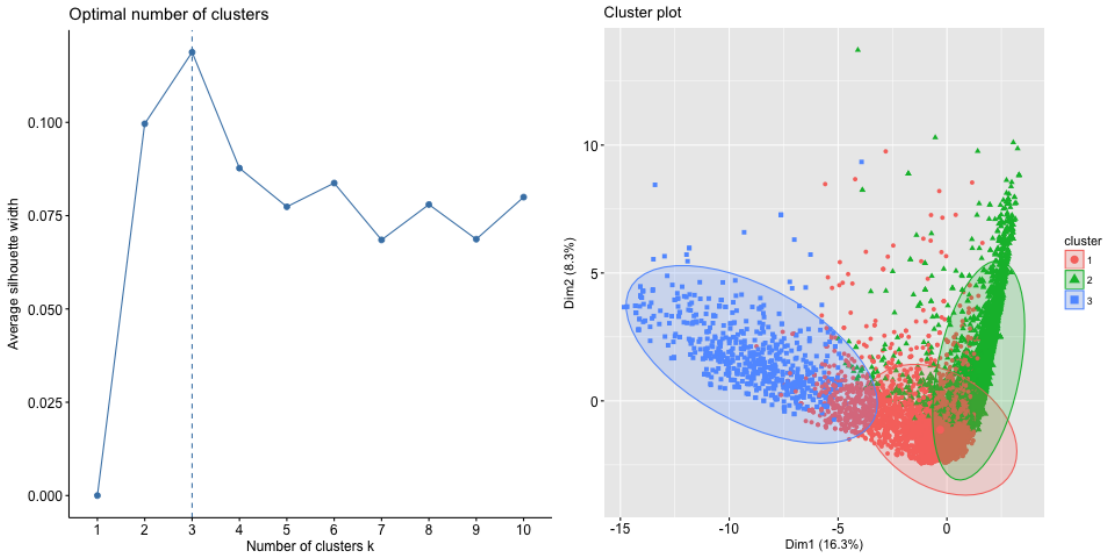
$\mu_{3i} = S(C_i)$). So as to parameterize AP and DS in Equation 4.20, we follow Di Falco & Chavas (2006) in assuming a CRRA utility function of the form $U(C) = -C^{1-r}$ with the relative risk aversion paramete $r = 2$, making $AP = 2/C$ and $DS = -6/C^2$.

Finally, the relationships between the risk premium and initial wealth is estimated via fractional polynomials. Initial wealth in this analysis is the total value of the household's physical asset holdings in the first period under analysis (2008-09, depending on the date on which the household was first interviewed). As above, all values are in 2012-13 real terms and have been spatially deflated as well as transformed via the asinh transformation.

4.6 Results

Three asset portfolios were identifed in the data. Average silhouette width per k and a visualization of the final cluster grouping are shown in Figure 4.1. The optimal average silhouette width at $k = 3$ of 0.119 suggests that the identified clusters are not strongly separated. The cluster plot in Figure 4.1 shows the projection of the data on to its first two principle components, which account for 16.3 and 8.3 percent of the total variation in the data, respectively. The ellipses provide confidence intervals on the clusters, containing 95 percent of each cluster's observations, assuming a multivariate normal distribution. The cluster plot shows three distinct but overlapping groups, which is what we would expect to find in the setting under analysis, as asset holdings will not be mutually exclusive among portfolios. Rather, portfolio cluster assignment is a matter of extent of asset holdings, a fact that becomes clearer in Table 4.3.

Figure 4.1: Cluster assignment



(a) Silhouette method

(b) Clusters

A comparison of household characteristics and asset holdings by assigned portfolio cluster for those households remaining in the same cluster across all waves is shown in Table 4.3. Figure 4.2 shows the distribution of the value of initial (2008-09) asset holdings as well as consumption for each of the identified asset portfolios. Despite the overlap in portfolio cluster assignment, portfolio means are generally statistically significantly different from one another and suggest distinct asset-based income generating activities in Tanzania. In particular, households holding asset portfolio 2 (46 % of the observations in the sample) are more urban and more likely to have migrated, have smaller household sizes, younger heads of household, and better educated heads; overall, they own less land, fewer livestock, and less farming equipment. Asset portfolio 2 is largely composed of business, communication, transportation, and entreprenuerial assets.

While households with portfolios 1 (52 % of the observations in the sample)

and 3 (2% of the observations in the sample) are both rural and farm based, those with portfolio 3 have significantly larger families, older heads of household, and a greater value of all land, livestock, and farming assets than those households with asset portfolio 1. Although both rural, those households holding asset portfolio 1 hold more fishing assets than do those holding portfolio 3.

Table 4.3: Comparison of asset portfolios

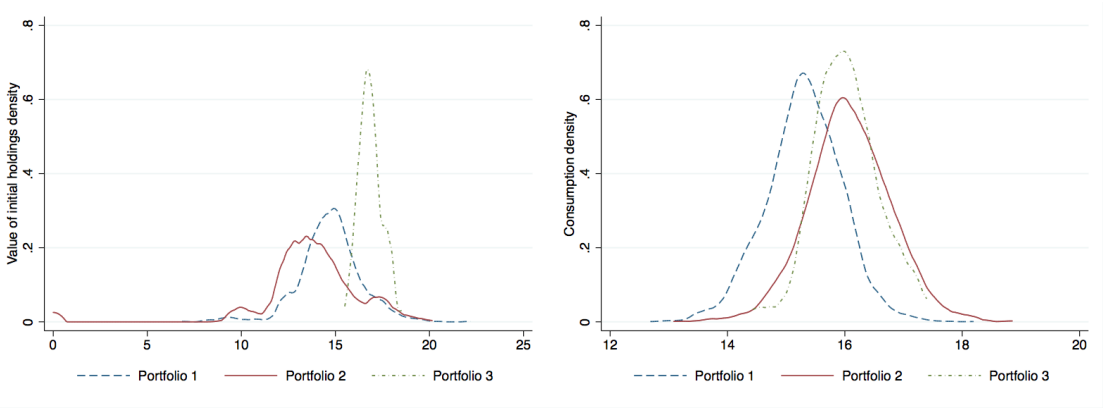
	Asset Portfolio 1 N=3,972 obs (or 1,324 hhs)	Asset Portfolio 2 N=3,553 obs (or 1,185 hhs)	Asset Portfolio 3 N=178 obs (or 59 hhs)	Wald test
rural	0.95	0.30	0.98	***
household size	5.08	4.67	8.98	***
head of household female	0.26	0.27	0.12	***
head age	49.30	42.16	49.18	***
head married	0.74	0.65	0.88	***
head migrated to this area	0.29	0.76	0.57	***
head_primaryed	0.42	0.49	0.47	***
head_secondaryed	0.02	0.23	0.00	***
head_university	0.00	0.02	0.00	NA
plotval	13.82	1.15	14.97	***
bullval	0.67	0.06	12.82	***
cowval	1.52	0.19	14.77	***
steerval	0.32	0.00	11.54	***
hefval	0.40	0.07	9.20	***
mcalfval	0.42	0.07	10.75	***
fcalfval	0.59	0.06	10.17	***
goatval	2.85	0.16	12.64	***
sheepval	0.44	0.02	8.40	***
pigval	0.89	0.04	1.32	***
chickenval	6.52	0.72	11.36	***
otlstkval	1.00	0.19	5.21	***
handhoeval	9.80	1.40	10.29	***
handsprayval	0.46	0.11	2.86	***
oxploughval	0.37	0.00	11.96	***
oxseedval	0.27	0.01	7.68	***

oxcartval	0.04	0.00	1.29	***
tractorval	0.02	0.01	2.74	***
tractploughval	0.01	0.02	0.00	
tractharrowval	0.00	0.02	0.00	*
thresherval	0.03	0.00	0.00	**
watercanval	0.09	0.02	0.30	***
farmbldgsval	0.51	0.07	3.25	***
gericanval	0.44	0.10	1.48	***
fishnetval	0.15	0.11	0.00	***
fishlineval	0.08	0.08	0.00	***
fishboatval	0.15	0.15	0.00	
fishmotorval	0.02	0.03	0.00	
bus_capitalval	2.70	5.36	2.41	***
bus_stockval	0.53	0.83	0.47	***
bus_goodsval	1.36	3.38	1.66	***
phone_landval	0.05	0.34	0.14	***
phone_mobileval	3.54	11.03	9.29	***
fridge_freezeval	0.12	4.55	0.24	***
sewmachval	0.43	2.42	0.50	***
computerval	0.09	1.23	0.09	***
stove_geval	0.13	2.04	0.26	***
stove_otherval	1.52	8.12	1.97	***
carval	0.09	1.33	0.00	***
motorcycleval	0.28	0.92	1.08	***
bicycleval	4.86	3.29	10.02	***

All monetary values have been transformed via the inverse hyperbolic sine function (asinh) and are reported in real, spatially delfated, 2012-13 Tanzanian Shilling (TSh). Standard deviations have been surpressed to curtail table length but are available on request. Number of households per cluster is average across the three time periods.

A comparison (Figure 4.2) of the value of initial (2008-09) asset holdings and consumption (all periods pooled) by portfolio reveals that households with asset portfolio 3 enjoy higher consumption throughout the panel and begin with greater asset holdings at the start of the panel. Households with portfolio 2 begin with smaller asset holdings in the first period but enjoy a level of consumption similar to those with portfolio 3. The high level of consumption despite the low value of productive assets exhibited by the households with portfolio 2 is due to the fact that some of the more valuable asset holdings of portfolio 2 households are their human capital assets (such as education, entreprenuerial skill, etc), which are not accounted for in the total asset valuation in this study.

Figure 4.2: Initial asset holdings and consumption densities by portfolio, values in asinh 2012-13 TSh



(a) Initial (2008-09) asset holdings (b) Consumption

The diagonal of the transition matrix in Table 4.5 shows that most households stay in their original portfolio cluster from year to year. However, the off-diagonals tell us about the direction of mobility throughout the panel. Those households leaving the low-asset, rural, farm-based portfolio cluster are more likely to go to the urban off-farm cluster (19.33 %) than to the high-asset, rural, farm-based cluster (3.58 %). Likewise, those leaving the urban off-farm cluster

are more likely to go to the low-asset, rural, farm-based portfolio cluster (13.06 %) than to the high-asset, rural, farm-based cluster (.80 %). Finally, the smallest portfolio cluster has the greatest mobility (and, as we have seen, the greatest initial welfare): % of those in the high-asset, rural, farm-based portfolio cluster will transition to the low-asset, rural, farm-based portfolio cluster while 12.47% will transition to the urban off-farm cluster. The high rate of transition from the high-asset farm portfolio to the low-asset farm portfolio suggests that it may not be easy to maintain the high-asset farm portfolio.

Due to the non-trivial mobility among portfolios, the analyses that follow are done using only those households – 60% of the total sample – that remain in the same portfolio cluster over the duration of the panel. Restricting the analysis to these households introduces selection bias in to the findings, as estimated returns, risk, and downside risk will apply only to those households that decided to stay within their original cluster over the five year time span either because they were successful in that portfolio cluster or because they didn't have the resources to transition out of that cluster. At this point it is not possible to identify the direction in which this selection issue will bias our results. Because the descriptive task of this paper is to estimate the within-portfolio risks and returns (rather than the risks and returns associated with portfolio changes, an objective of future work), we forge ahead with this restriction.

Table 4.5: Portfolio transition matrix

		Time t+1		
		Portfolio 1	Portfolio 2	Portfolio 3
Time t	Portfolio 1	77.09	19.33	3.58
	Portfolio 2	13.06	86.14	0.8
	Portfolio 3	29.1	12.47	58.44

The estimated conditional moments, calculated risk premia, and value of initial (2008-09) asset holdings are summarized in Table 4.6 for each portfolio, where pairwise statistical difference of means is assessed via a Tukey-Kramer test (common letters indicate results are statistically different at the 95% confidence level). On average, households holding portfolio 3 have greater expected consumption and face greater risk than those households holding portfolios 1 or 2. However, portfolio 3 also has the greatest positive skew, suggesting that this collection of assets reduces exposure to downside risk. Recall that the households holding portfolio 3 make up a very small share of the sample (2%) and that the portfolio is composed of large land, livestock, and other farming assets. We would expect the consumption of such households to be vulnerable to the long tail events, such as drought, witnessed in this dataset. However, several factors, each of which merit further investigation, may be contributing to the high positive skew of this portfolio: the diversity of the portfolio itself, the ability of portfolio 3 holders to access credit and/or insurance due to the value of their asset holdings, consumption smoothing.

Holders of portfolio 2 have conditional consumption that is slightly higher

but statistically indistinguishable from the holders of portfolio 1. In light of the high observed consumption of portfolio 2 households (see Figure 4.2) and the comparatively high education of portfolio 2 households (see Table 4.3), it is clear that the low conditional consumption estimates for portfolio 2 are due to the fact that these estimates do not properly account for expected returns to human capital assets. Likewise, the aggregate value of the portfolio 2 initial asset holdings is the lowest among all portfolios because portfolio 2 assets are largely held in terms of human, and not physical, capital. Holders of portfolio 2 face greater variance and greater downside risk than do holders of portfolio 1. Finally, portfolio 1, composing 52 % of the sample, appears to be the low risk, low return portfolio.

Across the three portfolios, we see the risk premium rising in expected consumption, which is what we would expect.[4].

[4]Di Falco & Chavas (2006) find that the estimation of risk premia via local approximation, as we have done, provides an underestimate of the true risk premia in the setting of their analysis (Sicilian wheat growers). Further analysis is needed to ascertain whether the risk premia we report are underestimates.

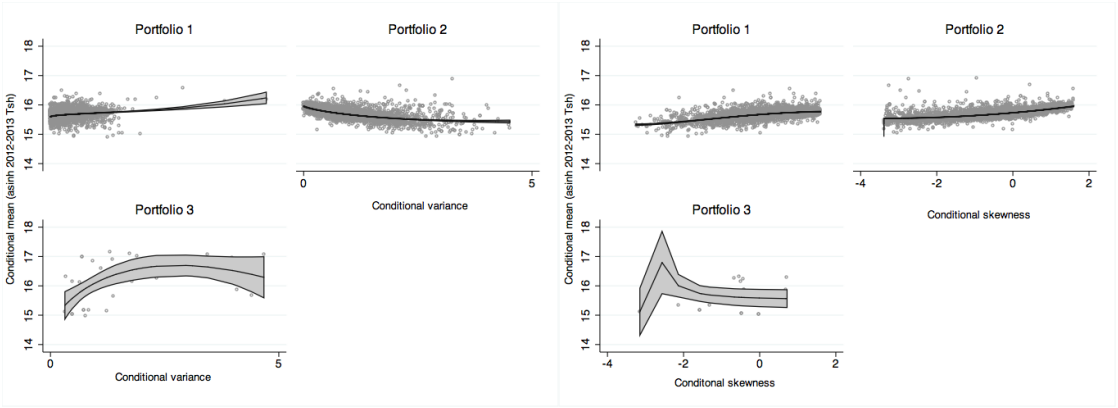
Table 4.6: Conditional moments, calculated risk premia, and value of initial (2008-09) asset holdings by portfolio

	Portfolio one (1,324 hhs)			Portfolio 2 (1,185 hhs)			Portfolio 3 (59 hhs)		
	Mean	Std dev	Median	Mean	Std dev	Median	Mean	Std dev	Median
Mean, μ_1	15.66(a)	0.17	15.68	15.68(b)	0.24	15.68	16.72(a)(b)	2.33	16.78
Variance, μ_2	0.36(a)	0.50	0.36	1.10(a)	1.02	0.94	2.71(a)	9.37	1.90
Skewness, μ_3	0.10(a)	1.39	0.16	-1.01(a)	3.46	-0.56	8.81(a)	60.12	2.88
Risk premium, R	0.02(a)	0.03	0.02	0.07(a)	0.07	0.06	0.14(a)	0.73	0.07
Initial holdings	14.72(a)	1.67	14.73	13.67(a)	2.91	13.79	16.86(a)	0.61	16.83

Common letters indicate results are statistically different at the 95% confidence level using a Tukey-Kramer test.

Switching from an across portfolio to a within portfolio analysis, we turn to non-parametric estimation of the relationship between the conditional mean and the conditional variance, skewness, and risk premium within each portfolio displayed in Figure 4.3. Figure 4.3 displays the relationship between the conditional mean and conditional variance estimates within each portfolio. We see conditional variance (risk) rising slightly in conditional mean (expected returns) in the case of portfolios 1 and 3. In the case of portfolio 2, we see a slight decline which is likely due to poor accounting for human capital assets in the estimation. Figure 4.3 shows the relationship between the conditional mean and conditional skewness by portfolio. Positive skewness appears to increase slightly with expected returns in the case of portfolio 1; there is no clear relationship in the case of portfolio 3. Finally, there appears to be no clear within portfolio relationship between the risk premium and expected returns or between the risk premium and the value of initial asset holdings (Figure 4.4).

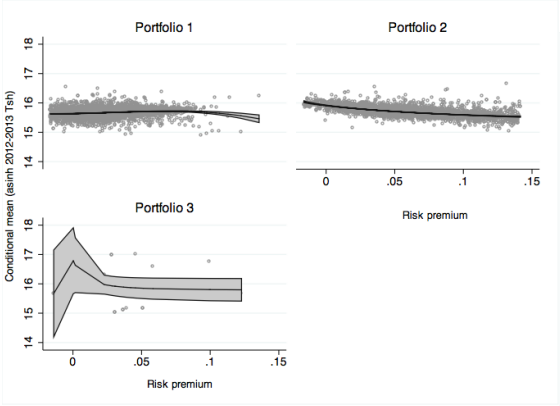
Figure 4.3: Conditional mean, variance, and skewness, fractional polynomial estimate



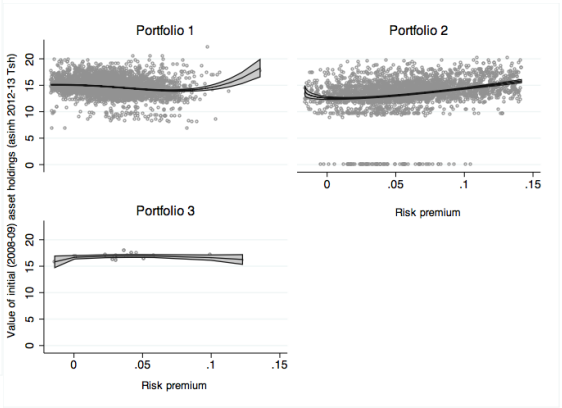
(a) Conditional mean and conditional variance

(b) Conditional mean and conditional skewness

Figure 4.4: Conditional mean, risk premium, and initial holdings by portfolio, fractional polynomial estimate



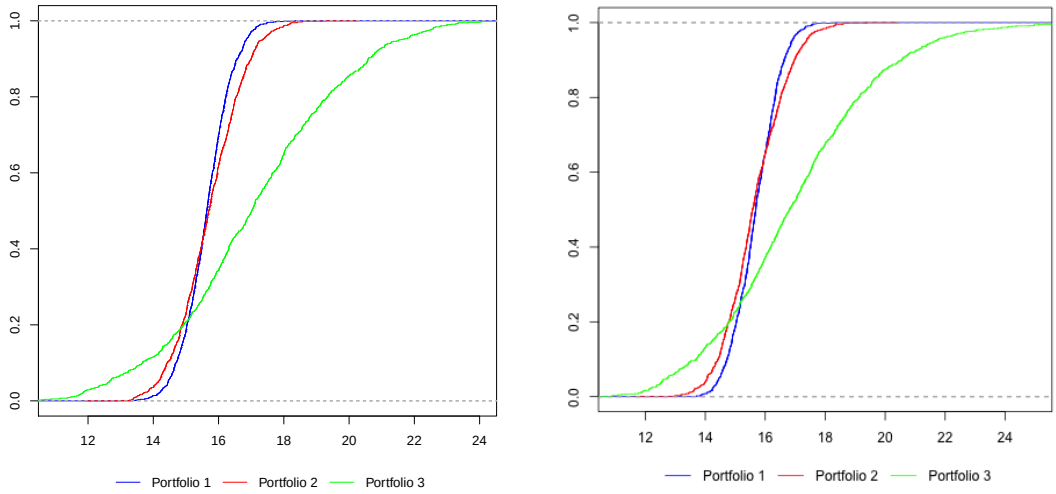
(a) Conditional mean and risk premium



(b) Risk premium and initial holdings

Finally, the log normal and gamma distributions for each portfolio, fit via the optimization specified Equation 4.26, are shown in Figure 4.5. No one portfolio first order stochastically dominates the others. However, a comparison of the areas under the CDFs for both the gamma and log normal distributions shows that portfolio 3 (gamma area of 8.87; log normal area of 11.22) second order stochastically dominates portfolio 2 (3.40; 3.74) which second order stochastically dominates portfolio 1 (2.58; 2.47).

Figure 4.5: CDFs of distributions fit on estimated moments of the conditional consumption distribution



(a) Gamma

(b) Log normal

4.7 Conclusion

Do initial asset holdings determine the riskiness (both upside and downside risk) and expected returns of a household's asset portfolio? We have not satisfactorly answered this question here because human capital investments, which play an important role in the off-farm asset portolio, are not properly accounted for in the aggregated initial asset holdings values or in the estimates of risk and returns. Moreover, we are able to consider only a snapshot in time – a, at most, five year time span – and do not account for long run welfare dynamics. These limitations aside, across agricultural portfolios (those holding portfolios 1 and 3), we do find evidence consistent with a pattern in which households with greater initial asset holdings also hold a riskier portfolios and enjoy greater returns to their assets. However, we do not find clear within-portfolio relation-

ships between initial asset holdings and risk.

APPENDIX A

APPENDIX FOR CHAPTER 1

Table A.1: Comparison of IRIS, Cross-Validation, and Stochastic Ensemble Accuracy Results

Country	Source	Estimation	TA	PA	UC	LE	BPAC
Bolivia (2005 EH)	IRIS	1) QR (0.42)-In sample (half)	83.65	67.18	32.82	33.29	66.71
		2) QR (0.42)[a]	81.88	57.58	42.42	34.3	49.33
		3) Std. err.	1.02	2.61	2.61	3.6	6.11
		4) QR (0.42)[b]	[79.78, 83.68]	[52.51, 62.65]	[37.35, 47.49]	[27.6, 41.66]	[36.73, 60.48]
	Rep	5) QR (0.42) rep.-In sample (full)	82.45	60.69	39.30	33.71	55.10
	Cross-validation	6) QR (0.38)[a]	81.76	63.37	36.63	41.61	57.94
		7) Std. err.	0.86	2.25	2.25	3.39	3.04
		8) QR (0.38)[b]	[80.10, 83.32]	[58.84, 67.88]	[32.12,41.15]	[35.15, 48.24]	[50.61, 63.44]
	Stochastic ensemble	9) QRF (0.41)[a]	80.17	55.33	44.67	40.44	50.18
		10) Std. err.	0.95	2.44	2.44	3.78	5.14
		11) QRF (0.41)[b]	[78.25, 82.03]	[50.51, 60.12]	[39.88, 49.49]	[33.59, 48.11]	[39.26, 58.57]
East Timor (2001 TLSS)	IRIS	1) Probit-In sample (full)	77.14	75.08	24.92	26.20	73.79
		2) Probit[a,c]	75.56	69.32	30.68	28.71	65.56
		3) Std. err.	1.52	2.56	2.56	3.38	4.33
		4) Probit[b,c]	[72.63, 78.50]	[64.38, 74.33]	[25.67, 35.62]	[22.40, 35.58]	[55.57, 72.08]
	Rep	5) Probit rep.-In sample (full)	77.16	71.41	28.59	27.63	70.45
	Cross-validation	6) QR (0.46)[a]	76.19	73.01	26.99	30.97	68.34
		7) Std. err.	1.42	2.32	2.32	3.33	2.96
		8) QR (0.46)[b]	[73.43, 77.51]	[73.43, 77.51]	[22.49, 31.46]	[24.35, 37.99]	[61.84, 73.38]
	Stochastic ensemble	9) QRF (0.47)[a]	75.05	71.75	28.25	32.51	66.85
		10) Std. err.	1.50	2.42	2.42	3.55	3.17
		11) QRF (0.47)[b]	[72.19, 78.03]	[67.12, 76.72]	[23.28, 32.88]	[25.94, 39.90]	[59.77, 72.22]
Malawi (2004/5 IHS2)	IRIS	1) QR (0.57)-In sample (half)	80.15	84.12	15.88	16.43	83.57
		2) QR (0.57)[a]	79.69	83.47	16.53	17.06	82.56
		3) Std. err.	0.55	0.65	0.65	0.76	0.74
		4) QR (0.57)[b]	[78.6, 80.84]	[82.2, 84.77]	15.23, 17.79]	[15.53, 18.56]	[80.95, 83.82]
	Rep	5) QR (0.57) rep.-In sample (full)	80.82	84.88	15.11	14.39	84.17
	Cross-validation	6) QR (0.55)[a]	80.79	85.72	14.28	15.07	84.75
		7) Std. err.	0.52	0.55	0.55	0.69	0.64
		8) QR (0.55)[b]	[79.79, 81.84]	[84.68, 86.86]	[13.14, 15.32]	[13.73, 16.38]	[83.42, 85.86]
	Stochastic ensemble	9) QRF (0.57)[a]	80.10	85.53	14.47	15.93	83.99
		10) Std. err.	0.58	0.67	0.67	0.75	0.73
		11) QRF (0.57)[b]	[78.93, 81.19]	[84.22, 86.80]	[13.20, 15.78]	[14.51, 17.47]	[82.46, 85.25]

Note: QR(#) = quantile regression estimated at the #th quantile; QRF(#) = quantile regression forest estimated at the #th quantile.

[a] *Bootstrapped 1,000 times, with replacement, mean reported.*

[b] *Bootstrapped 1,000 times, with replacement; 95% bootstrap confidence interval reported, where lower bound is 2.5% and upper bound is 97.5%.*

[c] *Because these bootstrapped estimates were not available in materials made public by IRIS, the estimates reported here were calculated by the authors based on the replication sample and model.*

Source: Authors and IRIS centers estimates using data and procedures detailed in the text.

Table A.2: Comparison of IRIS, Cross-Validation, and Stochastic Ensemble Accuracy Results under Halved and Doubled Poverty Lines

Data		Estimation	TA	PA	UC	LE	BPAC	Poverty line	Poverty rate (%)
Bolivia (2005 EH)	IRIS	2) QR (0.22)[a]	94.55	41.65	58.35	65.89	30.07	Half	4.92
		3) Std. err.	0.54	5.45	5.45	11.72	9.53		
		4) QR (0.22)[b]	[93.44, 95.56]	[30.72, 52.63]	[47.37, 69.28]	[45.54, 92.19]	[6.73, 44.38]		
	Cross-validation	6) QR(0.24)[a]	94.53	41.20	58.80	66.31	29.65		
		7) Std. err.	0.56	5.44	5.44	11.85	9.50		
		8) QR (0.22\4)[b]	[93.39, 95.61]	[31.14, 52.47]	[47.53, 68.86]	[44.93, 91.90]	[7.90, 44.58]		
	Stochastic ensemble	9) QRF (0.26)[a]	94.39	43.65	56.35	71.03	26.94		
		10) Std. err.	0.56	5.71	5.71	13.25	11.70		
		11) QRF (0.26)[b]	[93.24, 95.43]	[32.00, 55.00]	[45.00, 68.00]	[47.66, 100.00]	[0.00, 45.27]		
	IRIS	2) QR (0.54)[a]	78.90	82.64	17.36	16.65	81.10	Double	62.26
		3) Std. err.	0.94	1.12	1.12	1.35	1.86		
		4) QR (0.54)[b]	[77.11, 84.79]	[80.40, 84.79]	[15.21, 19.60]	[14.14, 19.17]	[76.61, 83.88]		
	Cross-validation	6) QR (0.52)[a]	79.01	83.60	16.40	17.45	81.92		
		7) Std. err.	0.94	1.11	1.11	1.39	1.31		
		8) QR (0.52)[b]	[77.17, 80.83]	[81.38, 85.67]	[14.33, 18.62]	[14.84, 20.09]	[79.05, 84.14]		
	Stochastic ensemble	9) QRF (0.54)[a]	77.89	83.66	16.34	19.32	80.62		
		10) Std. err.	1.02	1.14	1.14	1.38	1.33		
		11) QRF (0.54)[b]	[75.99, 79.79]	[81.41, 85.77]	[14.22, 18.59]	[16.68, 21.99]	[78.01, 83.10]		
East Timor (2001 TLSS)	IRIS	2) Probit[a]	90.82	28.51	71.50	23.37	-19.62	Half	10.65
		3) Std. err.	1.03	5.30	5.30	6.03	12.06		
		4) Probit[b]	[88.69, 92.75]	[18.79, 39.13]	[60.87, 81.21]	[13.37, 36.72]	[-42.95, 3.82]		
		2) QR (0.27)[a]	89.02	49.26	50.74	62.58	35.45		
		3) Std. err.	1.11	5.91	5.91	10.95	9.64		
		4) QR (0.27)[b]	[86.81, 91.25]	[38.03, 61.41]	[38.59, 61.97]	[43.63, 85.61]	[14.04, 51.27]		
	Cross-validation	6) QR (0.28)[a]	88.76	46.09	53.91	61.67	35.04		
		7) Std. err.	1.05	5.44	5.43	10.50	8.71		
		8) QR (0.28)[b]	[86.71, 90.81]	[35.29, 56.88]	[43.12, 64.71]	[42.44, 83.23]	[16.26, 48.94]		
	Stochastic	9) QRF (0.28)[a]	89.34	39.20	60.80	48.97	23.91		
		10) Std. err.	1.20	5.80	5.80	11.73	13.89		
		11) QRF (0.28)[b]	[86.99, 91.70]	[27.77, 50.55]	[49.45, 72.23]	[29.46, 74.37]	[-5.68, 45.75]		
	IRIS	2) Probit[a]	84.15	93.04	6.96	13.72	86.28	Double	80.20
		3) Std. err.	1.08	0.67	0.67	1.37	1.37		
		4) Probit[b]	[82.75, 85.75]	[92.20, 94.07]	[5.93, 7.80]	[12.12, 15.51]	[84.49, 87.88]		
		2) QR (0.60)[a]	83.34	89.27	10.73	11.16	87.72		
		3) Std. err.	1.33	1.27	1.27	1.43	1.61		
		4) QR (0.60)[b]	[80.75, 85.75]	[86.70, 91.68]	[8.32, 13.30]	[8.33, 14.04]	[83.82, 90.33]		
	Cross-validation	6) QR (0.57)[a]	83.86	91.18	8.82	12.40	87.58		
		7) Std. err.	1.21	1.06	1.06	1.44	1.41		
		8) QR (0.57)[b]	[81.61, 86.10]	[89.16, 93.30]	[6.70, 10.84]	[9.65, 15.33]	[84.67, 90.17]		
	Stochastic	9) QRF (0.58)[a]	82.63	89.96	11.04	11.78	87.28		
		10) Std. err.	1.28	1.26	1.26	1.44	1.46		
		11) QRF (0.58)[b]	[80.04, 85.09]	[86.52, 91.29]	[8.71, 13.48]	[8.99, 14.59]	[84.00, 89.58]		

Malawi (2004/5 IHS2)	IRIS	2) QR (0.41)[a]	79.67	58.19	41.81	45.48	54.25		
		3) Std. err.	0.56	1.15	1.48	2.42	2.17		
		4) QR (0.41)[b]	[78.58, 80.64]	[55.38, 61.15]	[38.85, 44.62]	[40.77, 50.31]	[49.63, 58.19]		
	Cross-validation	6) QR (0.40)[a]	79.59	56.97	40.02	47.48	52.52		
		7) Std. err.	0.56	1.36	1.36	2.40	2.40	Half	23.43
		8) QR (0.40)[b]	[78.54, 80.67]	[57.21, 62.88]	[37.11, 42.79]	[43.06, 52.16]	[47.84, 56.91]		
	Stochastic	9) QRF (0.42)[a]	79.24	56.04	43.96	45.15	53.43		
		10) Std. err.	0.58	1.56	1.56	2.41	2.13		
		11) QRF (0.42)[b]	[78.09, 80.32]	[53.04, 59.10]	[40.91, 46.96]	[40.63, 49.76]	[48.74, 57.10]		
	IRIS	2) QR (0.66)[a]	92.12	95.95	4.05	4.65	95.32		
		3) Std. err.	0.38	0.29	0.29	0.33	0.31		
		4) QR (0.66)[b]	[91.37, 92.86]	[95.36, 96.51]	[3.50, 4.64]	[4.05, 5.32]	[94.67, 95.88]		
	Cross-validation	6) QR (0.64)[a]	92.34	96.26	3.74	4.72	95.28		
		7) Std. err.	0.35	0.27	0.27	0.31	0.30	Double	90.65
		8) QR (0.64)[b]	[91.63, 93.01]	[95.75, 96.76]	[3.24, 4.24]	[4.14, 5.36]	[94.64, 95.84]		
	Stochastic	9) QRF (0.66)[a]	92.11	95.76	4.23	4.48	95.36		
		10) Std. err.	0.37	0.30	0.30	0.31	0.33		
		11) QRF (0.66)[b]	[91.33, 92.81]	[95.17, 96.33]	[3.82, 5.0]	[3.89, 5.11]	[94.62, 95.91]		

Note: QR(#) = quantile regression estimated at the #th quantile; QRF(#) = quantile regression forest estimated at the #th quantile.

[a] Bootstrapped 1,000 times, with replacement, mean reported.

[b] Bootstrapped 1,000 times, with replacement; 95% bootstrap confidence interval reported, where lower bound is 2.5% and upper bound is 97.5%.

Source: Authors and IRIS centers estimates using data and procedures detailed in the text.

APPENDIX B

APPENDIX FOR CHAPTER 2

Table B.1: Livelihoods 2004

	Cluster one (n=2216)		Cluster two (n=558)		Difference	
Variable	Mean	Std. Dev.	Mean	Std. Dev.	p-value	95% s
share of household members who completed school						
koranic	0.00	0.02	0.00	0.02	0.682	
primary	0.53	0.26	0.55	0.38	0.229	
secondary	0.04	0.12	0.17	0.31	0.000	*
advanced secondary	0.00	0.03	0.03	0.13	0.000	*
university	0.00	0.02	0.01	0.08	0.000	*
adult education	0.01	0.04	0.00	0.05	0.237	
share hh mbrs illness/injury free last 4 wks	0.47	0.37	0.55	0.47	0.000	*
total value of household business assets						
buildings	80493	2018033	1114888	23600000	0.042	*
vehicles	39839	691173	68549	987709	0.425	
equipment	65920	699336	131563	568487	0.040	*
total no. business operated by hh mbrs	0.52	0.61	0.50	0.59	0.488	
hours of household labor per capita per week allocated						
farm wage labor	0.84	3.43	1.18	7.14	0.108	
fishing wage labor	0.11	1.33	0.73	5.82	0.000	*
merchant wage labor	0.19	2.67	0.82	6.48	0.000	*
transportation wage labor	0.23	2.58	0.88	6.28	0.000	*
construction wage labor	0.34	2.46	0.74	5.55	0.011	*
education professional wage labor	0.14	1.22	1.01	5.55	0.000	*
health professional wage labor	0.04	0.66	0.45	4.79	0.000	*
other professional wage labor	0.14	1.71	0.65	4.85	0.000	*
clerical wage labor	0.05	0.78	0.02	0.48	0.456	
factory wage labor	0.05	0.88	0.32	3.55	0.001	*

Table B.1 continued from previous page

bar/hotel wage labor	0.11	1.94	0.75	6.24	0.000	*
skilled wage labor	0.42	2.64	3.34	11.93	0.000	*
other wage labor	0.05	0.74	1.17	8.10	0.000	*
fish self employed labor	0.18	1.51	0.22	2.05	0.578	
merchant self employed labor	1.17	4.20	4.08	11.95	0.000	*
transportation self employed labor	0.06	1.51	0.40	4.12	0.002	*
construction self employed labor	0.13	1.56	0.06	0.76	0.270	
education professional self employed labor	0.01	0.26	0.03	0.67	0.375	
health professional self employed labor	0.00	0.09	0.01	0.28	0.220	
bar/hotel self employed labor	0.05	0.67	0.35	3.25	0.000	*
skilled self employed labor	0.34	1.76	1.29	6.20	0.000	*
other self employed labor	0.00	0.00	0.00	0.00		
own farm labor	6.53	6.94	0.85	3.84	0.000	*
own herd/her processing labor	0.89	1.87	0.12	1.04	0.000	*
total household shamba area (acres)	3.62	4.00	0.17	0.63	0.000	*
total household farm expenditures						
hired labor	16642.15	88022.80	69.23	990.04	0.000	*
seeds	3827.66	9612.44	90.52	851.22	0.000	*
fertilizer	577.97	6158.75	0.00	0.00	0.027	*
organic fertilizer	8588.16	46849.34	33.22	554.40	0.000	*
pesticide	1203.47	11328.85	0.00	0.00	0.012	*
transportation	999.77	9771.34	0.00	0.00	0.016	*
other	1675.51	9191.36	2.86	49.98	0.000	*
total quantity of farm asset owned by household						
hoes	2.87	1.84	0.17	0.58	0.000	*
axes	0.67	0.64	0.03	0.16	0.000	*
machetes	0.06	0.30	0.00	0.06	0.000	*
picks	0.07	0.38	0.01	0.15	0.000	*
shovels	0.31	0.59	0.03	0.19	0.000	*
wheelbarrows	0.05	0.25	0.00	0.04	0.000	*
sickles	1.77	31.86	0.04	0.26	0.202	

Table B.1 continued from previous page

pangas	1.23	0.71	0.07	0.27	0.000	*
mundu	0.16	0.47	0.00	0.04	0.000	*
pruning shears	0.06	0.27	0.01	0.09	0.000	*
other tools	1.11	6.44	0.05	0.30	0.000	*
total value of farm asset owned by household						
mill	13628.25	355668.00	0.00	0.00	0.366	
water equipment	1369.65	12995.18	113.31	2676.56	0.023	*
other	9896.79	31458.28	101.02	1486.71	0.000	*
farm buildings	1864.19	25021.18	172.72	4080.03	0.112	
total number of livestock owned by household						
sheep/goats	1.65	4.85	0.13	0.93	0.000	*
chicken/fowl	3.41	17.96	1.72	18.09	0.047	*
cattle	0.64	2.97	0.04	0.36	0.000	*
pigs	0.17	0.68	0.03	0.37	0.000	*
other	0.16	1.40	0.01	0.17	0.010	*
savings account (yes/no)	0.11	0.31	0.24	0.43	0.000	*
total value of non-labor income received by household						
pension	17950.50	713779.60	0.00	0.00	0.553	
insurance	388.49	6576.96	476.46	5284.65	0.770	
interest	1342.40	20115.75	9573.61	174182.70	0.030	*
lottery	3.09	85.86	0.00	0.00	0.395	
dowry	2051.23	21495.53	0.00	0.00	0.024	*
inheretence	21633.55	269920.50	13282.20	241119.00	0.505	
sale of durables	5005.16	62412.82	24032.94	375721.40	0.024	*
other	3272.79	42469.05	5495.46	127769.40	0.495	
remittances	18703.15	65523.39	34164.63	118195.60	0.000	*
share of crop in total household crop production						
coffee	0.08	0.06	0.00	0.01	0.000	*
tea	0.00	0.01	0.00	0.00	0.241	
tobacco	0.00	0.02	0.00	0.01	0.005	*
cotton	0.00	0.02	0.00	0.03	0.769	

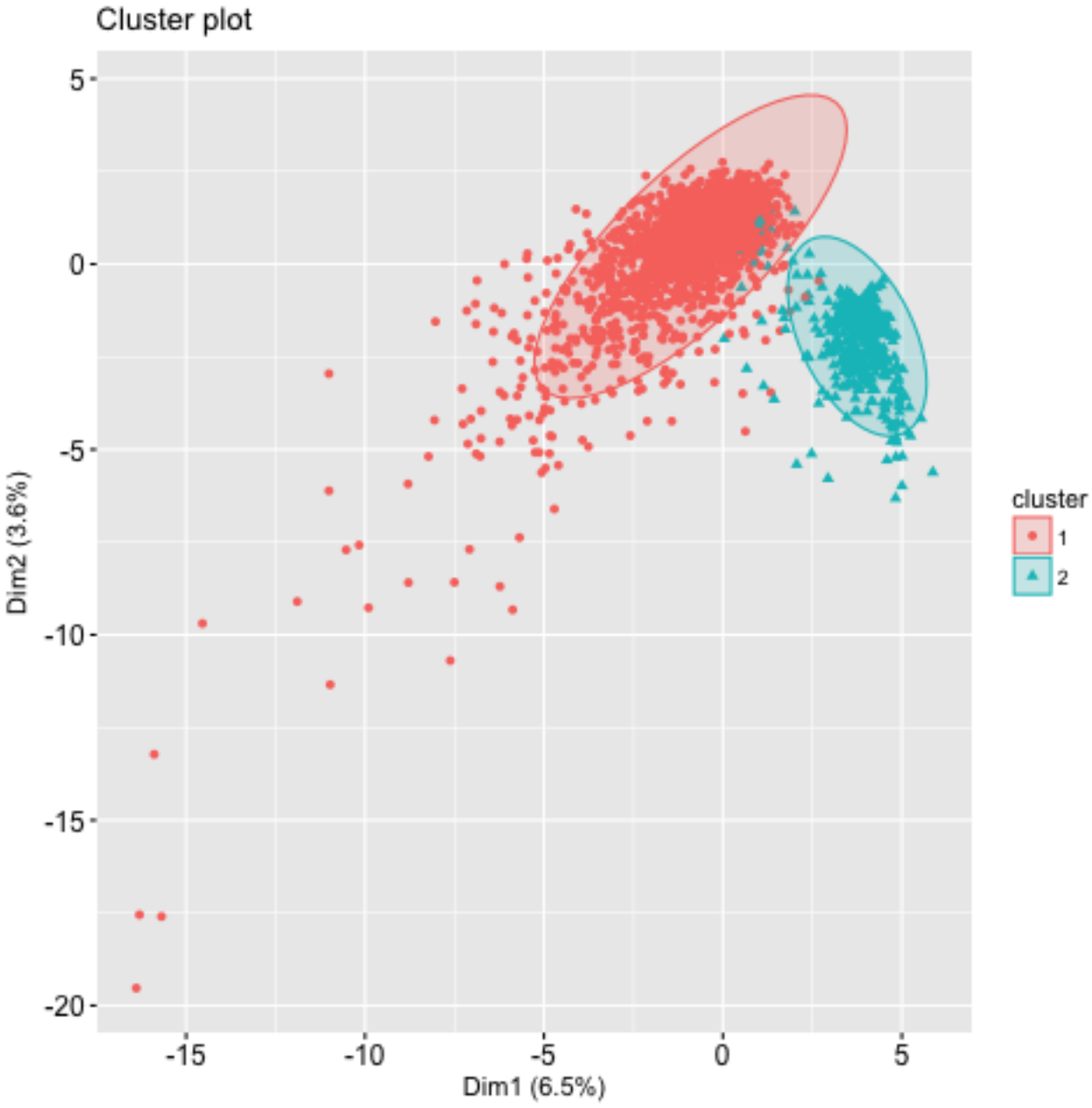
Table B.1 continued from previous page

lumber	0.04	0.05	0.00	0.03	0.000	*
banana	0.11	0.06	0.01	0.03	0.000	*
cassava	0.11	0.07	0.01	0.04	0.000	*
yam	0.05	0.06	0.00	0.00	0.000	*
sweet potato	0.10	0.06	0.01	0.04	0.000	*
potato	0.01	0.03	0.00	0.01	0.000	*
maize	0.13	0.07	0.02	0.08	0.000	*
millet/sorghum	0.02	0.05	0.00	0.03	0.000	*
rice	0.00	0.03	0.01	0.06	0.188	
beans/pulses	0.12	0.06	0.01	0.05	0.000	*
sunflower seeds	0.00	0.01	0.00	0.00	0.001	*
mambara	0.02	0.04	0.00	0.02	0.000	*
fruit	0.10	0.06	0.00	0.02	0.000	*
vegetables	0.05	0.07	0.00	0.01	0.000	*
other	0.04	0.05	0.01	0.04	0.000	*
mushrooms	0.00	0.00	0.00	0.04	0.075	
peas	0.01	0.03	0.00	0.01	0.000	*
vanilla	0.01	0.02	0.00	0.01	0.000	*
Tanzania	0.99	0.11	0.96	0.20	0.000	*
Uganda	0.01	0.11	0.04	0.20	0.000	*
region						
Kagera	0.96	0.20	0.59	0.49	0.000	*
Dar Es Salaam	0.00	0.06	0.11	0.31	0.000	*
Arusha	0.00	0.00	0.01	0.09	0.000	*
Other	0.01	0.10	0.03	0.18	0.000	*
Dodoma	0.00	0.00	0.01	0.09	0.000	*
Kampala	0.00	0.02	0.00	0.06	0.044	*
Kigoma	0.00	0.05	0.01	0.09	0.018	*
Kilimanjaro	0.00	0.02	0.00	0.06	0.044	*
Kyotera	0.00	0.02	0.00	0.04	0.292	
Mara	0.00	0.05	0.01	0.11	0.001	*

Table B.1 continued from previous page

Masaka	0.00	0.02	0.00	0.00	0.616	
Mbeya	0.00	0.00	0.00	0.04	0.046	*
Morogoro	0.00	0.00	0.01	0.10	0.000	*
Mwanza	0.01	0.10	0.14	0.35	0.000	*
Pwani	0.00	0.02	0.01	0.07	0.006	*
Ruka	0.00	0.00	0.01	0.07	0.001	*
Shinyanga	0.01	0.09	0.04	0.21	0.000	*
Southern	0.00	0.03	0.00	0.04	0.568	
Tabora	0.00	0.05	0.01	0.08	0.068	
*variables not used in cluster analysis, denoted **						
total ann. cons per cap in 2010 Tsh**	396583.20	296019.30	977533.00	671399.60	0.000	*
poor (yes/no)**	0.03	0.16	0.01	0.09	0.009	*
share of hh mmbrs ever moved**	0.21	0.24	0.49	0.37	0.000	*
household migrated (outside of ea)**	0.43	0.50	0.82	0.38	0.000	*
average age all hh mmbrs**	23.34	11.86	22.26	8.00	0.041	*
share of hh mmbrs female**	0.51	0.21	0.45	0.35	0.000	*
age of head**	43.73	17.28	32.15	11.92	0.000	*
head is female (yes/no)**	0.21	0.41	0.22	0.42	0.570	
head is married (yes/no)**	0.79	0.41	0.51	0.50	0.000	*
total children $\leq$ 5 yrs**	1.10	1.03	0.52	0.79	0.000	*
household size**	5.03	2.55	3.05	2.36	0.000	*

Figure B.1: 2004 clusters



APPENDIX C

APPENDIX FOR CHAPTER 3

Figure C.1: Income density by portfolio, values in asinh 2012-13 TSh

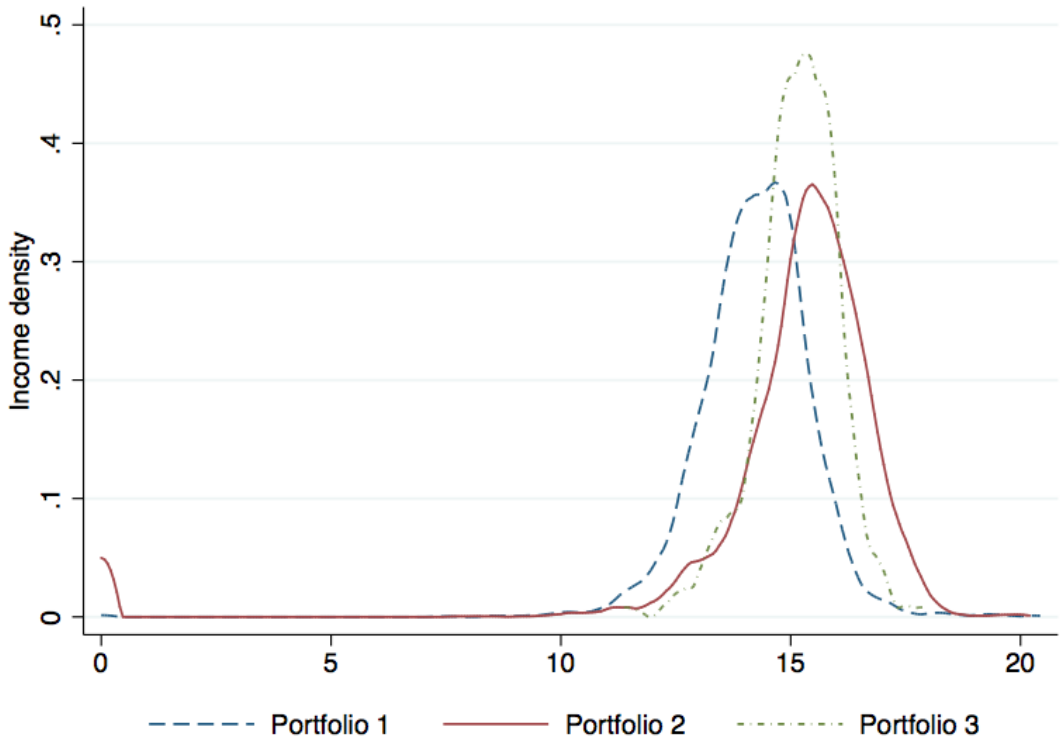


Table C.1: Contribution of assets to portfolio moments

	Mean	Variance	Skew
plotval	0.295***	0.580***	2.289***
	(0.0852)	(0.0940)	(0.208)
bullval	0.0100	0.563***	3.592***
	(0.0322)	(0.0434)	(0.0859)
cowval	0.642***	1.373***	7.453***
	(0.140)	(0.132)	(0.226)

steerval	-0.380**	-0.0104	-24.76***
	(0.168)	(0.271)	(0.560)
hefval	1.187***	0.0632	25.13***
	(0.205)	(0.126)	(0.195)
mcalfval	-0.995***	1.917***	-25.05***
	(0.195)	(0.160)	(0.304)
fcalfval	-0.261*	-0.694***	0.460
	(0.158)	(0.173)	(0.353)
goatval	-0.262**	-1.167***	-3.920***
	(0.108)	(0.152)	(0.268)
sheepval	-0.732***	-2.167***	-10.50***
	(0.242)	(0.158)	(0.300)
pigval	0.583	0.471	24.40***
	(0.578)	(0.380)	(0.727)
chickenval	0.142	-1.876***	7.843***
	(0.282)	(0.114)	(0.191)
otlstkval	0.642**	2.346***	11.63***
	(0.324)	(0.189)	(0.339)
handhoeval	-1.321***	0.488	-39.52***
	(0.430)	(0.420)	(0.856)
handsprayval	1.079***	0.985***	34.49***
	(0.381)	(0.112)	(0.201)
oxploughval	0.407**	1.833***	-7.508***
	(0.194)	(0.155)	(0.332)
oxseedval	2.039***	8.027***	9.721***
	(0.384)	(0.365)	(0.676)
oxcartval	-0.430**	3.560***	-9.323***

	(0.203)	(0.397)	(0.718)
tractorval	-0.768***	-2.576***	-13.21***
	(0.186)	(0.303)	(0.567)
tractharrowval	-0.664*	1.843***	-2.899***
	(0.385)	(0.433)	(0.727)
thresherval	0.179		
	(0.334)		
watercanval	0.653**	-0.152	0.0123
	(0.320)	(0.192)	(0.256)
farmbldgsval	0.781***	1.267***	13.34***
	(0.213)	(0.125)	(0.231)
gericanval	0.0959	-3.344***	8.074***
	(0.128)	(0.138)	(0.282)
fishnetval	-0.0671	-0.103**	0.0266
	(0.0553)	(0.0500)	(0.112)
fishlineval	-0.0340	0.0637	-0.688**
	(0.127)	(0.127)	(0.290)
fishboatval	0.163*	0.0583	0.291
	(0.0834)	(0.0974)	(0.188)
fishmotorval	-0.458***	0.236	-0.0607
	(0.127)	(0.225)	(0.562)
bus_capitalval	-0.539	-2.681***	6.795***
	(0.500)	(0.169)	(0.294)
bus_stockval	0.0275	10.97***	13.52***
	(0.0277)	(0.503)	(0.883)
bus_goodsval	-0.169	-1.278***	-3.699***
	(0.310)	(0.363)	(0.612)

phone_landval	0.184	0.0741	-0.464
	(0.400)	(0.354)	(0.847)
phone_mobileval	-0.600***	1.490***	1.490***
	(0.232)	(0.249)	(0.403)
fridge_freezeval	-0.165*	0.268	0.635*
	(0.0996)	(0.354)	(0.378)
sewmachval	-0.449	-0.0445	-0.795
	(0.316)	(0.368)	(0.842)
computerval	0.197	-0.0787	-0.134
	(0.135)	(0.210)	(0.607)
stove_geval	-0.290	-0.940	-1.696
	(0.376)	(0.615)	(1.240)
stove_otherval	0.0625	5.663***	-35.14***
	(0.0586)	(1.006)	(2.136)
carval	-0.234**	0.270	-0.749
	(0.0991)	(0.251)	(0.798)
motorcycleval	-0.154	0.393	-1.786
	(0.340)	(0.866)	(1.558)
bicycleval	0.687***	5.741***	-16.74***
	(0.201)	(0.237)	(0.432)
plotval2	-0.0207***	-0.0662***	-0.486***
	(0.00672)	(0.00703)	(0.0146)
bullval2	0.0154**	-0.0262***	-0.511***
	(0.00663)	(0.00522)	(0.00990)
cowval2	-0.104***	-0.245***	-0.621***
	(0.0196)	(0.0179)	(0.0293)
steerval2	0.0549**	0.00989	3.150***

	(0.0245)	(0.0337)	(0.0688)
hefval2	-0.177***	-0.0401**	-3.480***
	(0.0329)	(0.0185)	(0.0289)
mcalfval2	0.142***	-0.251***	3.705***
	(0.0299)	(0.0233)	(0.0446)
fcalfval2	0.0409	0.114***	-0.0852*
	(0.0263)	(0.0248)	(0.0507)
goatval2	0.0462***	0.167***	0.623***
	(0.0173)	(0.0233)	(0.0403)
sheepval2	0.103***	0.345***	1.599***
	(0.0354)	(0.0288)	(0.0537)
pigval2	-0.0917	-0.136**	-3.471***
	(0.0858)	(0.0569)	(0.108)
chickenval2	-0.0000222	0.330***	-1.004***
	(0.0511)	(0.0192)	(0.0336)
otlstkval2	-0.134***	-0.381***	-2.063***
	(0.0507)	(0.0318)	(0.0572)
handhoeval2	0.118*	-0.106*	4.840***
	(0.0620)	(0.0547)	(0.106)
handsprayval2	-0.181***	-0.0879***	-6.021***
	(0.0662)	(0.0188)	(0.0333)
oxploughval2	-0.0519*	-0.233***	1.010***
	(0.0296)	(0.0197)	(0.0422)
oxseedval2	-0.300***	-1.182***	-1.462***
	(0.0568)	(0.0556)	(0.103)
oxcartval2	0.0698**	-0.544***	1.456***
	(0.0340)	(0.0607)	(0.110)

tractorval2	0.110***	0.323***	2.208***
	(0.0217)	(0.0395)	(0.0735)
tractploughval2	0.0837**	-0.155***	-0.0994
	(0.0412)	(0.0465)	(0.0843)
thresherval2	-0.0323		
	(0.0576)		
watercanval2	-0.217***	0.432***	-3.390***
	(0.0696)	(0.0462)	(0.0652)
farmbldgsval2	-0.134***	-0.218***	-2.136***
	(0.0352)	(0.0204)	(0.0373)
gericanval2	-0.0181	0.612***	-1.558***
	(0.0239)	(0.0230)	(0.0462)
fishnetval2	0.00685	0.0165	-0.0895*
	(0.00901)	(0.0217)	(0.0537)
fishlineval2	0.000500	-0.0131	0.0913**
	(0.0207)	(0.0201)	(0.0463)
fishboatval2	-0.0248*	0.0171	-0.0755*
	(0.0132)	(0.0192)	(0.0435)
fishmotorval2	0.0650***	-0.124***	0.297***
	(0.0179)	(0.0261)	(0.0994)
bus_capitalval2	0.100	0.426***	-0.570***
	(0.0784)	(0.0280)	(0.0482)
bus_stockval2	0.00614	-1.637***	-3.134***
	(0.0143)	(0.0798)	(0.139)
bus_goodsval2	0.0249	0.198***	0.506***
	(0.0509)	(0.0612)	(0.103)
phone_landval2	-0.0261	0.126**	-0.663***

	(0.0676)	(0.0620)	(0.147)
phone_mobileval2	0.118***	-0.287***	-0.466***
	(0.0428)	(0.0453)	(0.0737)
fridge_freezeval2	0.0708***	-0.0917*	0.193***
	(0.0209)	(0.0528)	(0.0578)
sewmachval2	0.0271	0.0647	-0.419***
	(0.0523)	(0.0585)	(0.133)
computerval2	-0.0451**	-0.0834***	0.395***
	(0.0222)	(0.0304)	(0.0831)
stove_geval2	0.0651	0.103	0.816***
	(0.0615)	(0.0966)	(0.195)
stove_otherval2	0.0172	-1.043***	7.323***
	(0.0189)	(0.209)	(0.444)
carval2	0.0271**	-0.0266	0.0771
	(0.0114)	(0.0292)	(0.0935)
motorcycleval2	0.0338	-0.00677	0.820***
	(0.0455)	(0.115)	(0.205)
bicycleval2	-0.115***	-0.927***	2.760***
	(0.0325)	(0.0383)	(0.0707)
port1 × plotval	-0.299***	-0.587***	-2.275***
	(0.0875)	(0.0942)	(0.209)
port1 × bullval	0.200	-1.139***	-2.671***
	(0.161)	(0.220)	(0.449)
port1 × cowval	-0.706***	-1.435***	-7.780***
	(0.149)	(0.124)	(0.211)
port1 × steerval	0.312	-0.246	25.11***
	(0.233)	(0.317)	(0.644)

port1 × hefval	-1.135***	-0.0888	-25.45***
	(0.261)	(0.267)	(0.406)
port1 × mcalfval	0.869***	-1.534***	24.81***
	(0.264)	(0.262)	(0.423)
port1 × fcalfval	0.0856	0.662***	-0.801**
	(0.181)	(0.211)	(0.373)
port1 × goatval	0.245**	1.031***	3.807***
	(0.110)	(0.155)	(0.279)
port1 × sheepval	0.804***	2.240***	10.44***
	(0.261)	(0.209)	(0.366)
port1 × pigval	-0.744	-0.555	-24.93***
	(0.581)	(0.393)	(0.742)
port1 × chickenval	-0.151	1.898***	-7.788***
	(0.281)	(0.118)	(0.202)
port1 × otlstkval	-0.606*	-2.342***	-11.55***
	(0.326)	(0.190)	(0.338)
port1 × handhoeval	1.368***	-0.351	39.70***
	(0.431)	(0.424)	(0.860)
port1 × handsprayval	-1.034***	-1.069***	-33.85***
	(0.385)	(0.204)	(0.447)
port1 × oxploughval	-0.295	-1.977***	7.927***
	(0.228)	(0.178)	(0.380)
port1 × oxseedval	-1.965***	-8.193***	-9.323***
	(0.462)	(0.410)	(0.732)
port1 × oxcartval	0.367	-3.422***	9.440***
	(0.224)	(0.430)	(0.871)
port1 × tractploughval	0.0495	-0.236***	1.006***

	(0.0319)	(0.0466)	(0.0889)
port1 × watercanval	-0.684*		
	(0.354)		
port1 × farmbldgsval	-0.801***	-1.226***	-13.53***
	(0.217)	(0.146)	(0.261)
port1 × gericanval	-0.0568	3.414***	-8.057***
	(0.146)	(0.175)	(0.328)
port1 × fishlineval	0.0186	0.00659	1.126**
	(0.150)	(0.204)	(0.531)
port1 × bus_capitalval	0.538	2.699***	-6.755***
	(0.500)	(0.167)	(0.285)
port1 × bus_stockval	-0.0135	-10.84***	-13.29***
	(0.0428)	(0.501)	(0.876)
port1 × bus_goodsval	0.162	1.160***	3.674***
	(0.310)	(0.364)	(0.613)
port1 × phone_mobileval	0.644***	-1.888***	-1.338***
	(0.243)	(0.263)	(0.415)
port1 × fridge_freezeval	0.505*		
	(0.279)		
port1 × computerval	-0.362*	-0.500	-1.492
	(0.217)	(0.540)	(1.364)
port1 × carval	0.244	-0.504	-0.146
	(0.414)	(0.444)	(0.957)
port1 × bicycleval	-0.666***	-5.850***	17.07***
	(0.211)	(0.253)	(0.451)
port1 × plotval2	0.0211***	0.0666***	0.484***
	(0.00697)	(0.00745)	(0.0157)

port1 × bullval2	-0.0460**	0.110***	0.378***
	(0.0227)	(0.0330)	(0.0686)
port1 × cowval2	0.113***	0.253***	0.669***
	(0.0199)	(0.0173)	(0.0275)
port1 × steerval2	-0.0475	0.0273	-3.205***
	(0.0331)	(0.0410)	(0.0823)
port1 × hefval2	0.170***	0.0428	3.525***
	(0.0404)	(0.0386)	(0.0583)
port1 × mcalfval2	-0.125***	0.194***	-3.670***
	(0.0398)	(0.0383)	(0.0615)
port1 × fcalfval2	-0.0151	-0.109***	0.134**
	(0.0292)	(0.0301)	(0.0535)
port1 × goatval2	-0.0422**	-0.147***	-0.602***
	(0.0177)	(0.0239)	(0.0421)
port1 × sheepval2	-0.112***	-0.358***	-1.594***
	(0.0387)	(0.0358)	(0.0618)
port1 × pigval2	0.118	0.149**	3.554***
	(0.0863)	(0.0592)	(0.110)
port1 × chickenval2	0.000786	-0.336***	0.991***
	(0.0511)	(0.0200)	(0.0358)
port1 × otlstkval2	0.126**	0.379***	2.042***
	(0.0509)	(0.0319)	(0.0571)
port1 × handhoeval2	-0.125**	0.0861	-4.869***
	(0.0622)	(0.0552)	(0.106)
port1 × handsprayval2	0.173***	0.105***	5.905***
	(0.0669)	(0.0340)	(0.0747)
port1 × oxploughval2	0.0382	0.253***	-1.065***

	(0.0344)	(0.0239)	(0.0504)
port1 × oxseedval2	0.287***	1.207***	1.398***
	(0.0706)	(0.0637)	(0.113)
port1 × oxcartval2	-0.0621*	0.525***	-1.474***
	(0.0372)	(0.0657)	(0.133)
port1 × tractorval2	0.00519	0.0857***	-0.163***
	(0.0112)	(0.00868)	(0.0159)
port1 × tractploughval2	-0.0877**	0.234***	-0.206**
	(0.0439)	(0.0503)	(0.0905)
port1 × watercanval2	0.219***	-0.400***	3.381***
	(0.0760)	(0.0222)	(0.0381)
port1 × farmbldgsval2	0.136***	0.212***	2.166***
	(0.0360)	(0.0242)	(0.0427)
port1 × gericanval2	0.0112	-0.627***	1.554***
	(0.0271)	(0.0307)	(0.0558)
port1 × fishlineval2	0.00359	0.000795	-0.186*
	(0.0270)	(0.0378)	(0.101)
port1 × bus_capitalval2	-0.1000	-0.430***	0.564***
	(0.0784)	(0.0277)	(0.0469)
port1 × bus_stockval2	-0.00753	1.615***	3.093***
	(0.0150)	(0.0794)	(0.138)
port1 × bus_goodsval2	-0.0256	-0.179***	-0.510***
	(0.0510)	(0.0616)	(0.104)
port1 × phone_landval2	-0.00292	-0.156	0.536***
	(0.00863)	(0.101)	(0.192)
port1 × phone_mobileval2	-0.126***	0.354***	0.436***
	(0.0444)	(0.0472)	(0.0754)

port1 × fridge_freezeval2	-0.118***	0.0494***	-0.281***
	(0.0433)	(0.00983)	(0.0177)
port1 × sewmachval2	0.0462***	-0.0577***	0.551***
	(0.0172)	(0.00848)	(0.0152)
port1 × computerval2	0.0659**	0.161**	-0.184
	(0.0305)	(0.0700)	(0.176)
port1 × stove_geval2	-0.0160	0.0487***	-0.543***
	(0.0136)	(0.00824)	(0.0144)
port1 × stove_otherval2	-0.0312**	1.077***	-7.368***
	(0.0142)	(0.208)	(0.444)
port1 × carval2	-0.0284	0.0520	0.0213
	(0.0468)	(0.0508)	(0.111)
port1 × motorcycleval2	-0.0127*	-0.0442***	-0.584***
	(0.00694)	(0.00349)	(0.00633)
port1 × bicycleval2	0.111***	0.943***	-2.817***
	(0.0342)	(0.0410)	(0.0737)
port2 × plotval	-0.290***	-0.611***	-2.324***
	(0.0876)	(0.103)	(0.230)
port2 × bullval	-1.831	0.253	-6.927
	(1.513)	(2.896)	(6.817)
port2 × cowval	-0.900***	-1.982***	-6.092***
	(0.293)	(0.477)	(0.980)
port2 × hefval	-1.051***	1.237*	-26.59***
	(0.330)	(0.638)	(1.058)
port2 × mcalfval	0.841**	-0.729	23.73***
	(0.332)	(0.612)	(0.935)
port2 × fcalfval	0.820**	1.266*	0.934

	(0.405)	(0.675)	(1.387)
port2 × goatval	0.408**	-0.0366	3.525***
	(0.206)	(0.615)	(0.830)
port2 × sheepval	2.659*	1.130	16.66***
	(1.444)	(1.062)	(3.370)
port2 × chickenval	-0.181	1.719***	-7.841***
	(0.290)	(0.140)	(0.254)
port2 × otlstkval	-0.542	-2.119***	-11.87***
	(0.338)	(0.212)	(0.381)
port2 × handhoeval	1.305***	-0.557	39.74***
	(0.432)	(0.429)	(0.877)
port2 × handsprayval	-0.787*	-1.612***	-32.41***
	(0.427)	(0.352)	(0.728)
port2 × tractorval	1.483***	5.672***	20.13***
	(0.378)	(0.590)	(1.264)
port2 × farmbldgsval	-0.627***	-1.634***	-10.65***
	(0.226)	(0.294)	(1.411)
port2 × gericanval	0.0364	3.280***	-7.533***
	(0.161)	(0.213)	(0.469)
port2 × fishnetval	0.349***	-0.0559	0.791*
	(0.128)	(0.163)	(0.404)
port2 × fishboatval	-0.0741	-0.0671	0.0964
	(0.149)	(0.185)	(0.395)
port2 × fishmotorval	0.420*		
	(0.219)		
port2 × bus_capitalval	0.545	2.684***	-6.735***
	(0.500)	(0.170)	(0.296)

port2 × bus_goodsval	0.164	1.200***	3.850***
	(0.311)	(0.364)	(0.616)
port2 × phone_landval	0.171		
	(0.454)		
port2 × phone_mobileval	0.673***	-1.729***	-1.078**
	(0.235)	(0.258)	(0.426)
port2 × sewmachval	0.376	0.170	0.357
	(0.326)	(0.389)	(0.886)
port2 × stove_geval	0.359	0.679	2.317*
	(0.394)	(0.655)	(1.372)
port2 × stove_otherval	0.0652	-5.951***	35.81***
	(0.0700)	(1.007)	(2.136)
port2 × motorcycleval	0.0220	-0.125	1.249
	(0.382)	(0.939)	(1.794)
port2 × bicycleval	-0.879***	-5.462***	16.27***
	(0.208)	(0.245)	(0.447)
port2 × plotval2	0.0196**	0.0692***	0.488***
	(0.00849)	(0.00897)	(0.0197)
port2 × bullval2	0.252	-0.0922	0.981
	(0.222)	(0.422)	(0.987)
port2 × cowval2	0.141***	0.333***	0.417***
	(0.0465)	(0.0685)	(0.140)
port2 × steerval2	0.00406	-0.0190**	0.572***
	(0.00428)	(0.00862)	(0.0183)
port2 × hefval2	0.158***	-0.152*	3.699***
	(0.0495)	(0.0918)	(0.153)
port2 × mcalfval2	-0.122**	0.0702	-3.518***

	(0.0499)	(0.0924)	(0.140)
port2 × fcalfval2	-0.119**	-0.193**	-0.105
	(0.0585)	(0.0943)	(0.194)
port2 × goatval2	-0.0667**	0.0172	-0.553***
	(0.0327)	(0.0966)	(0.129)
port2 × sheepval2	-0.399*	-0.189	-2.530***
	(0.222)	(0.170)	(0.523)
port2 × pigval2	-0.00186	0.318**	2.506***
	(0.00519)	(0.135)	(0.285)
port2 × chickenval2	0.00556	-0.306***	1.002***
	(0.0523)	(0.0231)	(0.0432)
port2 × otlstkval2	0.118**	0.340***	2.096***
	(0.0532)	(0.0356)	(0.0637)
port2 × handhoeval2	-0.115*	0.122**	-4.882***
	(0.0627)	(0.0572)	(0.112)
port2 × handsprayval2	0.124*	0.218***	5.609***
	(0.0748)	(0.0633)	(0.129)
port2 × oxploughval2	0.0150**	0.0704***	0.323***
	(0.00643)	(0.00852)	(0.0407)
port2 × tractorval2	-0.264***	-0.916***	-3.691***
	(0.0616)	(0.0892)	(0.191)
port2 × watercanval2	0.0858***	-0.168*	2.446***
	(0.0301)	(0.0947)	(0.320)
port2 × farmbldgsval2	0.110***	0.276***	1.715***
	(0.0370)	(0.0461)	(0.218)
port2 × gericanval2	-0.00232	-0.603***	1.474***
	(0.0295)	(0.0361)	(0.0825)

port2 × fishnetval2	-0.0426**		
	(0.0183)		
port2 × fishboatval2	0.00925		
	(0.0198)		
port2 × fishmotorval2	-0.0528*	0.0800**	-0.269**
	(0.0295)	(0.0394)	(0.124)
port2 × bus_capitalval2	-0.102	-0.425***	0.558***
	(0.0784)	(0.0282)	(0.0485)
port2 × bus_stockval2	-0.00993	1.638***	3.127***
	(0.0133)	(0.0800)	(0.139)
port2 × bus_goodsval2	-0.0244	-0.187***	-0.529***
	(0.0510)	(0.0614)	(0.104)
port2 × phone_landval2	-0.0335	-0.139***	0.746***
	(0.0767)	(0.0173)	(0.0327)
port2 × phone_mobileval2	-0.129***	0.325***	0.398***
	(0.0432)	(0.0466)	(0.0773)
port2 × fridge_freezeval2	-0.0464***	0.0887	-0.148*
	(0.0146)	(0.0568)	(0.0778)
port2 × sewmachval2	-0.0158	-0.0822	0.484***
	(0.0550)	(0.0617)	(0.140)
port2 × computerval2	0.0169	0.0968***	-0.383***
	(0.0131)	(0.0134)	(0.0253)
port2 × stove_geval2	-0.0750	-0.0639	-0.909***
	(0.0642)	(0.103)	(0.215)
port2 × stove_otherval2	-0.0434**	1.101***	-7.460***
	(0.0206)	(0.209)	(0.444)
port2 × motorcycleval2	-0.0168	-0.0294	-0.749***

	(0.0507)	(0.124)	(0.237)
port2 × bicycleval2	0.146***	0.882***	-2.687***
	(0.0338)	(0.0398)	(0.0734)
Household size	0.0941***	0.195***	1.548***
	(0.0292)	(0.0469)	(0.0810)
Head of household female	-0.0806	-0.0859	-0.134
	(0.0778)	(0.0989)	(0.194)
Head age	0.0294	-0.223***	1.866***
	(0.0332)	(0.0116)	(0.0196)
Head married	-0.0201	0.0136	-0.140
	(0.0782)	(0.120)	(0.214)
Head migrated to this area	0.266	-0.596**	-0.420
	(0.278)	(0.290)	(0.477)
head_primaryed	0.826	2.699***	12.13***
	(0.522)	(0.258)	(0.518)
head_secondaryed	-0.0186	-0.128*	0.303*
	(0.160)	(0.0773)	(0.160)
head_university	0.0885	0.559	-1.687
	(0.136)	(0.462)	(1.181)
port1 × Household size	-0.111***	-0.273***	-1.777***
	(0.0327)	(0.0543)	(0.0998)
port1 × Head of household female	0.0748	0.267	0.496
	(0.121)	(0.189)	(0.374)
port1 × Head age	-0.0310	0.229***	-1.863***
	(0.0333)	(0.0119)	(0.0210)
port1 × Head migrated to this area	-0.294	0.493*	0.214
	(0.280)	(0.297)	(0.498)

port1 × head_primaryed	-0.821	-2.759***	-12.27***
	(0.525)	(0.274)	(0.552)
port2 × Household size	-0.111***	-0.0699	-1.828***
	(0.0303)	(0.0503)	(0.0911)
port2 × Head age	-0.0291	0.221***	-1.869***
	(0.0332)	(0.0126)	(0.0225)
port2 × Head married	0.0529	-0.0592	0.415
	(0.0919)	(0.138)	(0.257)
port2 × Head migrated to this area	-0.300	0.689**	0.262
	(0.283)	(0.296)	(0.497)
port2 × head_primaryed	-0.857	-2.638***	-12.16***
	(0.522)	(0.265)	(0.531)
port2 × head_secondaryed	0.0600		
	(0.167)		
port1 × thresherval		-0.433**	0.425
		(0.209)	(0.381)
port1 × fishmotorval		0.696**	-2.140**
		(0.294)	(0.920)
port1 × phone_landval		0.131	1.288
		(0.595)	(1.129)
port1 × stove_otherval		-5.816***	35.36***
		(1.002)	(2.137)
port1 × thresherval2		0.0788**	-0.0749
		(0.0382)	(0.0684)
port1 × fishnetval2		-0.00246	0.0807
		(0.0228)	(0.0563)
port1 × fishboatval2		-0.0241	0.0362

		(0.0240)	(0.0521)
port2 × pigval		-1.442	-18.24***
		(0.940)	(1.993)
port2 × watercanval		-1.027**	4.428***
		(0.462)	(1.586)
port2 × bus_stockval		-10.97***	-13.47***
		(0.504)	(0.885)
port2 × fridge_freezeval		-0.236	-0.962*
		(0.381)	(0.513)
port1 × head_secondaryed		-0.0957	-1.263***
		(0.343)	(0.471)
Constant	15.82***	0.440***	1.233***
	(0.111)	(0.171)	(0.322)
Observations	5017	5271	5271
Adjusted R^2	0.113	0.583	0.876

Standard errors in parentheses

* $p < .1$, ** $p < .05$, *** $p < .01$

BIBLIOGRAPHY

[1] Adato, M., Carter, M.R., & May, J. (2006). Exploring poverty traps and social exclusion in South Africa using qualitative and quantitative data. *Development Studies*. 42(2):22647

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

[3] Antle, J. M. (1987). Econometric estimation of producers' risk attitudes. *American Journal of Agricultural Economics*, 69(3), 509-522.

[4] Athey, S. (2017). Beyond prediction: Using big data for policy problems. *Science*, 355(6324), 483-485.

[5] Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27), 7353-7360.

[6] Banerjee, A. V., & Newman, A. F. (1993). Occupational choice and the process of development. *Journal of Political Economy*, 274-298.

[7] Barrett, C. B. (2005). Rural poverty dynamics: development policy implications. *Agricultural Economics*, 32(s1), 45-60.

[8] Barrett, C. B. (2008). Smallholder market participation: Concepts and evidence from eastern and southern Africa. *Food policy,* 33(4), 299-317

[9] Barrett, C. B., Bezuneh, M., Clay, D. C., & Reardon, T. (2000). Heterogeneous Constraints, Incentives and Income Diversification Strategies in Rural Africa. *USAID Working Paper*. Available at http://pdf.usaid.gov/pdf$_{d}ocs/PNACL435.pdf$

[10] Barrett, C. B., & Carter, M. R. (2013). The economics of poverty traps and persistent poverty: empirical and policy implications. *The Journal of Development Studies*, 49(7), 976-990.

[11] Barrett, C. B., Garg, T., & McBride, L. (2016). Well-being dynamics and poverty traps. *Annual Review of Resource Economics*, 8, 303-327.

[12] Barrett, C. B., & E. Lentz. 2013. Hunger and Food Insecurity. In D. Brady and L.M. Burton, eds., *The Oxford Handbook of Poverty and Society*. Oxford: Oxford University Press.

[13] Barrett, C.B., Marenya, P.P., McPeak, J., Minten, B., Murithi, F., et al. (2006). Welfare dynamics in rural Kenya and Madagascar. *Journal of Development Studies*. 42.2: 248-277

[14] Beegle, K., De Weerdt, J., & Dercon, S. (2011). Migration and economic mobility in Tanzania: Evidence from a tracking survey. *Review of Economics and Statistics*, 93(3), 1010-1033.

[15] Breiman, L. 2001. Random Forests. *Machine Learning*. 45: 532.

[16] Breiman, L. 2004. Consistency for a Simple Model of Random Forests. *Technical Report*. University of California-Berkeley.

[17] Brown, D. R., Stephens, E. C., Ouma, J. O., Murithi, F. M., & Barrett, C. B. (2006). Livelihood strategies in the rural Kenyan highlands. *African Journal of Agricultural and Resource Economics*, 1(1).

[18] Bryan, G., Chowdhury, S., & Mobarak, A. M. (2014). Underinvestment in a profitable technology: The case of seasonal migration in Bangladesh. *Econometrica*, 82(5), 1671-1748.

[19] Buera, F. J. (2009). A dynamic model of entrepreneurship with borrowing constraints: theory and evidence. *Annals of Finance*, 5(3-4), 443-464.

[20] Carter, M. R. (1997). Environment, technology, and the social articulation of risk in West African agriculture. *Economic Development and Cultural Change*, 45(3), 557-590.

[21] Carter, M. R., Barrett, C. B. (2006). The economics of poverty traps and persistent poverty: An asset-based approach. *The Journal of Development Studies*, 42(2), 178-199.

[22] Carter, M., & Ikegami, M. (2009). Looking forward: theory-based measures of chronic poverty and vulnerability. *Poverty dynamics: Interdisciplinary Perspectives*, 128-153.

[23] Carter, M. R., Little, P. D., Mogues, T., & Negatu, W. (2007). Poverty traps and natural disasters in Ethiopia and Honduras. *World Development*, 35(5), 835-856.

[24] Carter, M. R., & Lybbert, T. J. (2012). Consumption versus asset smoothing: testing the implications of poverty trap theory in Burkina Faso. *Journal of Development Economics*, 99(2), 255-264.

[25] Chavas, JP. (2004). *Risk Analysis in Theory and Practice*. Academic Press.

[26] Christiaensen, L., Weerdt, J., & Todo, Y. (2013). Urbanization and poverty reduction: the role of rural diversification and secondary towns. *Agricultural Economics*, 44(4-5), 435-447.

[27] Clemens, M. A. (2011). Economics and emigration: Trillion-dollar bills on the sidewalk?. *The Journal of Economic Perspectives*, 25(3), 83-106.

[28] Clemens, M. A., Montenegro, C. E., & Pritchett, L. (2009). The place premium: wage differences for identical workers across the US border. *HKS Faculty Research Working Paper Series RWP09-004*, John F. Kennedy School of Government, Harvard University.

[29] Coady, D., M. Grosh, & J. Hoddinott. 2004. *Targeting of Transfers in Developing Countries: Review of Lessons and Experience*. Washington, DC: The International Bank for Reconstruction and Development.

[30] Deaton Angus, S. (1991). Saving and Liquidity Constraints. *Ecomometrica*, 59(5), 221-248.

[31] De Janvry, A., Fafchamps, M., & Sadoulet, E. (1991). Peasant household behaviour with missing markets: some paradoxes explained. *The Economic Journal*, 101(409), 1400-1417.

[32] De Janvry, A., & Sadoulet, E. (2005). Progress in the modeling of rural housholds behavior under market failures. *Poverty, inequality and development, essays in honor of Erik Thorbecke*, 8.

[33] Dercon, S. (1996). Risk, crop choice, and savings: Evidence from Tanzania. *Economic Development and Cultural Change*, 44(3), 485-513.

[34] Dercon, S. (1998). Wealth, risk and activity choice: cattle in Western Tanzania. *Journal of Development Economics*, 55(1), 1-42.

[35] Dercon, S., Krishnan, P. (1996). Income portfolios in rural Ethiopia and Tanzania: choices and constraints. *The Journal of Development Studies*, 32(6), 850-875.

[36] De Weerdt, J. (2010). Moving out of poverty in Tanzania: Evidence from Kagera. *The Journal of Development Studies*, 46(2), 331-349.

[37] De Weerdt, J, K Beegle, H Liller, S Dercon, K Hirvonen, M Kirchberger & S Krutikova. (2012). *Kagera Health and Development Survey 2010: Basic Information Document*. Rockwool Foundation Working Paper Series, Study Paper No. 46.

[38] De Weerdt, J., & Hirvonen, K. (2016). Risk sharing and internal migration. *Economic Development and Cultural Change*, 65(1), 63-86.

[39] Di Falco, S., Chavas, J. P. (2006). Crop genetic diversity, farm productivity and the management of environmental risk in rainfed agriculture. *European Review of Agricultural Economics*, 33(3), 289-314.

[40] Di Falco, S., Chavas, J. P. (2009). On crop biodiversity, risk exposure, and food security in the highlands of Ethiopia. *American Journal of Agricultural Economics*, 91(3), 599-611.

[41] Easterly, W. (2006). *The White Man's Burden: Why the West's Efforts to Aid the Rest Have Done So Much Ill and So Little Good*. New York: The Penguin Press.

[42] Eswaran, M., Kotwal, A. (1990). Implications of credit constraints for risk behaviour in less developed economies. *Oxford Economic Papers*, 42(2), 473-482.

[43] Filmer, D., & L. H. Pritchett. 2001. Estimating Wealth Effects without Expenditure Data or Tears: An Application to Educational Enrollments in States of India. *Demography*, 38 (1): 11532.

[44] Galor, O., & Zeira, J. (1993). Income distribution and macroeconomics. *The Review of Economic Studies*, 60(1), 35-52.

[45] Giesbert L, & Schindler K. 2012. Assets, shocks, and poverty traps in rural Mozambique. *World Development*. 40(8):15941609

[46] Ghatak, M. (2015). Theories of poverty traps and anti-poverty policies. *The World Bank Economic Review*, 29: S77-S105.

[47] Gollin, D. (2014). The lewis model: A 60-year retrospective. *The Journal of Economic Perspectives*, 28(3), 71-88.

[48] Gollin, D., Lagakos, D., & Waugh, M. E. (2013). The agricultural productivity gap. *The Quarterly Journal of Economics*, 129(2), 939-993.

[49] Grosh, M., & J. Baker. (1995). Proxy Means Tests for Targeting Social Programs. *LSMS Working Paper* No. 118. The World Bank, Washington, DC.

[50] Hastie, T., R. J. Tibshirani, & J. Friedman. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer.

[51] Hennig, C. (2007). Cluster-wise assessment of cluster stability. *Computational Statistics & Data Analysis*, 52(1), 258-271.

[52] Herrendorf, B., & Schoellman, T. (2018). Wages, human capital, and barriers to structural transformation. *American Economic Journal: Macroeconomics*, 10(2), 1-23.

[53] Hoddinott, J. (2006). Shocks and their consequences across and within households in rural Zimbabwe. *The Journal of Development Studies*, 42(2), 301-321.

[54] Ikegami, M., Carter, M. R., Barrett, C. B., & Janzen, S. A. 2016. Poverty Traps and the Social Protection Paradox (No. w22714). *National Bureau of Economic Research*.

[55] IRIS Center. 2005. Note on Assessment and Improvement of Tool Accuracy. Poverty Assessment Tools. USAID. Accessed January 2014. http://www.povertytools.org/tools.html.

[56] IRIS Center. 2007. Poverty Assessment Tool Accuracy Submission. USAID/IRIS Tool for Timor-Leste. Poverty Assessment Tools. USAID. Accessed January 2014. http://www.povertytools.org/tools.html.

[57] IRIS Center. 2009. Poverty Assessment Tool Accuracy Submission. USAID/IRIS Tool for Bolivia. Poverty Assessment Tools. USAID. Accessed January 2014. http://www.povertytools.org/tools.html.

[58] IRIS Center. 2012. Poverty Assessment Tool Accuracy Submission. USAID/IRIS Tool for Malawi. Poverty Assessment Tools. USAID. Accessed January 2014. http://www.povertytools.org/tools.html.

[59] Jalan, J., & Ravallion, M. (2002). Geographic poverty traps? A micro model of consumption growth in rural China. *Journal of Applied Econometrics*, 17(4), 329-346.

[60] Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301), 790-794.

[61] Jensen, N. D., Barrett, C. B., & Mude, A. G. (2017). Cash transfers and index insurance: A comparative impact analysis from northern Kenya. *Journal of Development Economics*, 129, 14-28.

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

[63] Kaufman, L. & Rousseeuw, P. J. (1990). *Partitioning Around Medoids Program PAM, in Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley & Sons, Inc., Hoboken, NJ, USA. doi: 10.1002/9780470316801.ch2

[64] Koenker, R. 2005. *Quantile Regression*. Cambridge: Cambridge University Press.

[65] Kraay A, & McKenzie D. (2014). Do poverty traps exist? Assessing the evidence. *Journal of Economic Perspectives*. 28(3):12748

[66] Kshirsagar, V., Wieczorek, J., Ramanathan, S., & Wells, R. (2017). Household poverty classification in data-scarce environments: A machine learning approach. *arXiv preprint* arXiv:1711.06813.

[67] Kwak, S., & Smith, S. C. (2013). Regional agricultural endowments and shifts of poverty trap equilibria: Evidence from Ethiopian panel data. *The Journal of Development Studies*, 49(7), 955-975.

[68] Lagakos, D., Waugh, M. E. (2013). Selection, agriculture, and cross-country productivity differences. *American Economic Review*, 103(2), 948-80.

[69] Lee, D. 2014. Measuring Poverty Using Asset Ownership: Developing a Theory-Driven Asset Index Incorporating Utility and Prices. *Unpublished Job Market Paper.* University of California-Berkeley. Accessed January 2014.

[70] Liaw, A., & M. Wiener. 2002. Classification and regression by randomForest. *R News* 2:1822.

[71] Lin, Y., & Y. Jeon. 2006. Random Forest and Adaptive Nearest Neighbors. *Journal of the American Statistical Association*, 101 (474): 578590.

[72] Lokshin, M., & Sajaia, Z. (2004). Maximum likelihood estimation of endogenous switching regression models. *Stata Journal*, 4, 282-289.

[73] Lybbert, T. J., Barrett, C. B. (2011). Risktaking behavior in the presence of nonconvex asset dynamics. *Economic Inquiry*, 49(4), 982-988.

[74] Lybbert TJ, Barrett CB, Desta S, & Coppock DL. (2004). Stochastic wealth dynamics and risk management among a poor population. *Economic Journal*. 114:75077

[75] Lybbert TJ, Just DR, Barrett CB. (2013). Estimating risk preferences in the presence of bifurcated wealth dynamics: Can we identify static risk aversion amidst dynamic risk responses? *Eur. Rev. Agric. Econ*. 40(2): 36177

[76] Maddala, G. S. (1986). Disequilibrium, self-selection, and switching models. *Handbook of Econometrics*, 3:1633-1688.

[77] Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K.(2017). cluster: Cluster Analysis Basics and Extensions. *R package version 2.0.6*.

[78] Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, 7(1), 77-91.

[79] McBride, L. (2018). Heterogeneous welfare dynamics and structural transformation in Tanzania. Working paper.

[80] McBride, L. & A. Nichols. (2016). Retooling poverty targeting using out-of-sample validation and machine learning. World Bank Economic Review, lhw056.

[81] McCullough, E. B. (2016), Occupational Choice and Agricultural Labor Exits in Sub-Saharan Africa, Working Paper Series N 244, *African Development Bank*, Abidjan, Cte dIvoire.

[82] McCullough, E. B. (2017). Labor productivity and employment gaps in Sub-Saharan Africa. *Food Policy*, (67), 133-152.

[83] McMillan, M. S., Rodrik, D. (2011). Globalization, structural change and productivity growth (No. w17143). *National Bureau of Economic Research*.

[84] Meinshausen, N. 2006. Quantile Regression Forests. *Journal of Machine Learning Research*, 7: 98399.

[85] Meinshausen, N. 2016. quantregForest: Quantile Regression Forests. *R package version 1.3-5*. Available at http://CRAN.R-project.org/package=quantregForest

[86] Menezes, C., Geiss, C., Tressler, J. (1980). Increasing downside risk. *The American Economic Review*, 70(5), 921-932.

[87] Meyer, J. (1987). Two-moment decision models and expected utlitly maximization. *Amer. Economic Rev.*, 421-430.

[88] Mullainathan, S., & Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31(2), 87-106.

[89] Murtazashvili, I., & Wooldridge, J. M. (2016). A control function approach to estimating switching regression models with endogenous explanatory variables and endogenous switching. *Journal of Econometrics*, 190(2), 252-266.

[90] National Bureau of Statistics, United Republic of Tanzania. (2014). Basic information document: National Panel Survey 2012-2013. *World Bank Open Data*. Accessed March 2018 at microdata.worldbank.org/index.php/catalog/2252/download/34053

[91] Narayan, D., & Petesch, P. (2007). *Moving Out of Poverty: Volume 1*. Cross-Disciplinary Perspectives on Mobility. Washington, DC: World Bank and Palgrave Macmillan.

[92] Naschold F. (2012). The poor stay poor: Household asset poverty traps in rural semi-arid India. *World Development*. 40(10):203343

[93] Naschold, F. (2013). Welfare Dynamics in Pakistan and EthiopiaDoes the estimation method matter?. *The Journal of Development Studies*, 49(7), 936-954.

[94] PAT (Poverty Assessment Tool). 2014. Quantifying the Very Poor. *Poverty Assessment Tools Website*. Accessed February 2014. http://www.povertytools.org.

[95] Pratt, J. W. (1964). Risk Aversion in the Small and in the Large. *Econometrica*, 32(1/2), 122-136.

[96] R Development Core Team. 2005. R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. Vienna, Austria.

[97] Ravallion, M., & Q. Wodon (1999). Poor areas, or only poor people?. *Journal of Regional Science* 39, no. 4: 689-711.

[98] Reynolds, A., Richards, G., de la Iglesia, B. & Rayward-Smith, V. (1992). Clustering rules: A comparison of partitioning and hierarchical clustering algorithms. *Journal of Mathematical Modelling and Algorithms* 5, 475504 (http://dx.doi.org/10.1007/s10852-005-9022-1).

[99] Rosenzweig, M. R., Binswanger, H. P. (1993). Wealth, Weather Risk and the Composition and Profitability of Agricultural Investments. *The Economic Journal*, 103(416), 56-78.

[100] Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53-65.

[101] Royston, P., & Altman, D. G. (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. *Applied Statistics*, 429-467.

[102] Royston, P., & Sauerbrei, W. (2008). *Multivariable model-building: a pragmatic approach to regression anaylsis based on fractional polynomials for modelling continuous variables* (Vol. 777). John Wiley & Sons.

[103] Sachs, J. (2005). *The End of Poverty*. New York: Penguin.

[104] Santos, P., & Barrett, C. B. (2016). Heterogeneous wealth dynamics: On the roles of risk and ability (No. w22626). *National Bureau of Economic Research*.

[105] SAS Institute Inc. 2009. SAS/STAT 9.2 Users Guide, Second Edition. *SAS Institute Inc.*, Cary, NC. Accessed May 13, 2012. https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htmtitlepage.htm.

[106] Schreiner, M. 2006. A Simple Poverty Scorecard for Bangladesh, Report to Grameen Foundation USA. *Working Paper*. Accessed 15 February 2016.

[107] StataCorp. (2009). *Stata Statistical Software*: Release 11. College Station, TX: StataCorp LP.

[108] Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2), 411-423.

[109] Timmer, C. P. (1988). The agricultural transformation. *Handbook of Development Economics*, 1, 275-331.

[110] Timmer, C. P. (2002). Agriculture and economic development. *Handbook of Agricultural Economics*, 2, 1487-1546.

[111] USAID MRR. USAID Microenterprise Results Reporting Portal. Accessed December 17, 2014.

[112] Varian, H. 2014. Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives*, 28 (2): 328.

[113] Wager, S., & Athey, S. (2017). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*: 1-15.

[114] WBG (World Bank Group). 2011. Targeting: Safety Nets and Transfers: Proxy Means Testing. Washington, DC: *The World Bank*. Accessed May 2014.

[115] Young, A. (2013). Inequality, the urban-rural gap, and migration. *The Quarterly Journal of Economics*, 128(4), 1727-1785.

[116] Zimmerman FJ, & Carter MR. (2003). Asset smoothing, consumption smoothing and the reproduction of inequality under risk and subsistence constraints. *Journal of Development Economics*. 71(2):2336