Heterogeneous welfare dynamics and structural transformation

Linden McBride [*]

December 7, 2019

Abstract

This paper is driven by the hypothesis that poverty trap-like welfare dynamics play a role in differing returns to assets among different livelihoods. Using a theoretically grounded, data-driven, approach to identifying a livelihood strategy choice set, I estimate livelihood and migration conditioned returns to assets and associated welfare dynamics using a long panel dataset from Kagera, Tanzania. I find that, between 1991 and 2004, a subset of households 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 assets by livelihood strategy, suggesting that households might realize locally increasing returns if they could switch livelihoods. Analysis of welfare dynamics within and across livelihoods does not identify poverty traps but does uncover heterogeneous welfare dynamics and suggest conditional convergence. Although beginning with a flexible framework and employing a data driven strategy, the 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.

Key Words: welfare dynamics, heterogeneity, livelihoods, unsupervised learning, productivity gap, migration, structural transformation

JEL classification: O12, I32, C10, D60, D31, O15

1 Introduction

Two empirical regularities – the returns to labor are greater outside of agriculture than within agriculture (Gollin *et al.* 2014) and the cost of living adjusted consumption is greater in urban than in rural areas (Young 2013) – present a compelling problem for agricultural economists. At first glance, these productivity and consumption gaps suggest that inequality could be addressed and growth spurred by simply correcting the missallocation of factors across sectors and locations (Young 2013, Lakagos & Waugh 2013, Gollin *et al.* 2014, McMillan & Rodrik 2011). However, the problem has resisted such easy resolution.

[*] I gratefully acknowledge insights from Chris Barrett, John Hoddinott, Finn Christensen, Leah Bevis, Jakina Debnam, Yanan Li, Vesall Nourani, Maulik Jagnani, and Kibrom Tafere; however, all errors are my own. 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 1522054. 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. Email: lemcbride@smcm.edu; Website: www.lindenmcbride.com

Gollin *et al.* (2014) attempt to identify the source of the agricultural productivity gap by assessing whether it is driven by 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 determine that their findings are consistent with a story of self-selection wherein those with sufficient skill switch to the non-agricultural sector. A strong case for the self-selection story 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 (2013) finds that those with unobserved skill (correlated with observed education) relocate to the urban environment. 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, and unobserved skill are randomly distributed and there are few barriers to movement among livelihoods, then the observed gaps are simply due to efficient sorting (selection) of labor rather than the consquences of barriers to mobility, market failures, or poverty traps. However, if there are path dependencies to the distribution of ability or the development of skill (which, e.g., Lagakos & Waugh (2013) and Young (2013) proxy for with educational attainment) – and there is strong evidence that human capital development is linked to both geography and parental resources (Chetty *et al.* 2014, 2016, Chetty & Hendren 2018a, 2018b) – then the possibility that multiple equilibria welfare dynamics are playing a role in these gaps cannot be dismissed. In this paper, I hypothesize that poverty trap-like welfare dynamics play a role in differing returns to assets among different livelihoods.

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 sufficiently 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. 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 the estimated welfare dynamics.

In particular, 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. 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 in this paper.

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 suggest the sort of welfare dynamics that give rise to poverty traps, offering 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). My approach additionally 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.

This paper makes several contributions to the productivity gap and poverty traps literatures. With some exceptions, most of the data used for analyses of the productivity gap rely on aggregate data. Gollin *et al.* (2014) and Young (2013) present the first approaches using micro-data, with Gollin *et al.* (2014) relying on LSMS data and Young (2013) on DHS data. In fact, using household level LSMS-ISA data from Tanzania 2010/11, McCullough (2017) finds that the productivity gap is smaller than reported using aggregate data and that at least half of the observed gap is due to fewer labor hours supplied in the agricultural sector (rather than lower productivity per hour-worker in the agricultural sector). 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 agricultural and non-agricultural 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 Gollin *et al.* (2014) and Young (2013) I rely on panel data. Moreover, the span of the KHDS panel survey (1991-2010) allows for a unique, longitudinal look at the returns to assets over time. 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 outcomes that are observed. Finally, I estimate welfare dynamics within and between livelihoods, allowing welfare dynamics to differ[1] for different sectors of the economy.

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 generally assumed in the literature (Timmer 1988, Gollin *et al.* 2014). I find evidence of differences in returns to business, labor, and human capital assets by livelihood strategy and by migration status. In addition, I find evidence of heterogeneous welfare dynamics and conditional convergence; however, the welfare equilibria appear to converge over time, suggesting a catch-up in returns to assets in the agricultural sector. This suggests that the observed welfare and productivity gaps are snapshots of the differentiated returns that emerge as part of the structural trans-

[1] See Appendix for additional details on this contribution.

formation. Finally, 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.

2 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 a household at time t has asset stock vector $\mathbf{A}_t$, with $\mathbf{A}_t \geq 0$; these assets might include labor, land, livestock, other physical capital (such as business and farm assets), and human capital (such as education and health). The asset stock can be used to produce commodity outputs, $o_j, j = 1, ...J$, where j indexes each commodity.

A set of livelihood strategies, $L_k, k = 1, \ldots K$ are available to the household; a given L_k could include a single production technology or combinations of production technologies and is therefore represented as a correspondence between the asset stock vector $\mathbf{A}_t$ and the vector of outputs, $\mathbf{O}_t$ (Equation 1).

$$L_{kt} : \mathbf{A}_t \rightarrow \mathbf{O}_t \tag{1}$$

To illustrate, consider two example livelihood strategies, L_1 and L_2. While L_1 might include both maize farming (using assets such as land and labor to produce the commodity maize) and running a small street food business (using assets such as labor, pots and pans, and sterno oven to produce the commodity *chapati*), L_2 might include maize farming alone.

There exist fixed costs, $FC_{L_{kt}}$, and transactions costs to employing 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, $TC_t(\mathbf{Z}_t, \mathbf{A}_t, \mathbf{E}_t)$, are a function of household characteristics, $\mathbf{Z}_t$, household asset stocks, $\mathbf{A}_t$, and characteristics of the local environment, $\mathbf{E}_t$. 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. With this simplifying assumption, I assume households select their optimal livelihood strategies conditional on associated fixed costs.

The household can either be a net seller, M_j^s, or a net buyer, M_j^b, of a given commodity, where M_j^s, $M_j^b \in \{0, 1\}$; a household can also be autarkic with respect to a commodity, in which case M_j^s, $M_j^b = 0$. A household cannot be both a net seller and net buyer; i.e., there is no case where M_j^s, $M_j^b = 1$.

The household faces market price, p_j, for each commodity it buys and sells; however, the household specific price, p_j^*, is modulated by transactions costs as well as the household's status relative to the market,

$$\begin{gathered} p_{jt}^* = p_{jt} + (M_{jbt} - M_{jst})TC(\mathbf{Z}_t, \mathbf{A}_t, \mathbf{E}_t),\ where\ M_{jbt} \neq M_{jst} \\ p_{jt}^* = p_{jt},\ where\ M_{jbt} = M_{jst} = 0 \end{gathered} \tag{2}$$

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 net revenue has the same properties as a production technology — it is a quasi-concave and monotone mapping from the goods and services sold into net revenues — allowing one to nest market participation decisions within the choice of production technologies. One can think of the fixed 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. Similarly, we can think of the decision to migrate as a technology adoption decision.

The household earns income, y_t, from the sale of commodities it produces using its optimal livelihood strategy, having selected[2] that optimal livelihood strategy conditional on the associated fixed costs,

$$y_t = (\mathbf{p}_t^{*\prime}\mathbf{O}_t | max\,\{L_{1t}, L_{2t}, ...L_{Kt}|FC_{L_{kt}}\}) \tag{3}$$

For all commodities, j, 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),

$$c_{jt} \leq o_{jt} \tag{4}$$

The household maximizes utility over consumption of the vector of agricultural, small enterprise, or labor produced commodities, c_t, as well as other tradables, x_t, that the household cannot produce. The household is subject to budget constraints. Let p_{xt} represent the price of commodities the household cannot produce, let I_t indicate household investments in additional assets at the price, p_{It}. Then the household budget is,

$$\mathbf{p}_{xt}^{\prime}\mathbf{x}_t + \mathbf{p}_t^{*\prime}\mathbf{c}_t + \mathbf{p}_{It}^{\prime}\mathbf{I}_t \leq y_t \tag{5}$$

The asset accumulation law of motion is

$$\mathbf{A}_{t+1} \leq \boldsymbol{\delta}_t^{\prime}\mathbf{A}_t + \mathbf{I}_t \tag{6}$$

where each $\delta_t > 0$ 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 $\mathbf{A}_t \geq -B(\mathbf{A}_t)$ 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.

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

Altogether, the household's dynamic welfare maximization problem can be represented as follows:

$$\begin{gathered}
max_{\mathbf{c}_t,\mathbf{x}_t,I_t}\Sigma_{t=1}^{\infty}\beta^t U(\mathbf{c}_t\mathbf{x}_t)\\
subject\,to:\\
y_t=(\mathbf{p}_t^{*\prime}\mathbf{O}_t|max\,\{L_{1t},L_{2t},....L_{Kt}|FC_{L_{kt}}\})\\
\mathbf{p}_{xt}'\mathbf{x}_t+\mathbf{p}_t^{*\prime}\mathbf{c}_t+\mathbf{p}_{It}'\mathbf{I}_t\le y_t\\
c_{jt}\le o_{jt},j\notin M\\
\mathbf{A}_{t+1}\le\boldsymbol{\delta}_t'\mathbf{A}_t+\mathbf{I}_t\\
\mathbf{A}_t\ge -B(\mathbf{A}_t)
\end{gathered}\tag{7}$$

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 ($\mathbf{A}_t \geq 0$), 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, conditional on livelihood, will be estimated via Taylor series expansion of the reduced form expression mapping welfare to assets in Equation 8,

$$exp_{it}=L_{kit}(\mathbf{A}_{it})+\epsilon_{it}\tag{8}$$

where exp_{it} represents household i's consumption expenditures, the best available representation of welfare in the KHDS data, at time t, and L_{kit} represents the household's livelihood conditioned returns to their current asset holdings. If the data are consistent with a multiple equilibria welfare dynamics scenario, we should observe the marginal returns to assets differ 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 Data and region of study

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 rates—92 percent of baseline households were tracked through to 2010—and they cover a period when Tanzania is undergoing structural transformation (Christiaensen, De Weerdt, & Todo 2013). The initial 1991 survey 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), and De Weerdt & Hirvonen (2016).

Due to the unique longitudinal panel data set collected there, the region of study in this paper, Kagera, Tanzania, has been extensively studied. 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 1991/94—in particular, initial stocks of land and human capital as well as location factors such as "the degree of connectedness of the place of residence"—determine 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 diversification, migration, and connectedness; my final results support this finding.

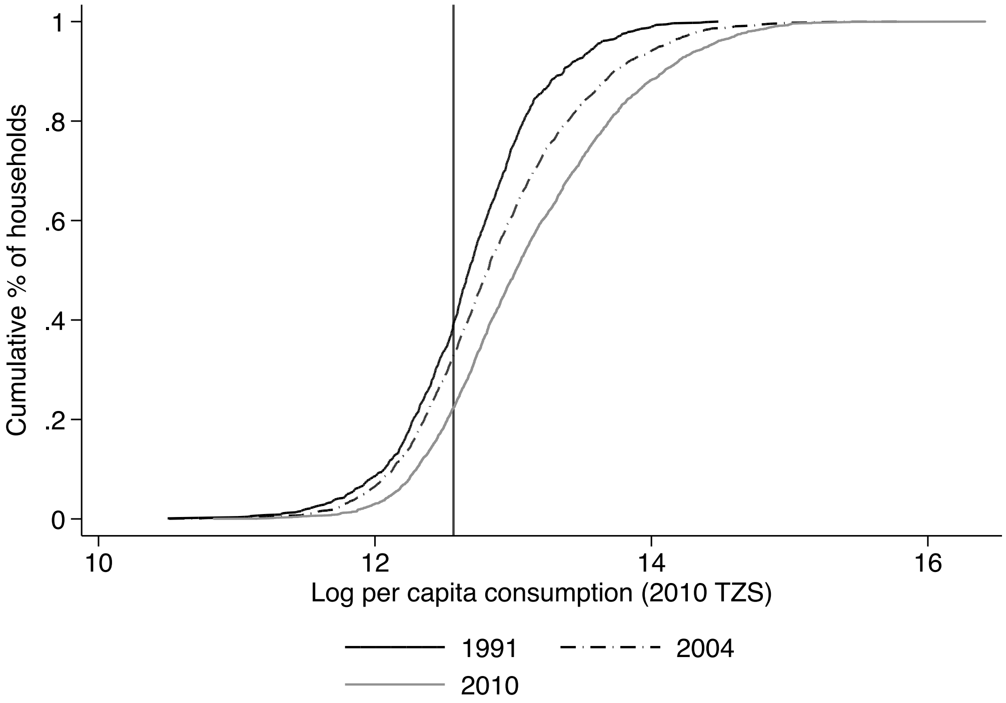
Households from and within Kagera have enjoyed growth in consumption over the course of the longitudinal study. Cumulative densities of per capita consumption in 1991, 2004, and 2010, using data in 2010 TZS value that has been deflated using a regional price index, are presented in Figure 1. The horizontal 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 transitioning out of poverty by 2010 (Table 1); however, 30 percent of the 1991 poor remain poor in 2010.

Table 1: 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 1: Cumulative consumption 1991, 2004, 2010



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

Table 2 suggests that financial market failures are a possible constraint in this setting: of those individuals starting businesses [3] in 2010, 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 2 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.

4 Empirical approach

My empirical approach is as follows, I: 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] 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 2: 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

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 for determining the optimal k. The cluster analysis procedure and gap statistic method are detailed in the Appendix.

The estimated clusters are displayed in Figure 2 and 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] To my knowledge, one paper exists: 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

So as to produce a set of livelihood strategies based on land and labor allocations and asset holdings in line with the theoretical model described above, I perform the cluster analysis over variables indicating household land and labor allocations 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. Note that 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. 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.

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. A vector of time-varying individual 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 the individual’s allocated labor hours per week, total land area, the log value of business assets, total TLU, and the individual’s years of education. In Equation 9, i indexes individual, t indexes time, l indexes location, and d indexes the productive assets included in the estimation. Standard errors are clustered at the household level.

$$\begin{aligned} exp_{itl} = \Sigma_d^5 \beta_d A_{itdl} + \frac{1}{2}\Sigma_d^5 \beta_{dd} A_{itdl}^2 + \Sigma_d^5 \Sigma_j^5 \beta_{dj} A_{itdl} A_{itjl} + \boldsymbol{\beta}_x \boldsymbol{X}_{it} + \\ \Sigma_d^5 \gamma_d A_{itdl} L_{it} + \frac{1}{2}\Sigma_d^5 \gamma_{dd} A_{itdl}^2 L_{it} + \Sigma_d^5 \Sigma_j^5 \gamma_{djl} A_{itdl} A_{itjl} L_{it} + \boldsymbol{\gamma}_x \boldsymbol{X}_{it} L_{it} + w_l + \alpha_i + \psi_t + \epsilon_{itl} \end{aligned} \tag{9}$$

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

$$m(A_r) = \hat{\beta_r} + \hat{\beta_{rr}} A_r + \Sigma_s^4 \hat{\beta_{rs}} \bar{A}_s + \hat{\gamma_r} + \hat{\gamma_{rr}} A_r L + \Sigma_s^4 \hat{\gamma_{rs}} \bar{A}_s L \tag{10}$$

where r indexes the support of the asset of interest and $\bar{A}$ indicates that a variable is being held at its mean. Standard errors are produced using the delta method.

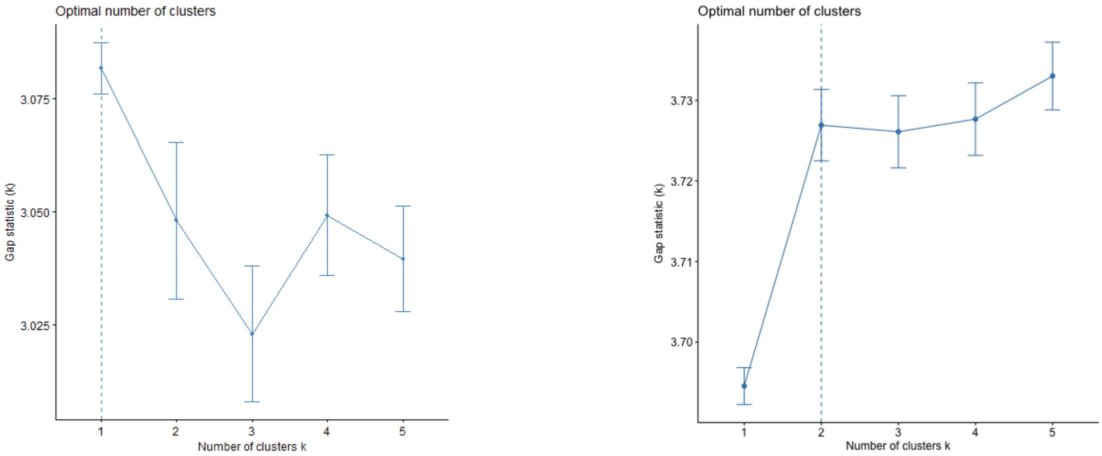
If marginal returns to assets differ by livelihood strategy, such that livelihoods requiring greater fixed costs offer higher returns to the same asset holdings, we would find locally increasing returns in any shift from a low return livelihood to a higher return livelihood. Such locally increasing returns combined with credit constraints would suggest multiple equilibria welfare dynamics.

Finally, 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

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 $k = 5$.

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

Figure 2: Optimal number of clusters 1991 ($N = 915, v = 99$) and 2004 ($N = 2774, v = 94$) using the gap statistic



(a) 1991

(b) 2004

2009) to graph these dynamics, regressing logged consumption in the later period on that of the earlier period.

5 Results

5.1 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 livelihoods were identified in the 2004 data. From Figure 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. 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. Summary statistics for, and a plot of, the 2004 clusters are available in Appendix Table 3 and Figure 13. The cluster plot in Figure 13, 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, and greater livestock assets, and allocate more land to every crop (excepting rice) than do those in cluster two (Appendix Table 3). Therefore I'll 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, I'll refer to cluster two as the wage labor/entrepreneur livelihood strategy.

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 3). 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 strategies—in 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 labor/entrepreneurs have no pension, no dowry, and do not play the lottery, perhaps reflecting the fact that these households have younger[7] 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 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 to some extent below via fixed effects estimation of the returns to assets.

Looking back to 1991 [8] 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 building assets (3.2 times greater in

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

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

wage labor/entrepreneur group), land holdings (0.55 acres smaller in wage labor/entrepreneur group), and financial assets—the 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.

5.2 Heterogenous and locally increasing returns

Marginal returns in consumption to each asset by livelihood strategy are shown in Figures 3 through 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), labor allocations (Figure 4), and human capital assets (Figure 7).

Marginal returns to business assets (Figure 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 allocated labor (Figure 4) than do the farmers.

While it appears that those in the wage labor/entrepreneur livelihood experience increasing returns to their land holdings (Figure 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 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 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: Marginal returns to business assets by livelihood strategy

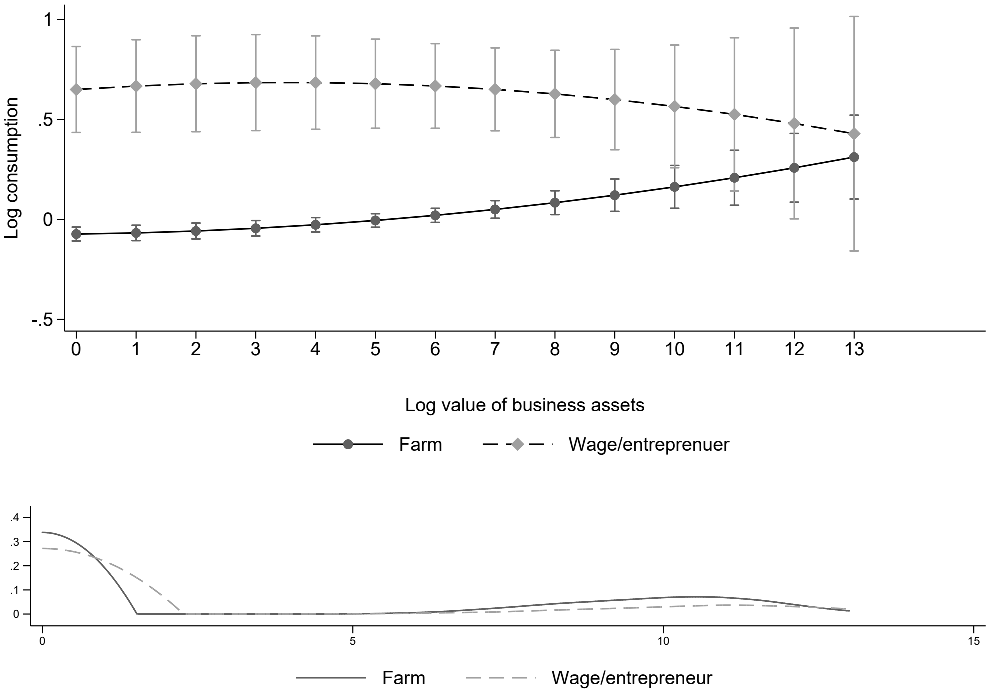


Figure 4: Marginal returns to labor by livelihood strategy

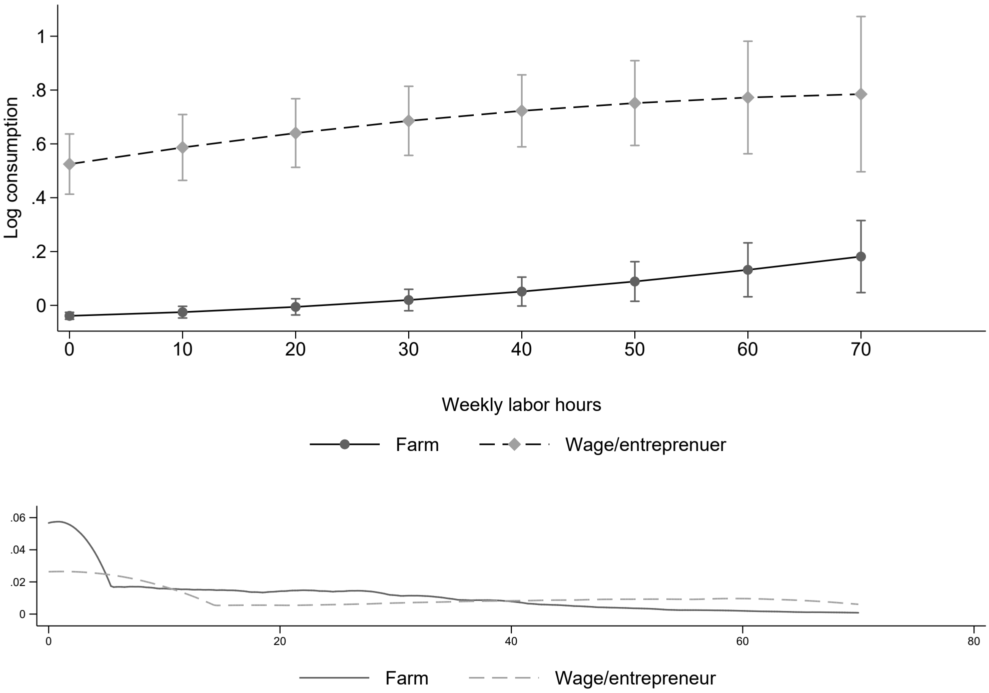


Figure 5: Marginal returns to land holdings by livelihood strategy

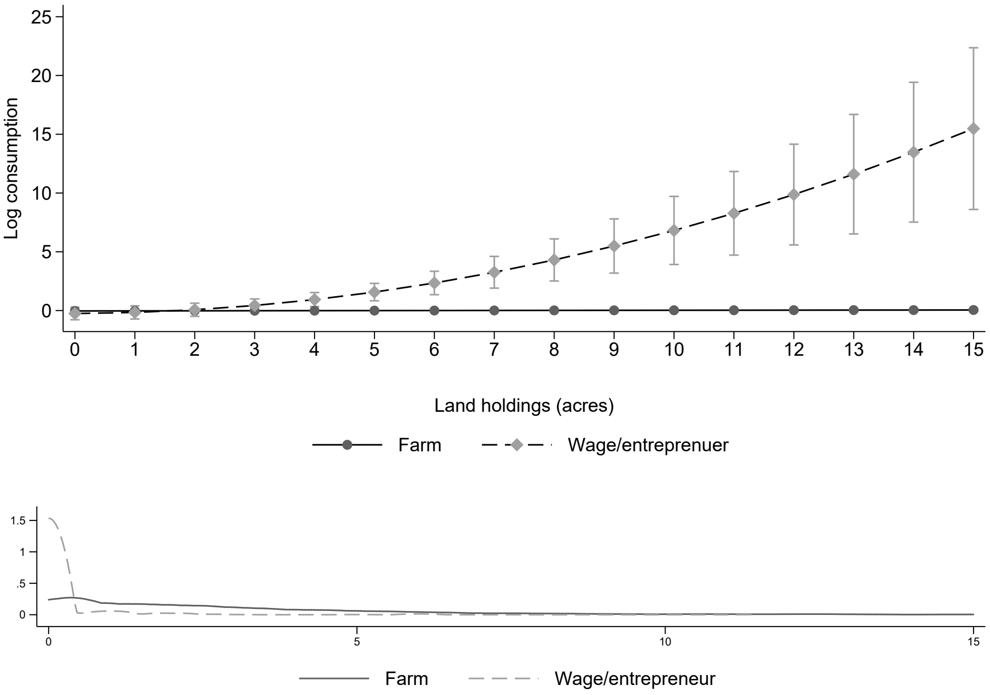


Figure 6: Marginal returns to livestock holdings by livelihood strategy

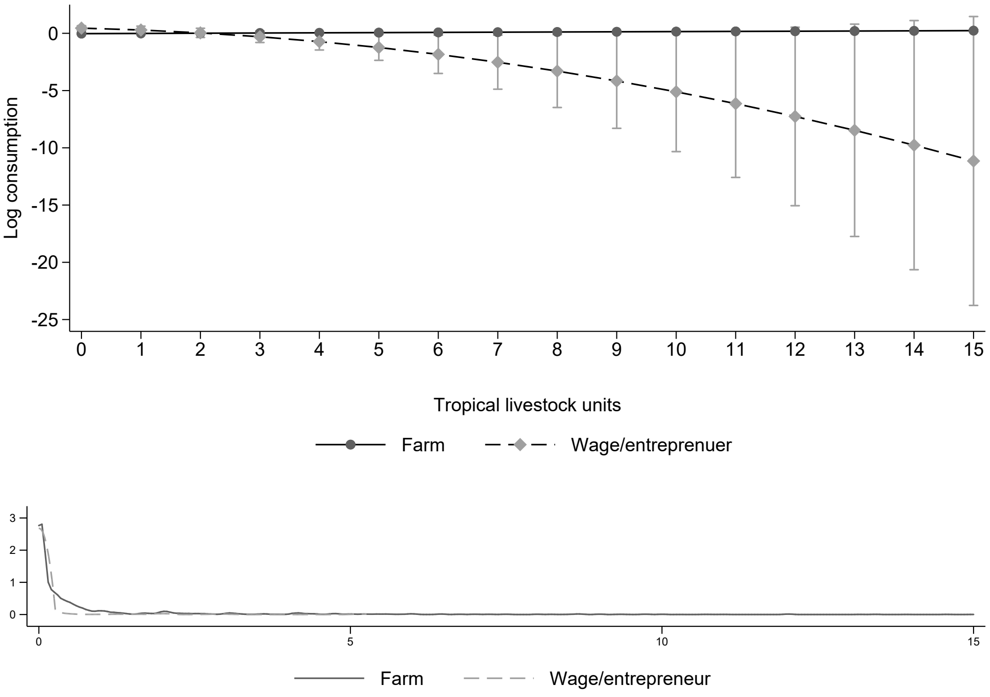
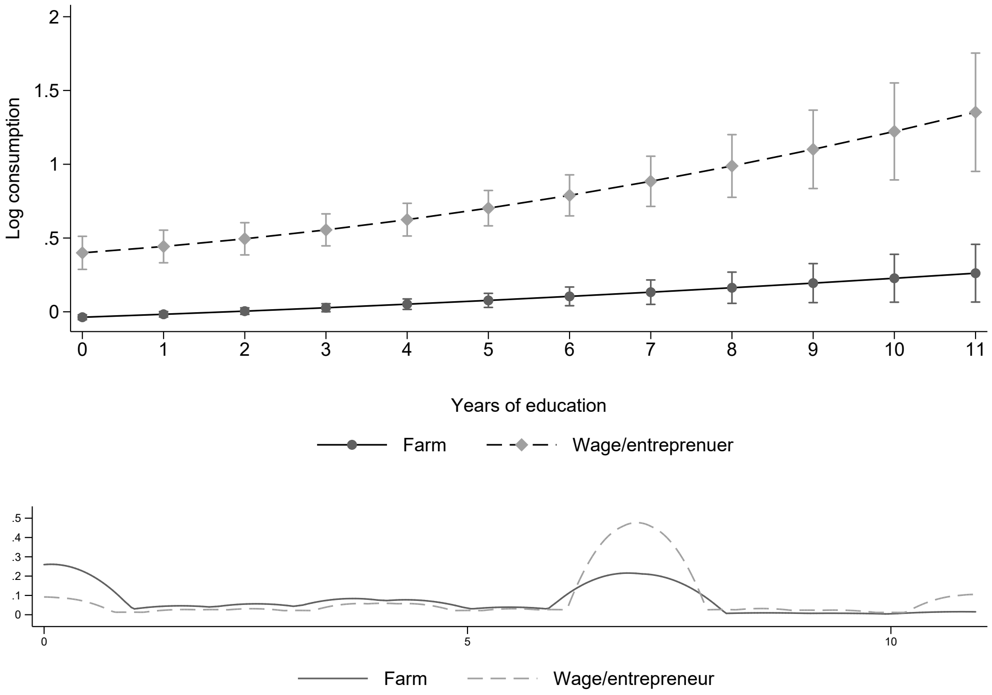


Figure 7: Marginal returns to education by livelihood strategy



Overall, the estimated marginal returns to assets by livelihood strategy suggest that, holding all else constant,[9] 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 the livelihood summary statistics as well as 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 9 with three livelihoods instead of the original two: Remain & Farm, Move & Farm, and Move & Wage/Entrepreneur. There are an insufficient number of observations of Remain & Wage/Entrepreneur to produce estimates for this group. The results are presented in Figures 8 through 10.

The returns to assets for those who move and switch livelihoods (Move & Wage/Entrepreneur) are greater than for those who remain in farming, regardless of whether or not they have moved. Comparing the estimated returns to assets by livelihood (Figures 3 through 7) with those interacted with migration (Figures 8 through 10) suggests that most of the differences in returns are driven by shifts in 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. 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

[9] It is important to note that the returns to livelihood shifts cannot be interpreted causally. The long panel data as well as the spell length between panel waves means that I may be observing the return to livelihood choices following a failed livelihood switch or a failed migration attempt from which the individual has since returned (to initial livelihood and/or location). In addition, the assets under consideration, e.g.– human capital acquisition and labor hour allocations – are endogenous to returns.

(2013), my findings are consistent with a selection story in that those with higher education in 2004 are found in the off-farm livelihood, where they enjoy higher returns and greater consumptions. However, as noted above, those households that ultimately switched livelihoods by 2004 already displayed higher levels of education and greater asset holdings in 1991 than did those households that did not switch livelihoods. Moreover, the marginal returns by livelihood estimates suggest locally increasing returns between livelihoods, a necessary but not sufficient condition for multiple equilibria welfare dynamics. Therefore, I'll next estimate welfare dynamics by livelihood strategy.

Figure 8: Marginal returns to business assets by migration status

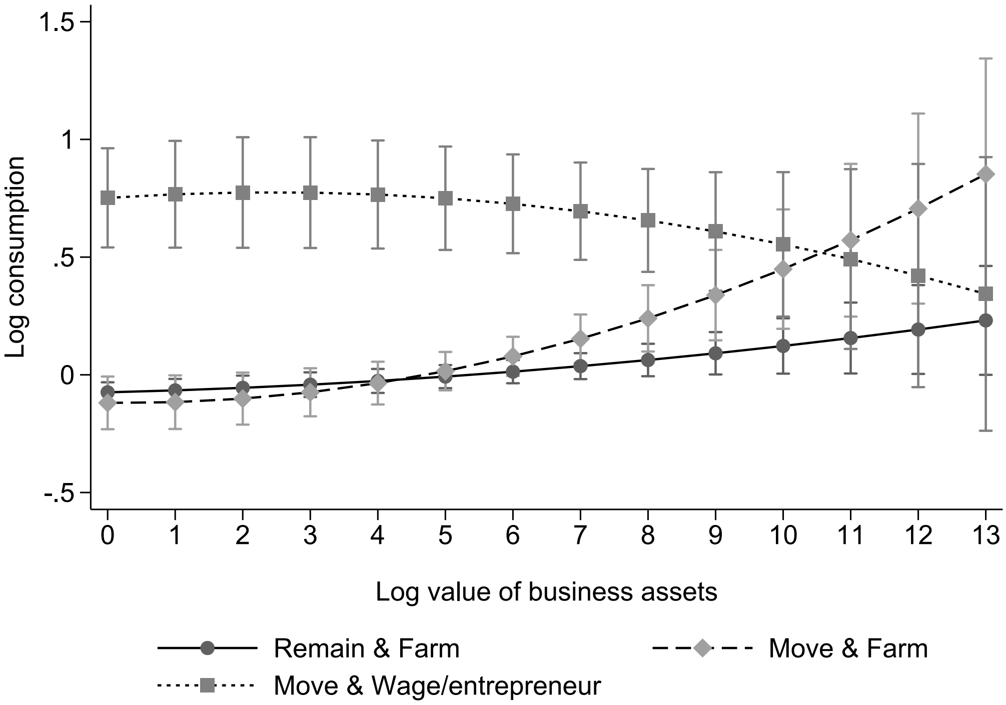


Figure 9: Marginal returns to labor by migration status

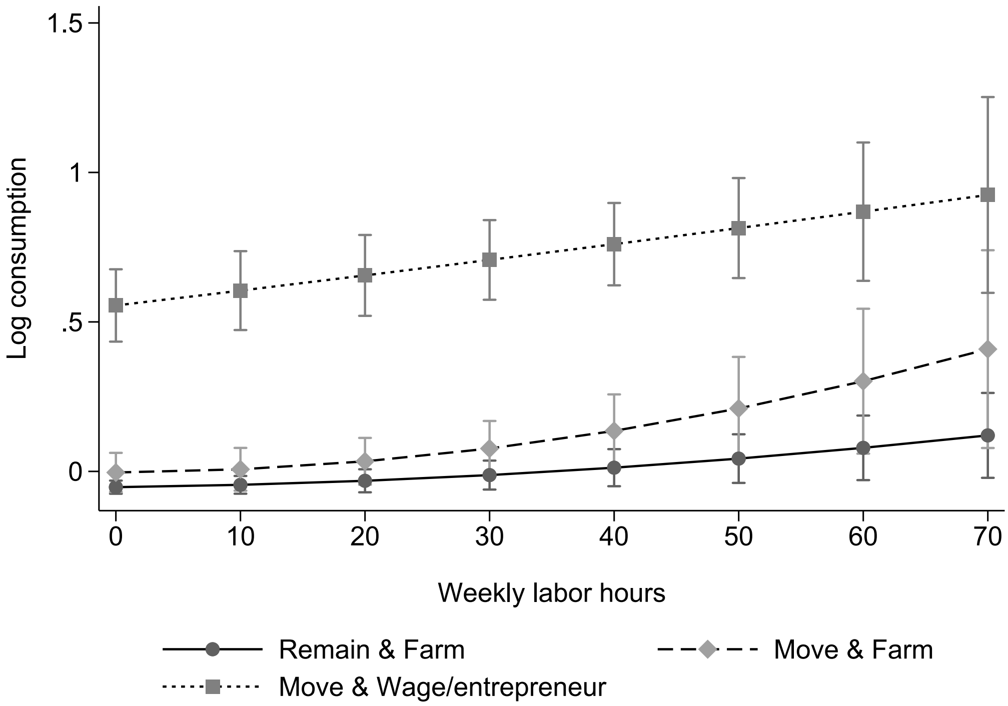
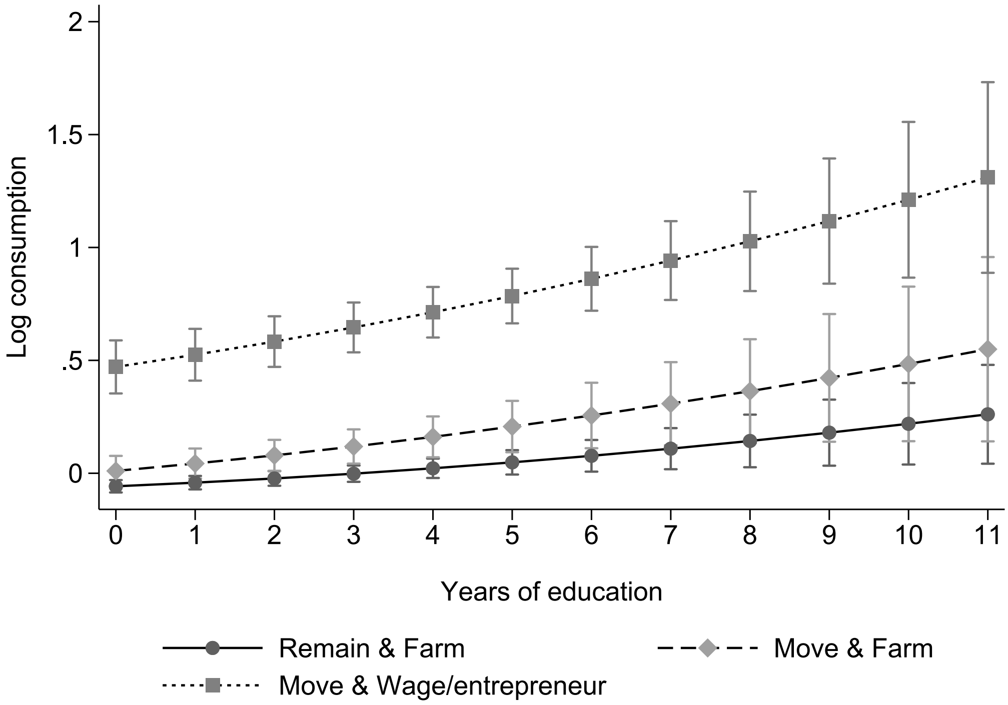


Figure 10: Marginal returns to education by migration status



5.3 Livelihood group welfare dynamics

Although the average household in the data is on a non-poor consumption dynamic path (Figure 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 11a) and 2010 (Figure

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 12a) with those of 2004 to 2010 (Figure 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 11: Mean consumption dynamics (a) 1991 to 2004 (b) 2004 to 2010

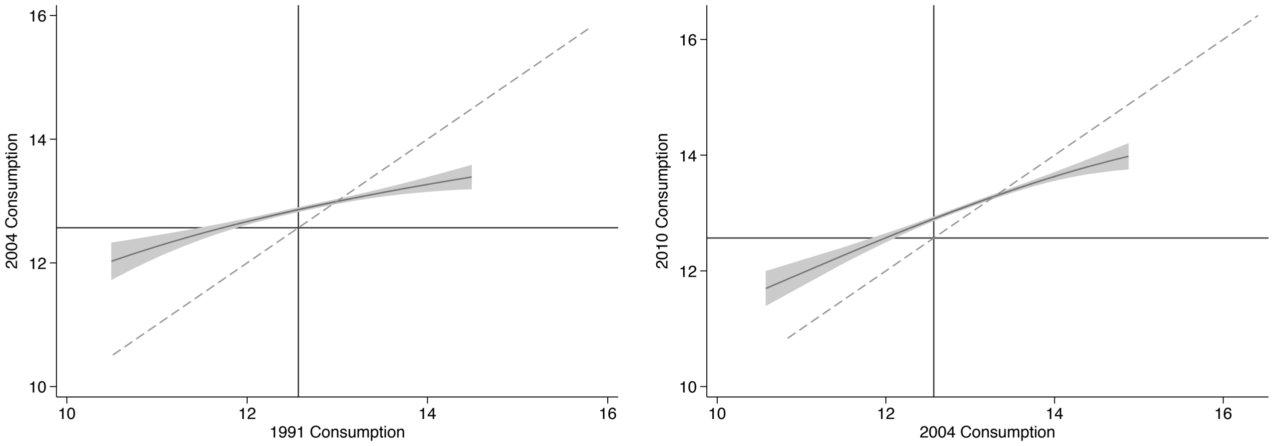
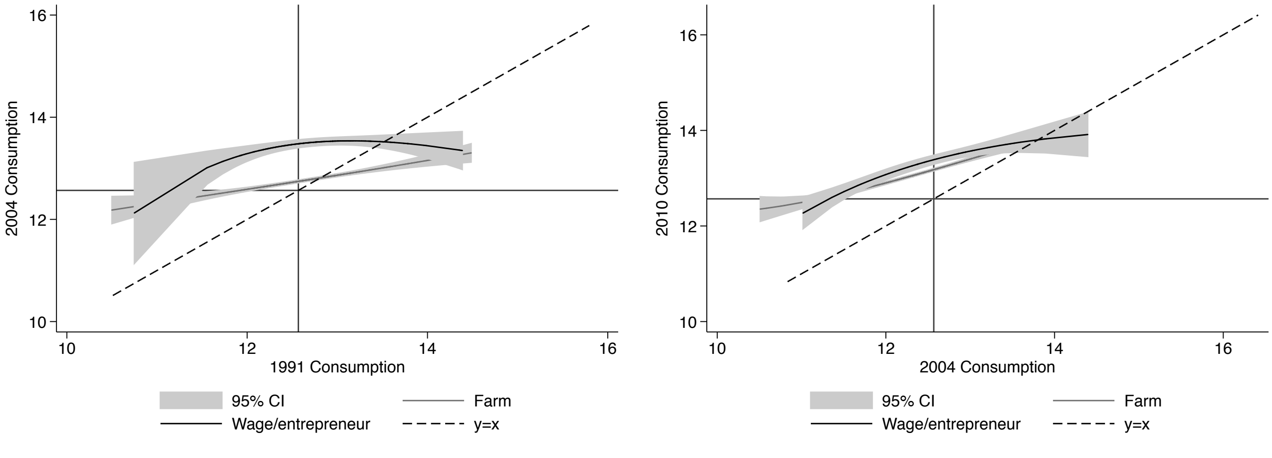


Figure 12: Consumption dynamics by 2004 livelihood strategy (a) 1991 to 2004 (b) 2004 to 2010



6 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 in Kagera, Tanzania. Estimated returns to key assets differ by livelihood, suggesting locally increasing returns in the move from one livelihood strategy to another. In line with the productivity and consumption gap literature (Young 2013, Gollin *et al.* 2014, Herrendorf & Schoellman 2018, Lakagos & Waugh 2013), the educated appear to select into the off-farm sector. I additionally find that those selecting into the off-farm sector in 2004 had greater asset holdings in 1991, suggesting bifurcating welfare dynamics.

However, the asset content of each of the identified livelihood strategies is diverse, reflecting mobility. In fact, according to the observed welfare dynamics, neither livelihood group is trapped in poverty; but when heterogeneity in livelihood strategies is allowed for in the estimation of welfare dynamics, conditional convergence is observed. By 2010, equilibrium welfare of 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, the 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. Finally, the findings suggest that livelihood change plays a greater role in increasing consumption than does geographic change.

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 that estimation of welfare dynamics at population means, without allowing for heterogeneity to emerge, masks policy relevant findings.

The absence of 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 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, 2017, Wager & Athey 2017) and Chernozhukov *et al.* (2018); this is an objective of my future work.

Finally, how can we reconcile the observed differences in returns to assets between livelihoods in the (likely) presence of market failures – 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

[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 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.

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.

References

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):226–47

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

Athey, S., & Imbens, G. W. (2017). The Econometrics of Randomized Experiments. In *Handbook of Economic Field Experiments* (Vol. 1, pp. 73-140). North-Holland.

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

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

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

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_docs/PNACL435.pdf`

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.

Barrett, C.B., Garg, T., & McBride, L. (2016). Well-being dynamics and poverty traps. *Annual Review of Resource Economics*, 8(1).

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

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.

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).

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.

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

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

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.

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.

Chetty, R., & Hendren, N. (2018a). The impacts of neighborhoods on intergenerational mobility I: Childhood exposure effects. *The Quarterly Journal of Economics*, 133(3), 1107-1162.

Chetty, R., & Hendren, N. (2018b). The impacts of neighborhoods on intergenerational mobility II: County-level estimates. *The Quarterly Journal of Economics*, 133(3), 1163-1228.

Chetty, R., Hendren, N., & Katz, L. F. (2016). The effects of exposure to better neighborhoods on children: New evidence from the Moving to Opportunity experiment. *American Economic Review*, 106(4), 855-902.

Chetty, R., Hendren, N., Kline, P., & Saez, E. (2014). Where is the land of opportunity? The geography of intergenerational mobility in the United States. *The Quarterly Journal of Economics*, 129(4), 1553-1623.

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.

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

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.

Chernozhukov, V., Demirer, M., Duflo, E., & Fernandez-Val, I. (2018). Generic machine learning inference on heterogenous treatment effects in randomized experiments. *National Bureau of Economic Research*, No. w24678.

Deaton, A. (1991). Saving and Liquidity Constraints. *Econometrica: Journal of the Econometric Society*, 1221-1248.

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

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.

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

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

De Weerdt, J, K Beegle, H Lilleør, 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.

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

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.

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

Giesbert L, & Schindler K. 2012. Assets, shocks, and poverty traps in rural Mozambique. *World Development.* 40(8):1594–1609

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

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

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

Hastie, T., R. Tibshirani & J. Friedman. (2009). *The Elements of Statistical Learning* (2nd Edition). New York: Springer-Verlag.

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

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

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

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.*

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.

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.

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

Kraay A, & McKenzie D. (2014). Do poverty traps exist? Assessing the evidence. *Journal of Economic Perspectives.* 28(3):127–48

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.

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

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

Lybbert TJ, Barrett CB, Desta S, & Coppock DL. (2004). Stochastic wealth dynamics and risk management among a poor population. *Economic Journal.* 114:750–77

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

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

McCullough, E. B. (2016), Occupational Choice and Agricultural Labor Exits in Sub-Saharan Africa, Working Paper Series N 244, *African Development Bank*, Abidjan, Côte d'Ivoire.

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

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

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.

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

Naschold F. (2012). "The poor stay poor": Household asset poverty traps in rural semi-arid India. *World Development.* 40(10):2033–43

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

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, 475–504 (http://dx.doi.org/10.1007/s10852-005-9022-1).

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

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.

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

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

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

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.

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

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

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

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

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):233–6

A Appendix

Table 3: Livelihoods 2004

	Cluster one (n=2216)		Cluster two (n=558)		Difference	
Variable	Mean	Std. Dev.	Mean	Std. Dev.	p-value	95% sgnf
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	*
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	
pangas	1.23	0.71	0.07	0.27	0.000	*
mundu	0.16	0.47	0.00	0.04	0.000	*

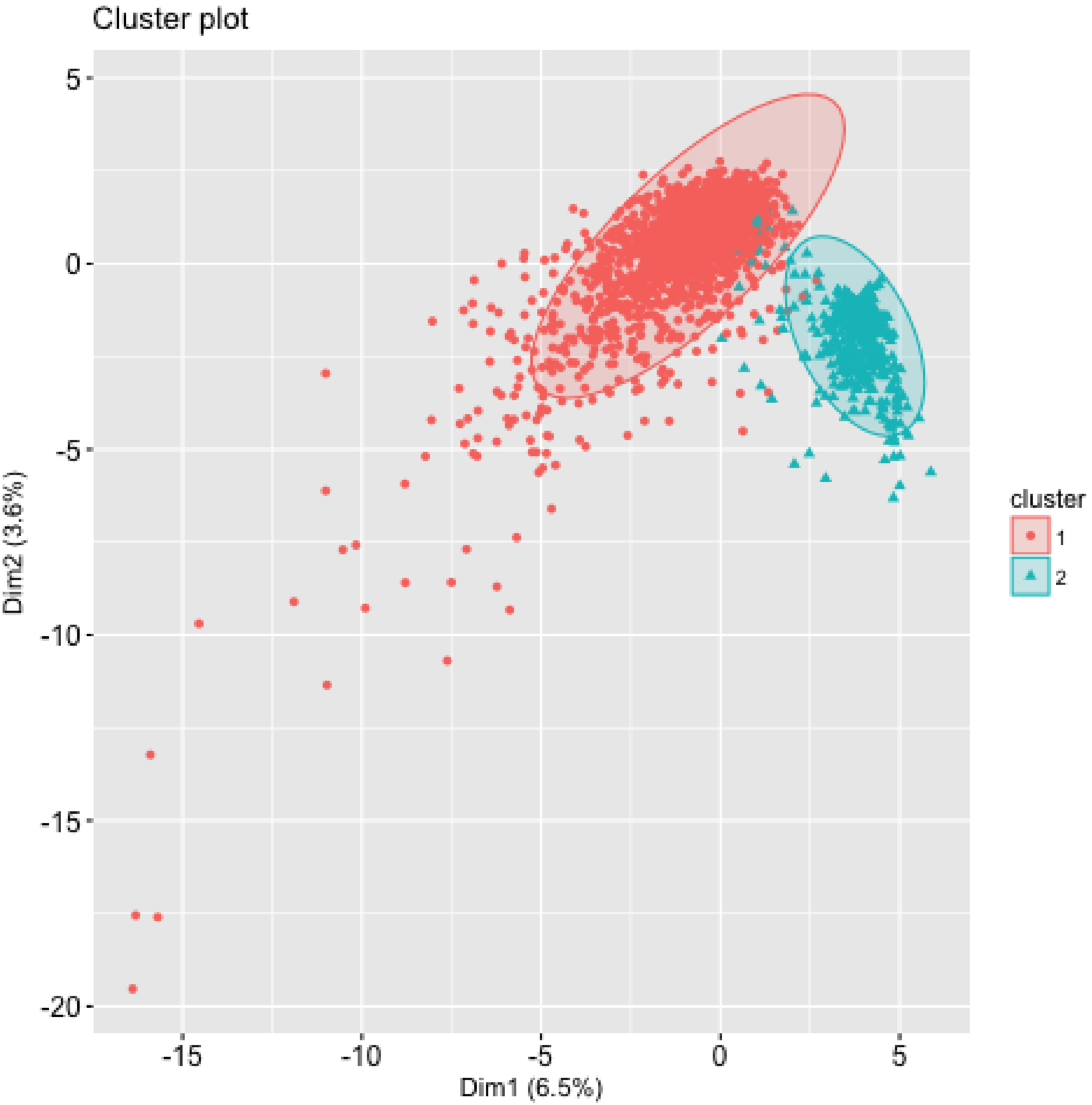
Table 3 continued from previous page

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	
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	*
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	*

Table 3 continued from previous page

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 13: 2004 clusters



A.1 Heterogeneous 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 depending on whether 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.

A.2 Cluster analysis and gap statistic

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^r_{kb}. 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,

$$gap(k) = \frac{1}{B}\Sigma_b log(W^r_{kb}) - log(W_k) \tag{11}$$

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

$$gap(k) \geq gap(k+1) - s_{k+1} \tag{12}$$

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

$$s_k = \left[\frac{1}{B}\Sigma_b(log(W^r_{kb}) - \frac{1}{B}\Sigma_b log(W^r_{kb}))^2\right]^{\frac{1}{2}} \sqrt{1 + \frac{1}{B}} \tag{13}$$

The first term of Equation 13 is the standard deviation of the B bootstrapped W^r_{kb}; 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 14, by iteratively choosing cluster medoids i_k (Hastie *et al.* 2009),

$$min_{C,\{i_k\}^{k^*}_1} \Sigma^{k^*}_1 \Sigma_{C(i)=k} d_{ii_k} \tag{14}$$

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.