Nutrition Demand, Subsistence Farming, and Agricultural Productivity

Stepan Gordeev*

November 4, 2021

click for latest version

Abstract

Subsistence agriculture is prevalent in many of the poorest countries. I propose nutrition demand as a mechanism that drives the production decisions of subsistence farmers and ultimately contributes to low aggregate agricultural productivity. I explore this mechanism in a model of farm-operating households facing explicit caloric needs and costly domestic trade, and test the model's predictions on Malawian household-level data. In the model and in the data, the smallest farmers focus their consumption on obtaining calories and specialize their production in unsold staple crops; medium farmers diversify both their diet and their subsistence production; the largest farmers shift consumption to purchased goods by producing and selling marketable farm products. I quantify the aggregate implications of this farm-level product choice using the model. It suggests that lowering trade frictions enough for the average share of output sold by farmers to reach even 50% would make the country's agricultural sector 42% more productive. Half of this increase is caused by the mechanically reduced erosion of output, and the other half by a better alignment of individual farmers' product choice with their comparative advantage rather than their family's nutritional needs or food preferences.

Keywords: agriculture, nutrition, productivity, trade costs

JEL classification: F11, I15, O12, O13, O18, O47, Q12, R40

I would like to thank Yan Bai for her guidance and support, as well as George Alessandria, Travis Baseler, and Mark Bils for their direction and countless suggestions. I am also grateful to Tillmann von Carnap, Gaston Chaumont, Nataliya Gimpelson, Kristina Manysheva, Roman Merga, Walter Steingress, and the participants of AFES21, NCDE21, YES, the international reading group and the lunch seminar at the University of Rochester for their comments and suggestions.

*Department of Economics, University of Rochester, sgordeev@ur.rochester.edu

1 Introduction

Some of the poorest countries in the world are still dominated by agriculture, which tends to be extremely unproductive. In Malawi, for example, 76% of workers labor in the agricultural sector, yet they produce only 23% of the nation's GDP[1]. The low agricultural productivity in low-income countries is puzzling but crucial in understanding the enormous income differences across countries (Gollin et al., 2014).

A frequent feature of unproductive agricultural sectors is the prevalence of subsistence farming. This condition arises when severe trading frictions make it difficult for farmers to sell their agricultural output and to buy the food they want with the revenue, forcing households to rely on their own production, not the market, for the food they need. As an illustration, I find that in Malawi, over three quarters of all households operate their own farm, while only 11% of these farmers sell most of their output.

In this paper, I explore whether the production decisions of subsistence farmers are relevant for understanding the aggregate agricultural productivity level of low-income countries. I propose nutrition demand by farmers themselves as a force that can explain much of subsistence farm production behavior and ultimately contribute to the low agricultural productivity of the whole economy.

In an environment where subsistence behavior at the level of individual farms is common, farmers may not choose to specialize in their profit-maximizing comparative advantage product, but rather choose farm products and the degree of specialization or diversification that directly satisfy their own family's demand for food, which is in large part a function of the family's nutritional needs. This interaction of trading obstacles and nutritional concerns of the farmers prevents them from fully exploiting the gains from specialization and trade, thus leaving the aggregate agricultural productivity of the economy depressed.

I use a government-run survey in Malawi that provides rich household-level consumption and production data. Most of these households operate their own farm. I find that there is significant subsistence behavior even at the level of individual households. Most sell almost none of their agricultural output and rely on their farm as an important source of food. This self-oriented production be-

[1] World Development Indicators

havior suggests that many farmers are unable to specialize in their comparative advantage and to commercialize the farm.

Motivated by this, I develop a model of household food consumption and production in a subsistence environment. Each household can engage in farming and jointly makes consumption, production, and trading decisions over many agricultural goods, which are heterogeneous in productivity and caloric content. In addition to having standard CES preferences over these agricultural goods as well as a single manufactured good, each household also faces a welfare *caloric deviation penalty* for significantly under- or overshooting its dietary energy requirement.

These food preferences make the household's consumption allocation across goods depend on the size of its farm and income relative to its caloric requirement. In the presence of a proportional trading cost, the nutritional situation of the household furthermore becomes relevant for what it ends up producing on its farm. The poorest farmers struggle to satisfy their most basic need for dietary energy, consequently specializing both their consumption and production in the most efficient sources of calories. Farmers that are able to cover most of their caloric needs can afford to diversify their diet to satisfy their love of variety in food, which they partly achieve by diversifying their production. Finally, the largest farmers easily satisfy their caloric needs and love of variety, permitting them to increasingly shift their consumption to purchased non-food goods, which requires growing and selling their comparative advantage product.

I calibrate the model to match several key moments from the survey of Malawian households. Model households are heterogeneous in their land, non-farm income, and good-specific productivity draws. They consume multiple agricultural goods whose caloric densities, taste weights, and productivity distributions are estimated in the data, and choose which good(s) to produce on their farm.

I investigate household behavior in the model and test the model's predictions in the data, starting with food consumption behavior. Both in the model and in the data, households with the smallest farms (relative to the family's caloric needs) have the most calorie-dominated and least diverse diets. Larger farms consume barely more calories—the empirical output elasticity of dietary energy intake, also targeted in the calibrated model, is just 0.09—but significantly more diverse (and, in the data, nutrient-rich) diets.

Next, I test the predictions of the model on selling behavior. Model households only ever sell at most one good—their comparative advantage one. However, unless trade costs are low enough to permit full specialization, they also produce several other goods for personal consumption. Likewise, in the data, farms' selling is far more specialized than their overall production: 69% of all sellers sell just one good, but only 9% produce just one. Furthermore, as the model predicts, survey households whose location offers better market access have more specialized production.

Both in the model and in the data, larger farmers are more likely to be sellers and sell more on average. The reason for this behavior in the model is that larger farmers are less calorically constrained and want to shift consumption to manufactured goods offered on the market, which require revenue to purchase.

Finally, the model predicts scale-dependent farm product choice that I also observe in the data. Small farms in the model specialize heavily in the most calorically-productive good they can grow: in the data, small farms specialize in maize, the dominant staple of the region. In the model and in the data, medium farms distribute their subsistence production more evenly across multiple edible products. At last, large model farms shift production to their comparative advantage goods and sell most of the output: in the data, large farms increasingly focus production on goods that end up sold. Thus, only the largest farmers in the model and in the data are market-oriented: small and medium ones target their own dietary needs with their farm product choice.

I use the model to quantify the importance of nutrition demand-driven product choice by subsistence farmers on the aggregate agricultural productivity. The model suggests that a reduction in trade costs sufficient to make Malawian farmers just balance between subsistence and commercialization (an increase in the average share of farm output sold from the currently observed 16% to a counterfactual 50%) would raise the aggregate agricultural productivity of the economy by 42%, realizing over half of productivity gains promised by a completely costless domestic trade. Just over half of the aggregate productivity gain happens because falling trade costs allow farmers to better align their product choice with their individual comparative advantages rather than with their individual nutritional needs or preferences for dietary diversity. The productivity and the consumption of the smallest farmers respond the most to this reduction, since their product choice realignment is the strongest. The remaining half of the aggregate productivity gain is caused by a mechanical reduction in trade cost losses happening between the harvesting of a crop and its ultimate consumption (if the two do not take place on the same farm). Thus, the model suggests that the limited extent of agricultural trade between Malawian farmers is a significant drag on the productivity of the agricultural sector, and the misalignment between farm product choice and farm comparative advantage is a big contributor to that.

Background. The relevance of subsistence farming for aggregate agricultural productivity is the subject of a growing literature on the interplay between subsistence farming, trade costs, and agricultural production. Gollin and Rogerson (2014) use a two-sector, three-region model with a subsistence requirement in food to show how transportation costs between regions of a country can generate subsistence behavior in remote regions and keep their agricultural productivity low. Rivera-Padilla (2020) argues that the low agricultural productivity in developing countries is in large part driven by low productivity in staples and the high share of staples in production, and develops a two-region model with two agricultural goods (a staple crop and a cash crop) and a subsistence requirement in staples to show that trade costs depress agricultural productivity by distorting crop choice in favor of staples. Sotelo (2020) develops a multi-region, multi-crop model where land within each region is heterogeneous, leading the region's representative farmer to allocate crops between plots according to their comparative advantage, but trade across regions is costly. He uses the model to evaluate the effects of Peru's road paving plans on agricultural productivity and farmer incomes. Kebede (2020a) builds on this Ricardian framework to obtain a model of crop allocation across heterogeneous plots at village level and uses it to evaluate the welfare effects of a large rural road-building project in Ethiopia.

I contribute to this literature firstly by exploring subsistence at the level of individual farms, rather than villages or regions. Using Malawian household-level data, I highlight significant subsistence behavior even at this disaggregated level and document scale-dependent product choice by Malawian farmers. For modeling purposes, this implies that agents may be unable to split production among multiple producers according to their comparative advantage: each production unit has to tailor production to its *own* demand, not to the demand of a representative consumer at a higher level of aggregation. Secondly, I propose demand for nutrition as a mechanism that can explain the heterogeneous production behavior of individual farmers observed in the data and can contribute to keeping the aggregate agricultural productivity low. Finally, I build a model in which farmers have explicit caloric needs, which generate a tradeoff between dietary energy, dietary diversity, and non-food consumption; this tradeoff is key in the model's ability to rationalize several salient features of subsistence farm behavior.

I also contribute to the literature on forces that can affect farm product choice in low-income countries. Blanco and Raurich (2019) explore the reallocation of farmland toward capital-intensive crops along the path of development. Allen and Atkin (2016) show that while trade liberalization increases farmer revenue volatility, farmers respond by reallocating resources toward less risky crops. The paper most related to mine is the aforementioned work by Rivera-Padilla (2020), in which he shows that trade costs induce farmers to reallocate production toward a staple crop, which is needed to satisfy a subsistence constraint and also has lower trade costs compared to a cash crop. I contribute to this literature by developing a model with explicit nutritional needs, in particular the demand for dietary energy and the demand for dietary diversity, and showing that these needs can explain several patterns in farm product choice.

Subsistence farmers' nutritional outcomes have been the subject of several studies in nutritional science. In particular, Sibhatu et al. (2015) and Jones (2017) show that the production diversity of subsistence farms is positively associated with the farmers' dietary diversity, and consequently with improved macro- and micronutrient intakes. While this literature is interested in the effect of farm production on farmers' nutritional outcomes, I use nutritional outcomes as studied in the nutrition discipline, focusing on energy intake and dietary diversity, as a driving force that may explain subsistence farm production behavior and aggregate agricultural outcomes.

Layout. The rest of the paper proceeds as follows. Section 2 describes the Malawian survey I use, supplementary data sources, and the construction of the main measures. Section 3 explores the extent of farm subsistence in Malawi. Section 4 develops a model of subsistence farming with nutrition demand and trade costs. Section 5 investigates farmer behavior in the model and tests the model's predictions in the data. Section 6 uses the model to evaluate the importance of the nutritional mechanism for aggregate agricultural productivity. Section 7 concludes.

2 Data

2.1 Household Survey

I use the Fourth Integrated Household Survey 2016/17, conducted by the National Statistical Office of the government of Malawi with assistance from the World Bank. It is a nationally representative survey that covers many aspects of household economic behavior. Of the 12,447 Malawian households surveyed, 9,799 (79%) reported producing agricultural goods on their own farm in the past year: it is this sample of farm-operating households that I use for my analysis. I describe the construction of the main measures in this section, and then define certain other variables in Sections 3 and 5 as they become needed for empirical exercises.

Farm Outputs & Inputs. Households report total agricultural output by product in the last rainy season and the last dry season[2] in physical quantities. If any products were sold, they report the quantity sold and the monetary value of this quantity. Many output/sales observations come in standard units easily convertible to kilograms, but far from all do[3]. I am able to convert most non-standard units into kilograms using a companion market survey that was conducted alongside the Third Integrated Household Survey in 2010/11 and provides average weights of measurement units common in Malawian markets[4]. This lets me construct household-product output and sales observations[5].

As will be discussed in Section 3, there is very little selling of agricultural products in this sample. Revenue is therefore not a good measure of output. To aggregate the physical production quantities of multiple goods into one farm-level measure, I weight the physical quantities of various products by the median

[2]Together these amount to roughly one year.

[3]For example, maize is often measured in pails, tobacco in bales, and bananas in bunches.

[4]Ultimately I am able to obtain a weight estimate for 95% of household-product output observations.

[5]For household-product measures, I aggregate multiple varieties of essentially the same good (e.g. *burley tobacco* and *oriental tobacco*) into one product (in this case, *tobacco*).

sale price[6] of each product, giving the market value of *farm output*: this measure roughly represents the revenue the farm could have obtained had it sold everything it produced.

Survey enumerators manually measured the total area of land cultivated by each household using GPS. I use this as a measure of farm area[7].

Non-farm Income. Households report the income from employment (often it is agricultural work on someone else's farm) of every member as well as the profits of any family-run enterprises (excluding the farm). I aggregate these income figures into a household-level measure of *non-farm income*.

Food Consumption. Each family reports the food it consumed in the last seven days. They break it down by good and source (produced, purchased, or received). The problem of non-standard units is even more acute in consumption data than in production data. For many units, the weight can be imputed from the companion market survey mentioned above. Some of the foods are sometimes reported in weights and sometimes in volumes. To convert volumes into weights, I use the FAO/INFOODS Density Database Version 2.0 (Charrondiere et al., 2012). Still, I am unable to produce a sensible weight estimate for certain rare food-unit combinations: they amount to 12% of household-food observations and are excluded from the intake calculations described below.

Food Composition. Based on the constructed weights of household-product and household-food observations, I can estimate the caloric value and the nutrient values (for several key macronutrients and micronutrients[8]) of each observation using the nutritional facts about each food.

I obtain nutritional contents for most of them from the Malawian Food Composition Table (Graan et al., 2019), filling in some gaps with the Tanzanian Food Composition Table (Lukmanji et al., 2008).

[6] I use median product-variety-level price, not product-level price, since for some goods, certain varieties are considerably more valuable than others.

[7] For some, GPS measurements are unavailable and I take the farmers' self-reported land area.

[8] Protein, fiber, vitamins A, C, and D, calcium, iron, potassium, magnesium, added sugar, saturated fat, and sodium.

Caloric Intake and Caloric Requirement. Aggregating the caloric values of foods consumed to the household level, I obtain estimates of total weekly *caloric intake* for each household.

Energy intakes can only be interpreted when compared to the weekly energy needs of each family. I construct the caloric needs of each person using FAO's Human Energy Requirements (FAO, 2004) based on their age and sex, which are reported in the survey. Body weight is not reported by households but is needed for an estimate of caloric requirements for adults. I use the average weight of adult men and of adult women in Malawi as measured by Msyamboza et al. (2013). Summing the energy needs of each person in a family gives the total *caloric requirement* of the household.

Whenever I use household-level farm output, non-farm income, or caloric intake in empirical exercises, I rescale these measures by the household's caloric requirement to obtain "per capita" (where different capita are weighted differently depending on their energy needs) versions of these variables.

Macro- and Micronutrient Intakes and Requirement. In the same way, I construct weekly intakes of several nutrients. For each nutrient, I estimate the daily recommended allowance for each person based on their age and sex using the Dietary Guidelines for Americans, 2020–2025 (USDA and HHS, 2020). Summing the recommended daily allowances of a given nutrient for all individuals in a family gives the total allowance of the household.

GAEZ Attainable Yields. The Global Agro-Ecological Zones (GAEZ) dataset produced by FAO and IIASA (2021) provides crop-location-specific estimates of attainable yield for a wide selection of crops at the level of 5-arc-minute grid cells spanning the globe. Estimates are obtained using an agronomic model fed information about local soil, terrain, and climate conditions. Predictions are conditional on water and input usage[9]. I use the crop-specific distribution of attainable physical yields across locations[10] in Malawi as a source of productivity distributions for the calibration of the model.

[9] I use estimates for rain-fed, low input usage agriculture.
[10] E.g. for maize, GAEZ provides yield estimates for 1374 grid cells in Malawi.

3 Subsistence in the Data

Malawian Farms are Small. The distribution of farm areas in Malawi is roughly lognormal, with the median farm spanning 1.22 acres[11]. Compare this to the area of a standard soccer pitch, 1.76 acres, or to the median US farm size, 45 acres (MacDonald et al., 2013). Malawian households operate farms on a tiny scale, but in spite of their size, these farms are of significant economic relevance to the families that operate them.

Farms are an Important Source of Food for Their Owners. How much do households rely on their own farms as a direct source of food? Some statistics are displayed in Table 1. I find that, on average, 24% of the foods (number of different food products) consumed by a household were produced on its farm[12]. Farm reliance is even more severe in energy: on average, 36% of the calories consumed by a household originated on its own land. The distribution is quite skewed, however: half of all farms produce less than 17% of their caloric intake, while a quarter produce more than 76% of their caloric intake. So while usually most of the consumed food is purchased or received by households rather than produced, the farm is a significant source of variety and energy for many households, with some families relying especially heavily on their farm as a source of energy. Thus Malawian households in this sample engage in semi-subsistence agriculture.

Farms are Primarily Used for Subsistence. The output of most farms is largely destined for own consumption. More than half of all farmers sell none of their agricultural output. The average share of output (by market value) sold is just 16%. Only 11% of farmers sell more than half of their output: all others primarily use their farm as a source of food for their family.

Farms constitute a crucial component of the households' overall economic output. I measure a given household's economic output as the sum of the market value of its farm's production and the non-farm income of the family from em-

[11]See Appendix Figure E.1.

[12]See Appendix Figure E.2 for a ranking of the most popular foods. Some of these, like salt or oil, are impossible or difficult to produce on a small family-operated farm, and so households have to rely on the market.

TABLE 1: Subsistence level summary statistics

	Mean	25 %-ile	Median	75 %-ile
# foods consumed	14.08	9.00	13.00	17.00
food diversity	6.21	3.59	5.31	7.89
share foods produced	0.24	0.12	0.22	0.33
relative kcal intake	1.08	0.68	0.94	1.31
share kcal produced	0.36	0.01	0.17	0.76
farm share in HH output	0.50	0.20	0.47	0.80
share output sold	0.16	0.00	0.00	0.25
production diversity	1.76	1.06	1.60	2.12

NOTE. Share foods produced and share kcal produced are the shares of consumed foods and calories respectively that were produced on the household's own farm. Relative kcal intake is measured as household kcal intake relative to household kcal requirement. Farm share in HH output is the value of farm output relative to the sum of farm output value and non-farm income of the household.

ployment and entrepreneurship. The farm's share is then

$$\frac{\text{farm output value}}{\text{farm output value } + \text{ non-farm income}}$$

The mean of this measure is 0.50, and the median is 0.47, suggesting that for roughly half of the farm-owning households their farm is the main component of the family's economic activity.

Many Farms are Not Specialized. As Table 1 shows, many farms procure a non-negligible fraction of consumed foods from their farm. This implies that these farms are definitely not specializing perfectly. But to what extent is their production diversified exactly?

I measure farm production diversity using the inverse of the Simpson index[13], which, in this setting, is the sum of squared product shares within a farm's output. For my benchmark diversity index, I measure product share as the share of the market value of each product's output in the total market value of the farm's

[13]The Simpson index is the same as the Herfindahl index, with the former name being more common in ecological and agricultural settings.

output[14]. Household h's production diversity index is then

$$\text{Production Diversity}_h = \left(\sum_{i=1}^{n} \left(\frac{\text{output}_{h,i}}{\sum_{j=1}^{n} \text{output}_{h,j}} \right)^2 \right)^{-1}$$

where n is the total number of agricultural products and $\text{output}_{h,i}$ is the market value of product i produced by h's farm. This measure gives the inverse of the probability that two randomly picked dollars from a farm's output value come from the same product. The minimum diversity value is 1, which is attained when the farm produces only one good.

The last row of Table 1 displays the summary statistics of this measure. The 25th percentile farm is almost perfectly specialized, but on average specialization is very incomplete: many farms have significantly diverse production. This result, together with the rarity of selling, could imply that farms do not specialize effectively within their village. For anonimity reasons, the survey does not contain information on the village that each household belongs to, but rather its enumeration area—the smallest geographical reporting unit for census purposes. The average number of households per enumeration area (EA) is 235 in the population (with 12.7 households in the sample I am working with[15]), thus roughly corresponding to the scale of a large village. To compare farm-level diversity to EA-level diversity, I define Normalized Diversity = Production Diversity − 1, so that complete specialization corresponds to 0. I compute this Normalized Diversity for every farm and also for every enumeration area. I find that the Normalized Diversity of a farm relative to the Normalized Diversity of the EA it belongs to is 0.54 on average. One way to interpret this figure is that roughly half of product diversification happens at the level of individual farms, not at the level of villages, suggesting that the extent to which farmers can specialize in one good and trade with their neighbors for other goods is limited.

[14] One simpler measure of diversity would be a count of different products that the farm produces. I prefer the diversity index because it is able to differentiate between two farms that produce the same number of goods but with one farm having a more uneven distribution of output across the goods. Still, results presented in this section are robust to measuring diversity with a product count.

[15] Roughly half of all enumeration areas were included in the survey, with 16 households sampled from each EA. Because I omit households without a farm, the average EA size falls to 12.7.

4 Model

The previous section showed that many households in Malawi engage in semi-subsistence farming: they use their farms primarily as a source of food for the family, and food grown on the farm often accounts for a sizeable portion of the family's diet. The broad goal of the paper is to investigate whether the production decisions that such farmers make are relevant for the aggregate agricultural output of the economy. As a stepping stone, we first need to understand what drives these production decisions. In this section, I develop a model aimed at fulfilling these two objectives.

To be useful for these purposes, the model needs to combine several features. Firstly, the individual agent of this model needs to be a farm-operating household that makes production and consumption decisions jointly. Models in the literature most often separate production and consumption decisions into different agents (e.g. a representative consumer and heterogeneous atomistic farms), which may be appropriate for village- or region-level studies, but isn't sufficient to explain subsistence behavior at the farm level[16]. Secondly, the model needs to feature heterogeneous agricultural products that farmers can choose from, so that it can make predictions on farm product choice. Thirdly, farmers in the model need to face significant trading frictions that actually force them into partial subsistence.

In addition to these basic features, it will be useful for the model to reflect the special role of food, produced by the farms and consumed by their owners, in satisfying the nutritional needs of individuals, with dietary energy needs being the most fundamental.

It is useful to consider energy demand separately because it adds some nuance to how preferences for food are formed. The utility function most often used to represent preferences over multiple goods is the constant elasticity of substitution aggregator. The CES composite over multiple foods captures the love of variety in food as well as the fact that different foods seem to be imperfect substitutes, but it is not well suited to capturing human energy needs. Every person needs to consume a certain number of calories a day in order to power the body's physi-

[16] See Kebede (2020b) for an empirical exploration of separability of production and consumption in the setting of Ethiopian smallholder farmers.

cal activities. Calories from different foods are *perfect* substitutes in their contribution to an individual's energy intake: there is no difference between 100 kcal worth of rice and 100 kcal worth of tomatoes in how much they will contribute to the body's energy allowance. Moreover, the overall utility from energy should be highly nonlinear: consuming far fewer calories than what is recommended for a given person based on their physical characteristics will impose huge costs on their physical condition and mental well-being, while consuming far more calories than needed for satiation is also physically difficult. Increasing daily energy intake from 1000 kcal to 2000 kcal is likely to make almost any person considerably better off, while increasing it from 2000 kcal to 3000 or even 4000 kcal is likely to either have little effect on a person's welfare or actually leave them worse off. Such caloric considerations may be of little importance for explaining aggregate economic behavior of developed countries, but I argue that they can be crucial for understanding the behavior of households and farms (which are usually one and the same) in primarily agricultural developing countries like Malawi. This is why I attempt to capture the energy channel of food demand explicitly in the household problem that follows.

4.1 Household Problem

Consider the problem of a household h. The household consumes multiple foods $\{c_{h,i}\}_{i=1}^{n}$ and a single manufactured good $c_{h,m}$. The manufactured good has taste weight φ_m and is purchased at price p_m. Each food i is characterized by its taste weight φ_i, caloric density k_i, and land productivity $z_{h,i}$. The household can choose to grow any combination of agricultural products with linear technology using its land endowment L_h. The production of each good i is denoted by $x_{h,i}$. The household can purchase $x^p_{h,i}$ or sell $x^s_{h,i}$ of any good i, with price p_i taken as given. Purchases and sales face a proportional trading cost $\tau > 0$. Define $d = 1 + \tau$ for convenience. The household supplies N_h units of labor inelastically to the manufacturing good producer, and is paid wage w for it.

The utility of the household consists of two components. The first is a standard CES aggregator with two layers: the inner layer combines the consumptions of different foods with elasticity of substitution σ, and the outer layer combines the food composite and the manufactured good with elasticity of substitution γ.

The second component is some cost function $f\left(\sum_{i=1}^{n} c_{h,i}k_i,\ K_{req,h}\right)$ whose first argument is the total caloric intake of the household, and the second is the household's exogenous caloric requirement. This term imposes a convex cost on the household's utility for deviating from its caloric needs $K_{req,h}$. I will refer to this cost function as the *caloric deviation penalty*.

The complete problem of the household h is:

$$\max_{\substack{\{c_{h,i}, x_{h,i}, x^p_{h,i},\\ x^s_{h,i}\}_{i=1}^n, c_{h,m}}} \quad \left((1-\varphi_m)\left(\sum_{i=1}^{n} \varphi_i c_{h,i}^{\frac{\sigma-1}{\sigma}}\right)^{\frac{\sigma}{\sigma-1}\frac{\gamma-1}{\gamma}} + \varphi_m c_{h,m}^{\frac{\gamma-1}{\gamma}} \right)^{\frac{\gamma}{\gamma-1}} - f\left(\sum_{i=1}^{n} c_{h,i}k_i,\ K_{req,h}\right) \tag{1}$$

s.t.

$$\sum_{i=1}^{n} \frac{x_{h,i}}{z_{h,i}} = L_h \tag{2}$$

$$\sum_{i=1}^{n} x^p_{h,i} p_i d + p_m c_{h,m} = \sum_{i=1}^{n} x^s_{h,i} \frac{p_i}{d} + wN_h \tag{3}$$

$$c_{h,i} = x_{h,i} + x^p_{h,i} - x^s_{h,i} \quad \forall i \tag{4}$$

$$c_{h,i}, x_{h,i}, x^p_{h,i}, x^s_{h,i} \geq 0 \quad \forall i \tag{5}$$

Caloric Deviation Penalty. The functional form that I choose for the caloric deviation penalty function f is

$$f\left(\sum_i c_{h,i}k_i,\ K_{req,h}\right) = \psi \left(\frac{\sum_i c_{h,i}k_i - K_{req,h}}{K_{req,h}}\right)^2 \frac{K_{req,h}}{\sum_i c_{h,i}k_i}$$

where ψ is a parameter. The function returns 0 when caloric intake equals caloric requirement, but imposes a symmetric convex cost on deviations from the caloric requirement in either direction. Appendix Figure E.3 illustrates how the function varies in caloric intake. This functional form has several appealing properties:

1. $f(aK_{in}, aK_{req}) = f(K_{in}, K_{req})$ (homogeneity of degree 0)
2. $f(aK_{req}, K_{req}) = f\left(\frac{K_{req}}{a}, K_{req}\right)$ (symmetry around K_{req} in ratios)

3. $\min_{K_{in}>0} f(K_{in}, K_{req}) = f(K_{req}, K_{req}) = 0$ (min and zero if intake = K_{req})

4. $f_{11}(K_{in}, K_{req}) = \frac{2\psi K_{req}}{K_{in}^3} > 0$ (convex in intake)

The caloric deviation penalty is the principal novelty of this model. Making the agent care about calories in this way, rather than using more standard ways of defining preferences over agricultural goods, does offer some advantages in the context of modeling the consumption and production behavior of subsistence farmers. For example, a subsistence constraint in a composite agricultural good, common in the structural transformation literature, would be silent on cross-household heterogeneity in the relative consumptions and productions *across* agricultural goods within the agricultural composite. Defining good-specific subsistence constraints, or introducing good-specific income elasticities in a different way, would offer more flexibility, but would be silent on *why* some goods should have stricter subsistence constraints or lower income elasticities than others. In contrast, the model I define above is able to determine *endogenously* which goods are preferred based on a household's circumstances. The caloric deviation penalty will also be key in reproducing the empirical patterns in farm consumption and production behavior discussed in Section 5.

4.2 General Equilibrium & Estimation

Manufacturer. A representative competitive manufacturer produces the manufactured good Y_m with linear technology in one input: labor N, which is aggregated from households' labor supplies: $N = \sum_h N_h$. The firm sells the manufacturing good back to households for p_m and pays them w for the labor.

Its profits are

$$p_m Y_m - wN$$

with $Y_m = z_m N$.

Imposing a zero profit condition implies

$$p_m = \frac{w}{z_m}$$

Both p_m and z_m are normalized to 1, which also implies $w = 1$.

Agricultural Goods. Each agricultural good i in the model is defined by four characteristics: taste weight φ_i, caloric density k_i, market price p_i, and the distribution of land yields $z_{h,i}$ across households. I select six agricultural goods common in the data to be represented in the model[17]: maize, pigeonpea, groundnut, tomato, soybean, and tobacco. These six goods account for, on average, 70% of the market value of household output and 43% of the market value of household food consumption.

Caloric densities k_i for all edible agricultural goods are obtained from the food composition tables. The estimation of taste weights φ_i of all edible goods and the $z_{h,i}$ distributions of all goods are described later in the section. I set $\varphi_{\text{tobacco}} = k_{\text{tobacco}} = 0$, since its consumption is not observed.

Market Clearing. For each agricultural good i except tobacco, the quantity delivered by households to the market needs to match the quantity purchased by households from the market, accounting for losses induced by the trade cost:

$$\frac{1}{d}\sum_h x^s_{h,i} = d\sum_h x^p_{h,i} \quad \forall i$$

In the data, tobacco is special not only because it is not edible, but also because it is the predominant export of Malawi: in 2017, raw tobacco accounted for 60% of Malawian export value[18]. To represent tobacco's unique role as a major export of Malawi, I set its price exogenously and treat $\bar{p}_{\text{tobacco}}$ as a parameter, meaning that tobacco can be traded internationally. The price is calibrated to match tobacco's share in aggregate farm output: $\frac{\sum_h x_{h,\text{tobacco}}\bar{p}_{\text{tobacco}}}{\sum_i \sum_h x_{h,i}p_i}$.

Tobacco is internationally traded at a fixed price $\bar{p}_t$, so its domestic market need not clear. Imposing that all other agricultural good markets clear, an application of Walras's Law implies

$$\underbrace{\bar{p}_{\text{tobacco}}\left(\frac{1}{d}\sum_h x^s_{h,\text{tobacco}} - d\sum_h x^p_{h,\text{tobacco}}\right)}_{\text{tobacco exports}} = \underbrace{p_m\left(\sum_h c_{h,m} - Y_m\right)}_{\text{manuf. good imports}}$$

[17]Using significantly fewer goods would leave little room for variation in consumed and produced diversity across households. Using more goods would make solving for prices that clear all markets simultaneously prohibitively costly.

[18]UN Comtrade Database.

Meaning that for trade to be balanced, the manufactured good also has to be tradeable internationally. I assume that the manufactured good can be imported from abroad at the same normalized price of 1.

Food Taste Weights and Elasticities of Substitution. The first order condition for household h, good i's consumption is

$$c_{h,i} = c_{h,m}\varphi_i^{\sigma}\left(\eta_{h,i} + f_1\left(\sum_j c_{h,i}k_i,\ K_{req,h}\right)k_i\right)^{-\sigma}\left(\frac{\varphi_m}{1-\varphi_m}\right)^{-\gamma}(\mu p_m)^{\gamma}$$
$$\cdot\left(\sum_{j=1}^{n}\varphi_j^{\sigma}\left(\eta_{h,j} + f_1\left(\sum_j c_{h,j}k_j,\ K_{req,h}\right)k_j\right)^{1-\sigma}\right)^{\frac{\gamma-\sigma}{\sigma-1}}$$

where λ_h and μ_h are Lagrange multipliers, and $\eta_{h,i} = \lambda_h / z_{h,i}$ if i is produced by h or $\eta_{h,i} = \mu_h p_i d_h$ if i is purchased by h.

Taking logs and doing a log-linear approximation on one of the terms, this FOC can be expressed as

$$\log c_{h,i} \approx \underbrace{\gamma(\log p_m - \log\frac{\varphi_m}{1-\varphi_m})}_{\text{constant}}$$
$$\underbrace{+\log c_{h,m} - \sigma\log\lambda_h + \gamma\log\mu_h + \frac{\gamma-\sigma}{\sigma-1}\log\left(\sum_{j=1}^{n}\varphi_j^{\sigma}\left(\eta_{h,j} + f_1\left(\sum_j c_{h,j}k_j,\ K_{req,h}\right)k_j\right)^{1-\sigma}\right)}_{\text{HH-produced FE}}$$
$$+\underbrace{\sigma\log\varphi_i}_{\text{good FE}} + \sigma\underbrace{\log z_{h,i}}_{X_{1,h,i}} - \underbrace{k_i z_{h,i}}_{X_{2,h,i}} \cdot \underbrace{\sigma\frac{f_1\left(\sum_j c_{h,j}k_j, K_{req,h}\right)}{\lambda_h}}_{\text{HH-produced FE2}} \qquad (6)$$

for HH-good combinations where the good is produced, and

$$\log c_{h,i} \approx \underbrace{\gamma(\log p_m - \log\frac{\varphi_m}{1-\varphi_m})}_{\text{constant}}$$

$$\underbrace{+\log c_{h,m} + (\gamma-\sigma)\log\mu_h + \frac{\gamma-\sigma}{\sigma-1}\log\left(\sum_{j=1}^{n}\varphi_j^{\sigma}\left(\eta_{h,j}+f_1\left(\sum c_{h,j}k_j,\ K_{req,h}\right)k_j\right)^{1-\sigma}\right)}_{\text{HH-purchased FE}}$$

$$+\underbrace{\sigma\log\varphi_i}_{\text{good FE}} + \sigma\underbrace{\log p_i d_h}_{X_{1,h,i}} - \underbrace{\frac{k_i}{p_i d_h}}_{X_{2,h,i}}\cdot\underbrace{\sigma\frac{f_1\left(\sum_j c_{h,j}k_j, K_{req,h}\right)}{\lambda_h}}_{\text{HH-purchased FE2}} \quad (7)$$

for HH-good combinations where the good is purchased.

Thus the model produces two expressions that can be estimated as one regression and provide estimates of the elasticity of substitution between foods σ (the coefficient on $X_{1,h,i}$) and $\{\varphi_i\}_i$ (which can be extracted from the good fixed effect conditional on the choice of σ). The only variables that need to be observed are $c_{h,i}$ (which I map to the physical quantity of i consumed by h), $z_{h,i}$ (map to the reported physical land yield), $p_i d_h$ (map to the reported purchase price), and k_i (map to the caloric density).

Estimating this regression yields $\sigma = 0.75$ (standard error 0.01), implying that foods are complements for the CES term in the household's utility. However, as will be shown in Section 4.3, the caloric deviation penalty effectively makes the foods more substitutable than the CES term dictates, especially for the poorest households that are forced to deviate far from their caloric requirement. For comparison, Kebede (2020a) estimates a σ of 1.3 in a similar setting, and Behrman and Deolalikar (1989) estimate it to be 1.25 for a sample of very poor countries and 0.28 for a sample of developing ones, supporting the idea that apparent substitutability between foods may be higher for poorer samples.

The elasticity of substitution between the food composite and the manufactured good, γ, cannot be estimated in this way because the manufactured good in the model is a fictitious good that captures all non-food consumption: its quantities and prices cannot be measured in the data. For lack of a good moment to calibrate this parameter to, I set $\gamma \approx 1$: the outer CES layer thus converges to the Cobb-Douglas utility function.

Household Heterogeneity. Households are heterogeneous in three dimensions: land endowment L_h, non-farm income endowment wN_h, and a set of productivity draws across agricultural products $\{z_{h,i}\}_i$. Caloric requirement $K_{req,h}$

is set to 1 for all farms, meaning that every household's endowments and outcome variables can be interpreted as being relative to the household's caloric requirement—matching how I measure these in the data.

Productivity Distributions. I assume a log-normal distribution of physical yield across households for each crop $(z_{h,i})$. I compute the crop-specific mean and variance of log yields across Malawian grid cells in GAEZ data, using predicted attainable yields for water-fed, low input usage agriculture.

Size and Income Distributions. The land endowment L_h is log-normally distributed. The parameters of the land distribution are not taken from the data, even though land area is observable. Instead, the mean of log land area distribution is calibrated to match the average $\frac{K_{in,h}}{K_{req,h}}$ ratio as a way to ensure a realistic scale of the solution[19]. The variance of log land area distribution is calibrated to match the variance of log output value (whose distribution is approximately normal in the data). The calibrated L_h distribution should therefore be thought of as "effective" acres, capturing the differences in household-level farming productivity common to all agricultural products.

Because w is the same for all households, the heterogeneity in non-farm income wN_h is entirely due to the distribution of N_h, household non-farm labor. The empirical distribution of non-farm income relative to caloric requirement is approximated well with a mass of households at zero income and a log-normal distribution of positive incomes. The mass at zero non-farm labor, $P(N_h = 0)$, is taken directly from the data. Likewise, the variance of log positive labor $V(\log N_h \mid N_h > 0)$ maps directly to the observed $V(\log wN_h \mid wN_h > 0)$. Because aggregate non-farm income and manufactured good consumption are directly linked in the model[20], mean log positive labor $\mathbb{E}(\log N_h \mid N_h > 0)$ can be normalized to 1, with the job of matching the observed scale of non-farm income falling to the φ_m parameter, as described below.

[19]By design, household behavior in the model is scale-dependent, with scale being defined relative to the caloric requirement $K_{req,h}$ of each household.

[20]In the absence of international trade in the manufactured good, $\sum_h c_{h,m} = \sum_h z_m N_h$. While the manufactured good can actually be imported, the value of imports is effectively limited by the need to match the observed share of tobacco in farm sales, and is comparatively minor in the estimated model.

Other Parameters. The taste weight of the manufactured good, φ_m, is calibrated to match the ratio of aggregate non-farm income to aggregate farm output value: $\frac{\sum_h wN_h}{\sum_h \sum_i x_{h,i} p_{h,i}}$ in model terms. For instance, a higher manufactured taste weight would raise the demand for c_m, necessitating a drop in food prices $\{p_i\}_i$ (relative to the normalized $p_m = 1$), which lowers the market value of agricultural output and raises the aforementioned ratio.

The trade cost d is what forces farms into subsistence: the higher the d, the more attractive it becomes to produce goods for own consumption, rather than sell farm output and buy food on the market. This parameter is calibrated to match the average share of farm output sold.

The strength of the caloric deviation penalty, ψ, is calibrated to match the farm output elasticity of energy intake. The relevance of the caloric deviation penalty for the relationship between farm size and household energy intake will be discussed in Section 5.

Simulation. 500 household types take independent draws from the N_h and the $\{z_{h,i}\}_i$ distributions. Within each of these types, I approximate the L_h distribution using 80 sub-types. This procedure yields an economy populated with 40,000 households. Due to the structure of the household's problem, each household needs to be solved separately. See Appendix C for details.

Table 2 summarizes the calibration of the model. Note that the parameter-moment mapping is only an approximation, as all model moments are determined by all parameters jointly.

4.3 Nutritional Mechanism

Nutrition demand in the model is driven by the introduction of the caloric deviation penalty and in turn drives much of the farm behavior discussed in Section 5. In what follows, I present several analytical results that help elucidate the mechanism.

First of all, note that any solution has to have $c_{h,i} > 0$ for all i with $\varphi_i > 0$. This means that the household has to either produce or purchase every good except tobacco (the only good with $\varphi_i = 0$).

For a simple comparison of two goods that are equally costly and equally liked

TABLE 2: Model Estimation

parameter	value	moment/source	**data** moment	**model** moment
Distributions				
$\mathbb{E}\,(\log L_h)$	-15	avg $K_{in,h}/K_{req,h}$	1.036	0.902
$V\,(\log L_h)$	1.5	$V\left(\log \text{output}_h\right)$	1.528	1.385
$P\,(N_h = 0)$	0.112	$P\,(\text{non-farm income}_h = 0)$	0.112	0.117
$\mathbb{E}\,(\log N_h \mid N_h > 0)$	1	normalization	—	—
$V\,(\log N_h \mid N_h > 0)$	2.103	$V\,(\log \text{non-farm income}_h)$	2.103	1.924
Parameters				
σ	0.75	estim. of Eqns 6, 7	—	—
γ	1	—	—	—
d	1.75	avg share sold	0.159	0.203
ψ	0.5	output elasticity of K_{in}	0.091	0.124
Good characterstics				
$\{\varphi_i\}_i$	...	estim. of Eqns 6, 7	—	—
φ_m	0.5	$\frac{\text{aggr. non-farm income}}{\text{aggr. farm output}}$	1.539	1.632
$\{k_i\}_i$	...	Food Comp. Tables	—	—
$\{z_i\}_i$	...	GAEZ attainable yields	—	—
z_m	1	normalization	—	—
$\bar{p}_{\text{tobacco}}/p_{\text{maize}}$	5.4	aggr. tobacco output share	0.091	0.094

by the household, the model makes straightforward predictions on how their relative consumption will be affected by calories:

Proposition 1 (Consume kcal-dense foods when undereating[a])
Suppose there are two goods i and j with equal taste weights $\varphi_i = \varphi_j$ and marginal costs $MC_{h,i} = MC_{h,j}$[b] for some household h, but one has a higher energy density: $k_i > k_j$.

1. if the household is undereating calories, consumption of the calorically denser good is preferred:

$$\sum_i c_{h,i}k_i < K_{req,h} \quad \Longrightarrow \quad c_{h,i} > c_{h,j}$$

2. if the household is overeating calories, consumption of the calorically emptier good is preferred:

$$\sum_i c_{h,i}k_i > K_{req,h} \quad \Longrightarrow \quad c_{h,i} < c_{h,j}$$

[a] See Appendix B for the proof of this proposition and of all that follow.

[b] If both goods i and j are optimally produced by the household ($x_{h,i}, x_{h,j} > 0$), then marginal costs are equal when $z_{h,i} = z_{h,j}$. If both goods are purchased by the household ($x^p_{h,i}, x^p_{h,j} > 0$), marginal costs are equal when $p_i = p_j$. If one (say, i) is produced and the other (j) is purchased, marginal costs are equal if $z_{h,i}p_j = \lambda/(\mu d)$, where λ and μ are the Lagrange multipliers associated with constraints (2) and (3) of the household problem respectively.

A related proposition helps illuminate the role that adding the caloric deviation penalty f to the standard CES utility has on the consumption bundle chosen by the household.

Proposition 2 (Calories skew the CES solution)
Suppose there are two goods i and j that are both produced by household h ($x_{h,i} > 0$, $x_{h,j} > 0$), and one has a higher calorie yield than the other: $k_i z_{h,i} > k_j z_{h,j}$[a]. Let “CES” denote the allocation in an auxiliary model with $\psi = 0$ (i.e. with no caloric deviation penalty).

1. if the household is undereating calories ($\sum_i c_{h,i}k_i < K_{req,h}$), relative consumption is skewed from the pure CES solution toward the calorically productive good:
$$\frac{c_{h,i}}{c_{h,j}} > \left(\frac{\varphi_i z_{h,i}}{\varphi_j z_{h,j}}\right)^{\sigma} = \frac{c_{h,i}^{CES}}{c_{h,j}^{CES}}$$

2. if the household is overeating calories ($\sum_i c_{h,i}k_i > K_{req,h}$), relative consumption is skewed from the pure CES solution toward the calorically unproductive good:
$$\frac{c_{h,i}}{c_{h,j}} < \left(\frac{\varphi_i z_{h,i}}{\varphi_j z_{h,j}}\right)^{\sigma} = \frac{c_{h,i}^{CES}}{c_{h,j}^{CES}}$$

[a] If both goods are instead purchased ($x^p_{h,i} > 0$, $x^p_{h,j} > 0$), the required condition is $\frac{k_i}{p_i} > \frac{k_j}{p_j}$.

The previous two propositions suggest that households that cannot afford to satisfy their caloric requirement will skew their consumption toward the cheaper sources of calories. The following proposition takes this logic to the extreme.

Proposition 3 (Poorest households maximize calories, specialize)
As land endowment L_h and non-farm income wN_h of household h approach 0, the solution of the full problem defined in (1)-(5) converges to the solution of the following problem:

$$\max_{\{c_{h,i},x_{h,i},x^p_{h,i},x^s_{h,i}\}_{i=1}^n,c_{h,m}} \quad \sum_{i=1}^n c_{h,i}k_i$$

s.t. the same constraints (2)-(5).

In the solution of this limiting problem, the number of goods consumed is either 1 or 2, and the number of goods produced is 1.

Which good(s) are produced and which are consumed in this limit depends on the properties of the goods and the trade cost d. In particular, there is a household-specific cutoff trade cost, call it $\bar{d}_h$:

$$\bar{d}_h = \sqrt{\frac{\max_i p_i z_{h,i}}{\min_i p_i/k_i \cdot \max_i k_i z_{h,i}}}$$

If $d < \bar{d}_h$, then the household produces only the $\arg\max_i p_i z_{h,i}$ good and consumes only the $\arg\min_i p_i/k_i$ good.

If $d > \bar{d}_h$, then the household produces only the $\arg\max_i k_i z_{h,i}$ good and consumes only the same $\arg\max_i k_i z_{h,i}$ good as well as the $\arg\min_i p_i/k_i$ good.

If $d = \bar{d}_h$, the household is indifferent between the two solutions.

The first statement of Proposition 3 arises because in the extreme poverty limit (L_h and wN_h approaching 0 while the energy requirement K_{req} stays fixed), the curvature on the caloric deviation penalty function f is such that the problem of the household effectively converges to maximizing caloric intake alone. Because calories from different foods are perfect substitutes and production is linear, the solution is simple. All non-farm income is devoted to purchasing the cheapest source of calories on the market (the $\arg\min_i p_i/k_i$ good). All land is devoted either to producing the most efficient source of calories (the $\arg\max_i k_i z_i$) and eating it at home, or to producing the revenue-maximizing good ($\arg\max_i p_i z_{h,i}$), selling it, and purchasing the cheapest source of calories on the market (the $\arg\min_i p_i/k_i$ good) with the proceeds. The choice between these two alternatives is driven by how costly trade is: at high levels of d, producing and selling the $\arg\min_i p_i/k_i$ good in order to purchase the $\arg\min_i p_i/k_i$ good becomes relatively less attractive, while the subsistence path of producing and consuming the $\arg\max_i k_i z_{h,i}$ good becomes more so.

The increasing concentration of consumption and production of poor households in the most efficient source(s) of calories (at the expense of diversity, tastes, and non-food consumption) is at the core of this model's nutrition demand mechanism. The next section explores the household behavior generated by this mechanism and compares it to observed patterns in the behavior of Malawian farmers.

5 Farm Behavior in the Model and in the Data

5.1 Larger Farms Shift Consumption from Energy to Diversity

In the model, households with vanishingly small farms and non-farm incomes dedicate all their resources to maximizing their caloric intake (Proposition 3), leading to extremely specialized consumption and production. However, this case is unrealistically extreme, however. How does food consumption differ across farms of different sizes in the estimated model, with realistic distributions of farm size and income?

I focus on two dimensions of food consumption that are straightforward to measure both in the model and in the data. The first is household caloric intake (relative to caloric requirement). The second is food diversity, measured using the inverse of the Simpson index, analogously to the production diversity measure defined previously. In this case, the index uses the market value of the consumed quantity of each food[21]:

$$\text{Food Diversity}_h = \left(\sum_{i=1}^{n} \left(\frac{c_{h,i} p_i}{\sum_{j=1}^{n} c_{h,j} p_j} \right)^2 \right)^{-1}$$

The summary statistics of this measure in the data are included in Table 1.

Model. First, consider how food consumption of a household depends on the size of the household's farm and income in the baseline calibration of the model. Column (2) of Table 3 shows the results of a regression of log energy intake on log farm output and log non-farm income in the model, while column (3) uses the food diversity index as the outcome variable. Energy intake, farm output, and non-farm income are relative to the caloric requirement $K_{req,h}$ of each household. Households with tiny farms struggle to satisfy their caloric requirement, which makes the marginal utility of each additional calorie comparatively large. This leads them to concentrate their modest resources in the most efficient source of calories, keeping dietary diversity low. Households with larger farms find it easier to satisfy their caloric needs: marginal utility of calories falls and they shift some resources from obtaining calories to diversifying the diet to satisfy their love of variety. This creates a positive relationship between farm size on one hand and both energy intake and food diversity on the other. However, because the smaller farms were directing a far larger share of their resources to obtaining calories, their caloric intake wasn't that much lower, so the relationship between output

[21] In the data, I map $c_{h,i}$ to the physical quantity of good i in kg consumed by household h, and p_i to the median purchase price of good i.

TABLE 3: Household food consumption vs farm size: model and data

	log kcal intake			food diversity		
	(1) **model**: $\psi = 0$	(2) **model**: benchmark	(3) **data**	(4) **model**: $\psi = 0$	(5) **model**: benchmark	(6) **data**
log output	0.663	0.124	0.091***	−0.082	0.428	0.395***
	(0.002)	(0.001)	(0.005)	(0.002)	(0.002)	(0.034)
log non-farm income	0.354	0.084	0.063***	0.041	0.396	0.857***
	(0.004)	(0.001)	(0.004)	(0.002)	(0.002)	(0.033)
N	29,168	33,613	8,674	29,168	33,613	8,675
Adj. R^2	0.919	0.393	0.063	0.085	0.762	0.131

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

NOTE. Kcal intake, output, and non-farm income are relative to the caloric requirement of the household. Food diversity index is calculated using product shares in a household's total food value.

and energy consumption is quite flat in the model: the output elasticity of kcal intake is just 0.124.

Data. In the data, the output elasticity of kcal intake, displayed in column (3) of Table 3, is similarly low: a 1% increase in output or income is associated with just a 0.09% or 0.06% rise, respectively, in energy intake. As an illustration, the average energy intake (again, relative to requirement) is 0.91 in the 1st quartile of farms by total shadow income (output + non-farm income) and 1.21 in the 4th quartile, while the corresponding difference in average shadow incomes is more than tenfold. As households become wealthier, they barely increase their consumption of calories. The output elasticity of caloric intake (0.091) was used as a moment in the estimation of the model, primarily to discipline the strength of the caloric deviation penalty, ψ. Other data coefficients in the table are non-targeted. Column (6) shows that, just like the model predicts, households running bigger farms or earning more income in the data do eat more diverse diets[22].

Another way to measure dietary diversity is to simply count the number of distinct foods[23] consumed by the household. This measure is highly correlated

[22] Behrman and Deolalikar (1989) find a similar relationship at the cross-country level: aggregate food price and quantity data suggests an increasing relationship between the consumers' taste for food variety and the country's development, at least among low-income countries.

[23] The left panel of Figure E.2 shows the most widely consumed foods.

with food diversity (see Appendix Table A.1) and has a similar relationship with farm size and non-farm income (see Appendix Table D.3): an increase of one log unit in farm output or non-farm income is associated with respectively roughly 0.7 and 1.4 extra food products consumed by the household. The same table also shows that the strong positive relationship between size/income and food diversity is almost unchanged when total energy intake is controlled for.

Finally, the average food diversity index is 4.9 in the 1st quartile of farms by total shadow income and 8.3 in the 4th quartile. The average count of foods goes from 11.8 to 17.6 over the same span.

One reason to consume a diverse diet is to satisfy the pure love of variety, which is the channel captured in the model. Another reason is to ensure a sufficient consumption of essential macro- and micronutrients. I find that the nutrient richness of a household's diet behaves similarly to the diversity of the diet: see Appendix A.

Importance of the Caloric Mechanism. The utility cost f of deviating far from the caloric requirement is key in the ability of the model to reproduce observed food consumption behavior. To illustrate this, I show the result of running the same regressions in a model without the caloric deviation penalty (equivalent to setting $\psi = 0$) in columns (1) and (4). In this "pure CES" model, intake would grow one-for-one with shadow farm income, resulting in large output and income elasticities of kcal intake: 0.660 and 0.362 respectively in this calibration. Moreover, because removing the caloric deviation penalty f makes preferences homothetic, changes in farm size and income do not alter the relative consumptions, meaning that consumption diversity becomes invariant in farm size. This leads to coefficients close to 0 when regressing food diversity on log farm output or non-farm income, in contrast to the empirical coefficient of 0.395 and 0.857.

Explicitly modeling dietary energy needs allows this model to reproduce the salient empirical facts on how the composition of a household's diet depends on the size of the farm that the household is operating. Below, I explore the predictions of the model on selling and production behavior of subsistence farmers, and compare them to the empirical patterns in the behavior of Malawian farmers.

5.2 Farm Sales are Specialized

Model. Households in the model don't just consume agricultural goods: they can also grow them and sell them on the market. The problem of which goods to grow and how much to sell is numerical and will be explored below, but the problem of choosing *what* to sell has a simple analytical solution in the model:

> **Proposition 4 (Sell the revenue-maximizing product)**
> If household h sells good j ($x^s_{h,j} > 0$), then it must be the most revenue-productive good: $j = \arg\max_i p_i z_{h,i}$.

The proposition implies that, as long as the maximizer is unique, the household would never sell more than one good, no matter how many it is producing[24]. Sales are perfectly specialized in the comparative advantage product.

Data. Farm behavior in the data is similar, albeit less extreme: farms are significantly more specialized in their sales than in their overall production. Among farms that sell something fully 69% of all sellers sell just one good, while only 9% produce just one good. On average, the good the household sells the most accounts for 91% of sales, but the good the household produces the most (usually the same) accounts for 67% of output[25].

5.3 Larger Farms Sell More

Model. As long as trade cost d is non-negligible, farms in the model are unlikely to fully specialize their production in the revenue-maximizing good and trade for the rest: at high levels of d, it's relatively cheaper to produce certain goods at home since subsistence production is not subject to the trade cost. But the choice of whether to sell any of the output and if so, how much, is strongly linked to the size of the farm, as columns (1) and (3) of Table 4 show. Among

[24] If multiple goods share the argmax, then the household is indifferent between any reallocation of production and sales across these goods (including only selling one of these goods), as long as the total revenue is preserved. This coincidence does not occur in the simulated model.

[25] The same point can be made with diversity indices: the average production index value among sellers is 2.04, which falls to 1.23 (close to 1, which denotes perfect specialization) when only their sales are considered.

TABLE 4: Farm size and selling: model vs data

output quartile	sold output share		fraction sellers	
	(1) **model**	(2) **data**	(3) **model**	(4) **data**
1	<1%	13%	<1%	14%
4	67%	31%	>99%	77%

NOTE. Output quartile based on market value of farm output. Sold output share is averaged within each quartile. Fraction sellers is the fraction of farms in each quartile with non-zero sales.

the simulated model farms, the average share of farm output that is sold is almost zero in the smallest quartile of farms but 67% in the largest one. Much of this effect is due to the extensive margin: almost no farms in the smallest quartile choose to sell anything, while almost all in the largest quartile choose to do so.

There are two related reasons why larger farms sell more in this model, despite it not having any fixed costs of selling. Firstly, larger farmers shift their consumption from calories to diversity. At least some of the foods are likely to be cheaper to buy on the market in exchange for the revenue-maximizing good, rather than to grow at home. To do that, farmers need revenue. Secondly, larger farmers shift their consumption away from food in general to the manufactured good, which has to be purchased on the market and so also requires getting cash. Both of these channels are generated by the caloric deviation penalty. Smallest farmers sell very little because the need to obtain calories eclipses all else for them, and at the calibrated level of trade cost d exceeds the $\bar{d}_h$ described in Proposition 3 for most simulated households, meaning that the cheapest way to get calories for most is to grow the good with the highest caloric productivity on their own farm.

Data. Only 47% of farms in the sample sell any agricultural output. The likelihood of being a seller increases significantly in size, however: columns (2) and (4) of Table 4 show that larger farms are more active sellers in the data as well, matching the prediction of the model. The empirical relationship is strong, although not as extreme as in the model.

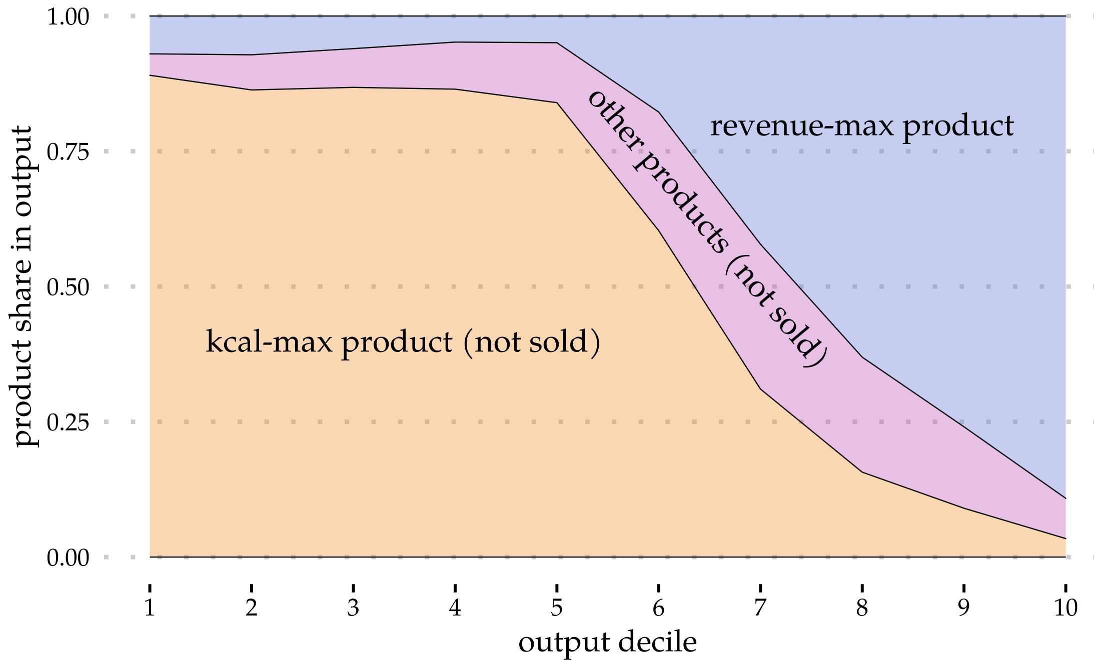
While the model calibrated to match only the average share of output sold reproduces also the upward relationship between farm size and selling, it overpredicts its strength. One missing element that could improve the model's performance is risk. Volatile harvests or market prices would make specialization in cash crops carry an increased risk of starvation. All but the largest farmers in the model would thus be reluctant to commercialize too heavily once their nutritional needs are satisfied: some precautionary subsistence production can help insure against negative shocks to the comparative advantage good's yield or price, which might leave normally affluent farmers with insufficient revenue to satisfy the family's nutritional needs.

Still, the overprediction of the size-selling relationship by the model is likely to bias down the aggregate results coming in Section 6. Because the model overstates how many of the large farmers are sellers and how much they sell, their predicted product choice response to counterfactual policy changes will be subdued—there is little extra commercialization that large farmers can undertake. And because of their size, the weaker response by large farmers is particularly important for attenuating the aggregate response in the model.

5.4 Larger Farms Diversify Production, Shift to Marketable Goods

Model. What ultimately matters for the aggregate productivity of the agricultural sector in this model is the product choice of individual farmers. The caloric deviation penalty makes farm product choice depend on the scale of the farm relative to its caloric requirement. Figure 1 summarizes how the goods that farmers decide to grow on their land depend on the size of the farm by splitting the total output of each farm size decile into three categories of goods. As Proposition 3 suggested should be the case for small farms facing sufficiently costly trade (most simulated farms indeed face $d > \bar{d}_h$), production of small farms is dominated by the $\arg\max_i k_i z_{h,i}$ good. The identity of this good depends on the productivity draws of each farm, but most do choose to specialize in growing the most calorie-productive good available.

Larger farms (those that can produce more output relative to their caloric requirement) can afford to diversify their diet. With a high trade cost, many of the agricultural goods available are cheaper to grow at home than to buy on the market, leading farmers to diversify their subsistence production: the share of other products grown but not sold grows relative to the share of the most calorie-



NOTE. Products are split into groups at farm level. "kcal-max product (not sold)" is the $\arg\max_i k_i z_{h,i}$ good, unless it's the same as the $\arg\max_i p_i z_{h,i}$ and is sold. "revenue-max product" is the $\arg\max_i p_i z_{h,i}$ good, unless it's the same as the $\arg\max_i k_i z_{h,i}$ and is not sold. "other products (not sold)" are all other goods. Product shares are the output value shares of each product group in the decile-level output value. Farms are grouped into deciles by output market value relative to caloric requirement.

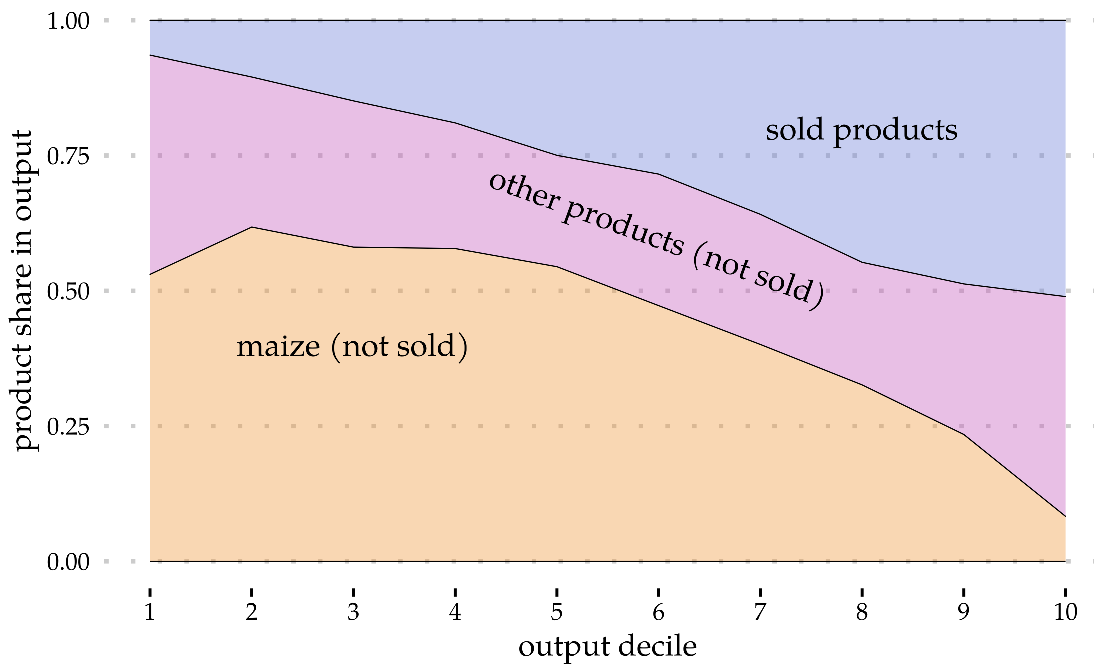
FIGURE 1: Farm size and product choice: model

productive good as farms get larger.

Finally, as Section 5.3 discussed, larger farmers can afford to shift resources from obtaining food to obtaining the manufactured good, for which they need revenue. The share of the most revenue-productive good in output thus grows in farm size, dominating farm output at the top of the farm size distribution[26].

Data. Figure 2 is the conceptually similar data analogue of Figure 1. It shows the evolution of the output share of maize (as long as it's not sold by the farm), products sold by the farm, and all other products with farm size. The smallest farms in the sample mostly produce maize, the dominant staple crop in the region and a rough analogue of the kcal-max product in the model. As the model pre-

[26] Appendix Figure E.4 shows similar patterns, but using average product shares across farms within a decile, rather than product shares within decile-level output.



Note. Products are split into groups at farm level. Product shares are the output value shares of each product group in the decile-level output value. Farms are grouped into deciles by output market value relative to caloric requirement.

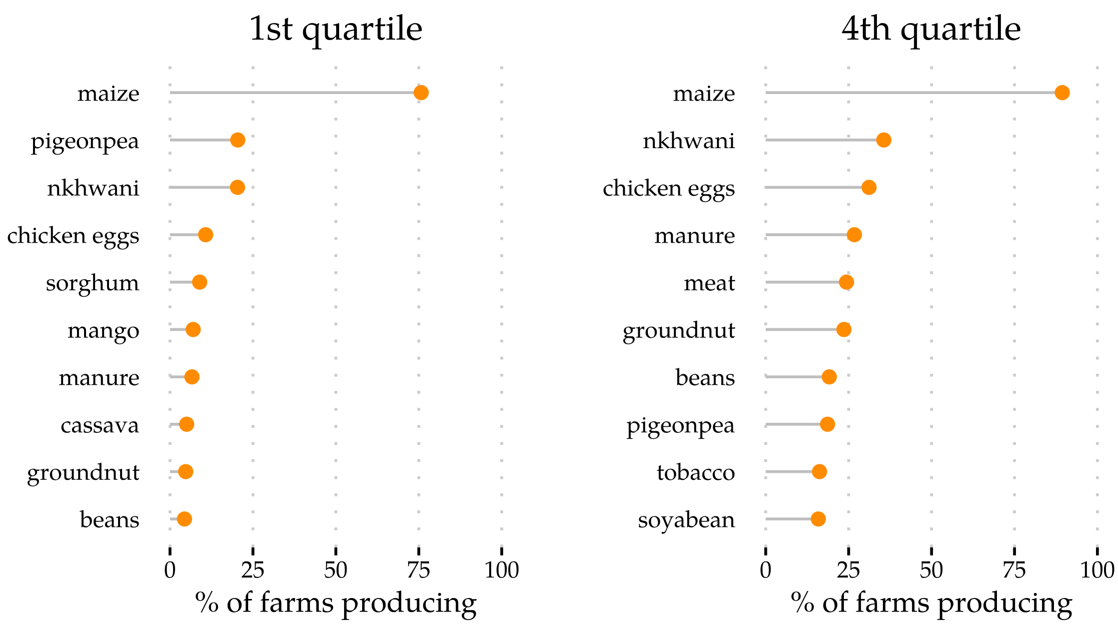
Figure 2: Farm size and product choice: data

dicts, larger farms diversify their subsistence production: output that is not sold is more evenly split between different products. Finally, larger farms increasingly focus on producing goods that are at least partially sold[27].

As with the selling behavior discussed in Section 5.3, the model reproduces the empirical relationship between farm size and product choice, but overpredicts its strength. Again, one missing feature of the model that could improve its performance is yield or price volatility, which would raise the incentives to diversify production for farmers across the entire size distribution.

Figure 3 shows how the probability a farm produces a specific product depends on its size. Virtually all farms grow at least some maize. But what makes small farms different is that most of them grow little else. Pigeonpea and nkhwani (two other common staple crops in Malawi) are grown by about a quarter of small farms each, but all the other products are exceedingly rare. Among large farms,

[27] Appendix Figure E.5 shows a similar pattern, but using average product shares across farms within a decile, rather than product shares within decile-level output.



Note. Farms are grouped by output quartile (output market value relative to the caloric requirement). For each group and each product, the percentage of farms in that group that produce that product is computed. Products are ranked by this measure within each group, and only the top 10 are displayed.

Figure 3: Products ranked by the % of farms within each output quartile producing each product

the frequency of all products is much higher. But not only do larger farms have more diverse production: their product selection also changes. Animal products like eggs and meat are far more common among large farms not only in absolute, but also in relative terms. Tobacco makes an appearance in the ranking for the top quartile, with about a fifth of the large farms producing it, while virtually none of the small farms do.

The data on households' consumption of various goods adds nuance to this picture. Figure E.2 in the Appendix shows the percentage of farms that have reported consuming a given food in the past week, and the percentage that have reported buying it. It suggests that maize is almost universally consumed, but is less frequently bought. Households seem to rely on their land to produce staples like maize and nkhwani, but rely on their income to purchase goods like salt, oil, or sugar, which are commonly used but are difficult to produce yourself.

5.5 Farms Facing Lower Trade Costs Specialize

Model.

Proposition 5 (Specialize production if trade costs are low)
Let j be household $h's$ revenue-maximizing good: $j = \arg\max_i p_i z_{h,i}$.
If $d < \sqrt{\frac{p_j z_{h,j}}{p_i z_{h,i}}}$ for some other good i, then selling j and buying i with the revenue is cheaper than growing i on the farm, implying $x^p_{h,i} > 0,\ x_{h,i} = 0$ (as long as $\varphi_i > 0)$[a].
If furthermore $d < \tilde{d}_h \equiv \sqrt{\frac{p_j z_{h,j}}{\max_{i \neq j} p_i z_{h,i}}}$, then h only produces j, sells it, and purchases all other goods: $x_{h,i\neq j} = 0$, $x^p_{h,i\neq j} > 0$, and $x_{h,j},\ x^s_{h,j} > 0$

[a]For any good i with $d \geq \sqrt{\frac{p_j z_{h,j}}{p_i z_{h,i}}}$ (flipping the inequality), whether i is produced or purchased cannot be determined analytically and becomes a numerical issue.

If d is low enough (below $\tilde{d}_h$), the household behaves in a canonical Ricardian way: it specializes its farm production fully in its comparative advantage good j and trades for all other goods, regardless of any characteristics of the farm.

Data. One implication of Proposition 5 above is that more active sellers should be more specialized, since they either drew a particularly good productivity in their comparative advantage good, or face lower trade frictions. I test this prediction in column (1) of Table 5, which shows the results of a regression of production diversity on farm commercialization, measured as the share of output value that is sold, controlling for farm size and non-farm income. It only includes farms that sell something, thus comparing more intensive sellers to less intensive ones, rather than sellers to non-sellers. More commercialized farms do indeed have more specialized production compared to their less commercialized peers of the same size and income.

A more direct prediction of the model is that farms facing lower trade costs should specialize their production. To test it in the data, I use a binary measure of market access based on the exact geographic distances between each farm and the nearest population center, agricultural market, and road, which are reported in the survey (whereas the exact physical location of each farm is not, to maintain anonymity). These can serve as measures of how costly it is for a given house-

Table 5: Commercialization, market access, and production diversity

	production diversity	
	(1)	(2)
sold output share	−0.044***	
	(0.016)	
$\mathbb{1}$[good mkt access]		−0.164***
		(0.018)
log(output value)	0.102***	0.180***
	(0.013)	(0.006)
log(non-farm income)	−0.051***	−0.042***
	(0.009)	(0.006)
N	4,042	8,675
Adj. R^2	0.025	0.099

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note. Output and non-farm income are relative to the household's kcal requirement. Sold output share is the share of output market value that the farm has sold. Farms with no sales are excluded from column (1). Good mkt access = 1 if the farm is in the bottom half of the three distance distributions (distance to nearest town, nearest agricultural market, and nearest road) at the same time.

hold to trade agricultural goods. I assign a given farm to a "good market access" group if the farm is in the bottom 50% of the distributions of all three distances at the same time, and to a "bad market access" group otherwise[28]. Column (2) of Table 5 shows that farms that are closer to towns, markets, or roads specialize their production relative to more remote farms of the same size and income.

6 Aggregate Productivity

Costly agricultural trade generates subsistence behavior in the model, tightly linking the production of a given farm with the consumption of its owners; explicitly modeled caloric needs create a tradeoff between calories, dietary variety, and non-food consumption. A combination of these two elements of the model generates

[28]Note that these groups are not balanced: the bad group includes more observations. Similarly constructed binary proxies that are balanced by design produce similar results.

predictions on the consumption behavior of households and on their farm product choice, many of which are borne out by the data.

Subsistence farmers produce not what they are best at relative to others, but what they want to see on their family's table. If the two do not coincide, the agricultural output of the economy will not be maximal. The model defined in Section 4 and explored and tested in Section 5 provides a framework to assess the quantitative importance of this mechanism.

In this section, I conduct counterfactual reductions in trade cost d—the ultimate driver of subsistence—and investigate how farmers' changing product choice—largely driven by nutrition demand—affects the economy.

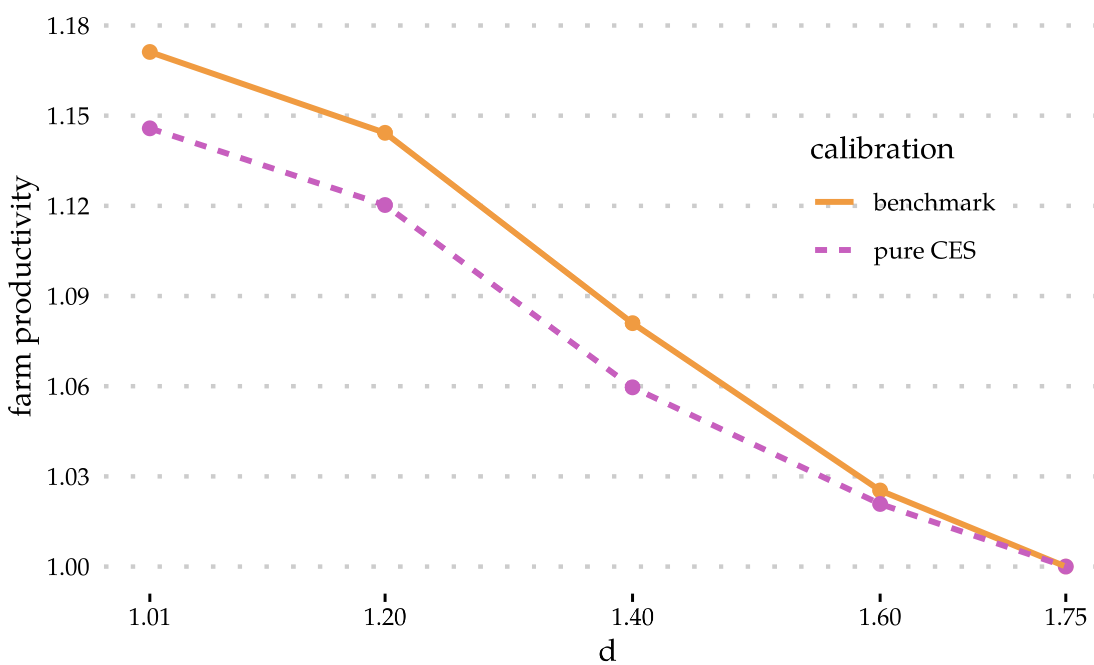
6.1 Falling Trade Costs Raise Farm Productivity

Figure 4 shows the effect of gradual trade cost reductions from the calibrated $d = 1.75$ level on real farm output productivity[29]. The output is measured at "farm gate": it does not account for any direct losses from d that sold goods are subject to.

Going to an allocation with costless trade in agricultural goods ($d = 1$) would raise farm productivities by 17.1%. All of this improvement is driven by the increasing alignment between the products individual farmers choose to grow and their comparative advantages.

Without the caloric channel ($\psi = 0$) this gain would be 14.6%. Reductions in trade costs still improve farm product choice in the pure CES model, since households in it still want to consume multiple agricultural products, which have to be grown at home when trade costs are high. But because farmers in the pure CES model do not prioritize calorie-rich goods above all else, they have more room to align their product choice with their productivity draws, meaning that the $d = 1.5$ allocation is already more efficient than in the benchmark model. Still, the overall difference to aggregate farm productivity introduced by the caloric channel specifically is modest, since the caloric deviation penalty affects most strongly the production decisions of the poorest farmers, who contribute little to the total agricultural output of the economy.

[29]Prices are used for weighting different goods, but changes between consecutive levels of d are deflated by chain-weighting.



NOTE. Farm productivity is the sum of farm outputs valued at market prices, divided by the sum of farm land endowments. Aggregate farm output is chain-weighted between each pair of consecutive d values to obtain real values. Productivity at the calibrated level of d is normalized to 1.

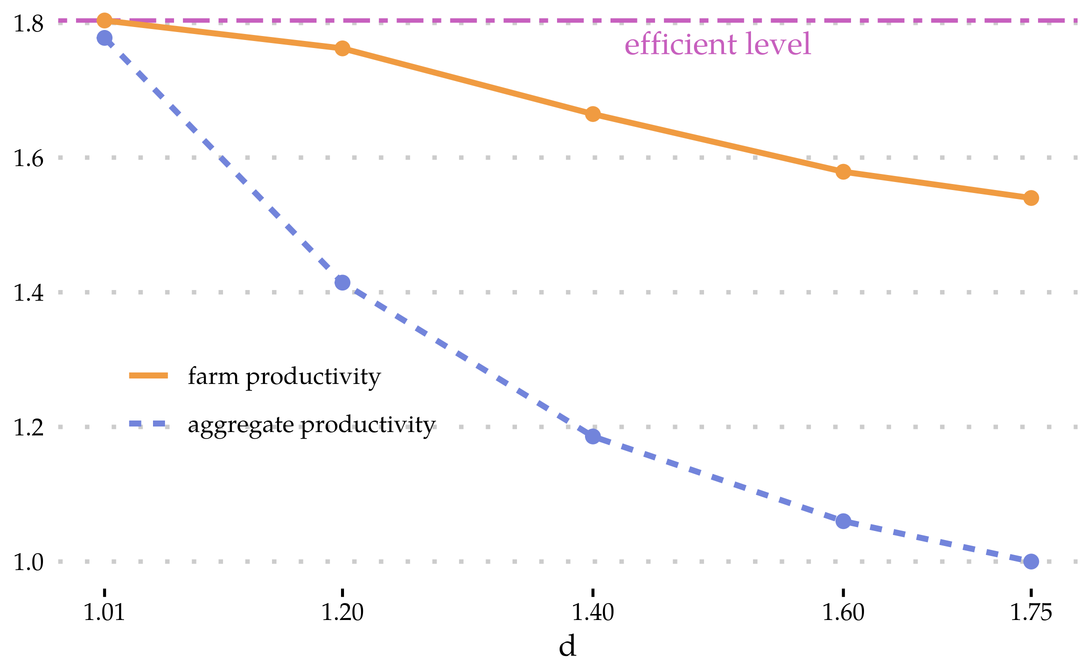
FIGURE 4: Farm productivity and trade cost

6.2 Mechanical Losses vs Product Choice

The response of farm output to trade cost changes captures the effect of shifting product choice, but not of changing mechanical losses to trade cost—the part of output that "melts" due to d. Final consumption of agricultural products[30], or agricultural GDP, captures both effects. Figure 5 displays the effect of trade cost reductions on both measures: the productivity of farm output and the productivity of aggregate agricultural GDP.

At the calibrated value of trade cost d, aggregate productivity is considerably more depressed than farm productivity alone, relative to the efficient level. Likewise, the potential gains from removing trade frictions are much greater: a 78% increase in agricultural GDP. Improved product choice is responsible for 34% of this, with the remaining 66% mechanically caused by lower losses to d.

[30] I include tobacco exports into final consumption.

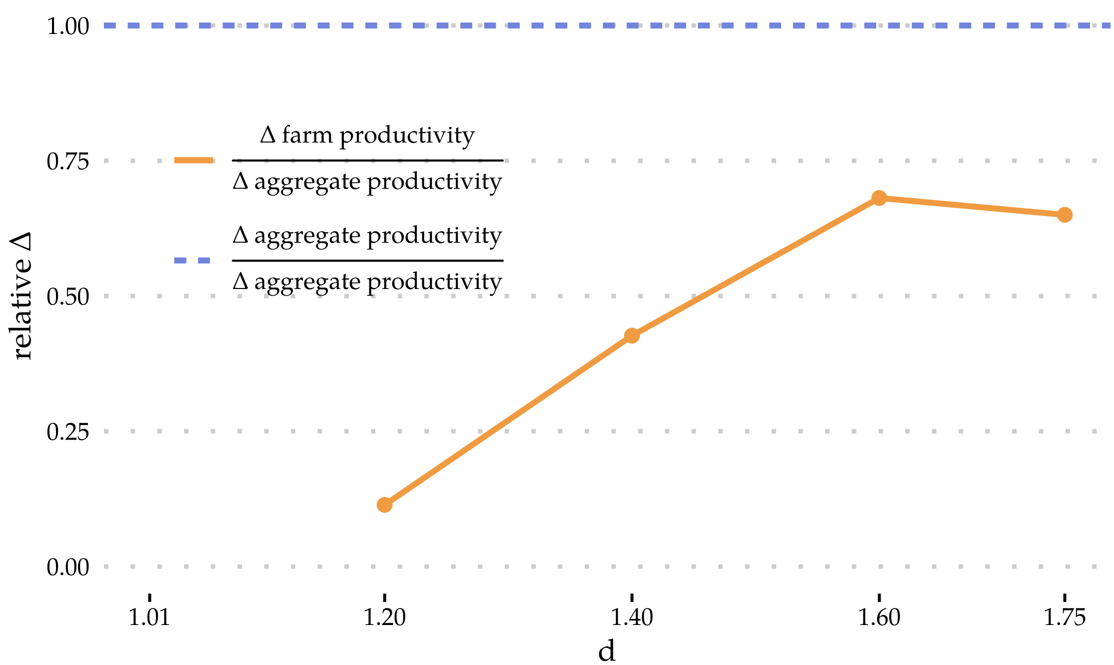


NOTE. Farm output and the final consumption of agricultural products are valued at market prices. Both series are chain-weighted between each pair of consecutive d values to obtain real values.

FIGURE 5: Farm productivity and aggregate agricultural productivity

Now consider a partial reduction in trade costs to the level that makes farms balance between subsistence and commercialized farming, with an average share of output sold of 50% (instead of the 16% at the calibrated $d = 1.75$). This intermediate stage is attained at $d = 1.20$. The model predicts that lowering the trade costs to that level would raise aggregate agricultural productivity by 42%, with improved product choice causing over half—53%—of this increase. Reductions in mechanical losses to trade costs play a smaller role in partial liberalizations simply because the volume of trade that d applies to is not yet that large.

This logic becomes even stronger for even smaller reductions in trade costs. Figure 7 displays the change in farm productivity relative to the change in aggregate productivity that results from marginal drops in trade cost d. At high levels of d, over 60% of productivity gains from marginal liberalizations are due to improvements in farmers' selection of products to grow.



NOTE. Productivity changes are between consecutive levels of d (marginal trade cost reductions).

FIGURE 6: Share of product choice changes in marginal productivity gains

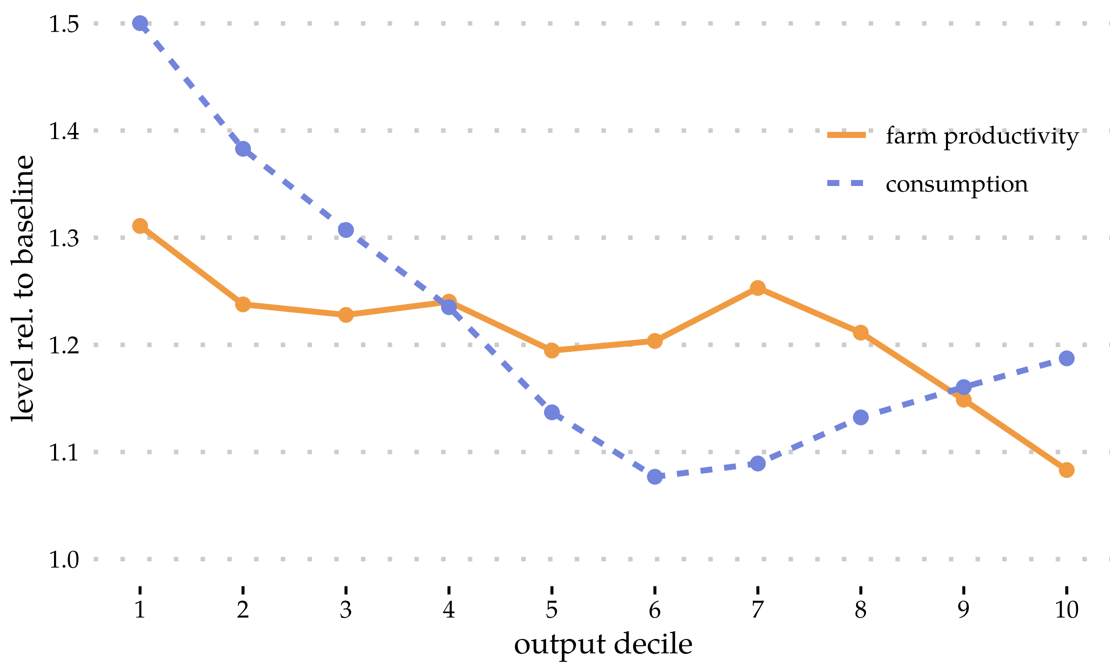
6.3 SMALLEST FARMERS ARE MOST AFFECTED

The product choice response of farms to a reduction in trade cost is heterogeneous among farms of different sizes. Figure 7 shows the heterogeneous responses to a trade cost reduction to an intermediate $d = 1.2$ level, which raises the average output share sold from 16% to 50%.

The smallest farmers respond the most, with their yields increasing by over 30% and their consumption value by 50%. The smallest farmers are the most calorically constrained, and at the baseline level of trade costs they specialize production heavily in the most calorically productive good. As trade costs fall to allow more commercialization, many small farmers switch to almost complete specialization in the most revenue-productive good, which allows them to buy more calories on the market than they used to manage to grow themselves.

Medium farmers respond less because they value food diversity relatively more: while partial trade cost reductions allow medium farmers to start buying some products at the market instead of growing them, certain products will still be cheaper to produce at your farm.

Finally, the productivity of the largest farmers responds the least, since their product choice was well-aligned with their comparative advantage even at the baseline high trade cost level. Their consumption value, however, responds stronger than that of the medium farmers, since falling trade costs benefit active sellers by increasing the revenue they get from the same quantity of goods shipped.



Note. Farm output value and household consumption value is deflated by chain-weighting between the two allocations for each household individually.

Figure 7: Response of farm productivity and household consumption to a 1.75 → 1.2 reduction in d, by farm size decile

7 Conclusion

Subsistence farming is prevalent in low-income countries. The production decisions of subsistence farmers are driven by the need to feed their family from their own land. I capture this idea in a model of farm-operating households that face a caloric deviation penalty in addition to standard CES preferences. Explicit nutritional needs generate a tradeoff between obtaining dietary energy, satisfying tastes and love of variety, and consuming purchased non-food goods. In the presence of domestic trade frictions, this nutrition-driven tradeoff in consumption, rather than comparative advantage alone, starts determining also the production decisions of these farmers.

I test the predictions of the model in Malawian household-level data and find that the model matches several patterns in the observed behavior of Malawian farmers. Both in the model and in the data, households with small farms have relatively specialized diets; they specialize their production heavily as well, usually in maize or another staple crop; and they consume virtually their entire output. Medium farmers shift the focus of their diet from calories to diversity and correspondingly diversify their production away from staples by increasing the range of edible products. Large farmers increasingly orient their farm to producing goods destined for the market, and are the most active sellers. The caloric deviation penalty is the key element of the model that lets it reproduce these empirical patterns in the consumption and production behavior of subsistence farmers.

Malawian farmers sell just a minor fraction of their output. The model suggests that lowering trade frictions to raise the average share of output sold to just 50% would make the country's agricultural sector 42% more productive—partly by reducing the direct burden of having to pay the trade cost, partly by allowing farmers to better align their product choice with their comparative advantage rather than their family's nutritional needs and food preferences.

The model was constructed to analyze how subsistence farmers allocate resources between different goods within one sector—agriculture. A useful development of the model would include household labor choice between working on the family farm, working for wages on someone else's farm, or working in the manufacturing sector. Extended in this way, the model would be useful for studying structural transformation patterns both within the agricultural sector and across sectors on the path of development, potentially adding nuance to both dimensions.

Risk would be another fruitful addition to the model. Extended with volatile harvests and prices, the model would permit the study of the interaction of riskiness and nutritional concerns in driving the product choice of farmers: e.g. adjusting the crop portfolio to reduce the risk of starvation.

The framework developed in this paper is well suited for analyzing nutritionally sensitive programs aimed at smallholder farmers. These are central to the public policy of low-income countries like Malawi. For instance, much of Malawi's agricultural policy has revolved around supporting smallholder farmers in the production of staples or, to a lesser extent, tobacco (Levy, 2005; Chibwana et al., 2014). Meanwhile, some researchers argue that promoting biodiversity can be a more effective way of bolstering the food security of smallholder farmers (Jones, 2017; Pingali and Sunder, 2017). Since the model this paper develops combines explicit nutritional needs with endogenous farm product choice that aims to fulfill those needs, it can be used to predict and compare the effects of encouraging staples, cash crops, or biodiversity, and to determine which environments each policy is best suited for.

References

Allen, Treb and David Atkin (2016) "Volatility and the Gains from Trade," working paper.

Behrman, Jere R. and Anil Deolalikar (1989) "Is Variety the Spice of Life? Implications for Calorie Intake," *The Review of Economics and Statistics*, Vol. 71, No. 4, p. 666.

Blanco, Cesar and Xavier Raurich (2019) "Agricultural Composition and Labor Productivity," working paper.

Charrondiere, U. Ruth, David Haytowitz, and Barbara Stadlmayr (2012) *FAO/INFOODS Density Database Version 2.0*.

Chibwana, Christopher, Gerald Shively, Monica Fisher, Charles Jumbe, and William Masters (2014) "Measuring the impacts of Malawi's farm input subsidy programme," *African Journal of Agriculture and Resource Economics*, Vol. 9, No. 2, pp. 132–147.

Drewnowski, Adam (2010) "The Nutrient Rich Foods Index helps to identify healthy, affordable foods," *The American Journal of Clinical Nutrition*, Vol. 91, No. 4, pp. 1095S–1101S.

Drewnowski, Adam, Colin Rehm, and Florent Vieux (2018) "Breakfast in the United States: Food and Nutrient Intakes in Relation to Diet Quality in National Health and Examination Survey 2011–2014. A Study from the International Breakfast Research Initiative," *Nutrients*, Vol. 10, No. 9, p. 1200.

FAO (2004) *Human energy requirements: report of a joint FAO/WHO/UNU Expert Consultation*, Rome.

FAO and IIASA (2021) "Global Agro Ecological Zones version 4 (GAEZ v4)."

Fulgoni, Victor L., Debra R. Keast, and Adam Drewnowski (2009) "Development and Validation of the Nutrient-Rich Foods Index: A Tool to Measure Nutritional Quality of Foods," *The Journal of Nutrition*, Vol. 139, No. 8, pp. 1549–1554.

Gollin, Douglas, David Lagakos, and Michael E. Waugh (2014) "Agricultural Productivity Differences across Countries," *American Economic Review*, Vol. 104, No. 5, pp. 165–170.

Gollin, Douglas and Richard Rogerson (2014) "Productivity, transport costs and subsistence agriculture," *Journal of Development Economics*, Vol. 107, pp. 38–48.

van Graan, Averalda, Joelaine Chetty, Jumat Malory, Sitilitha Masangwi, Agnes Mwangwela, Felix Pensulo Phiri, Lynne M. Ausman, Shibani Ghosh, and Elizabeth Marino-Costello eds. (2019) *MAFOODS. 2019. Malawian Food Composition Table.*, Lilongwe, Malawi, 1st edition.

Jones, Andrew D (2017) "On-Farm Crop Species Richness Is Associated with Household Diet Diversity and Quality in Subsistence- and Market-Oriented Farming Households in Malawi," *The Journal of Nutrition*, Vol. 147, No. 1, pp. 86–96.

Kebede, Hundanol A. (2020a) "The gains from market integration: The welfare effects of new rural roads in Ethiopia," working paper.

——— (2020b) "Market Integration and Separability of Production and Consumption Decisions in Farm Households," working paper.

Levy, Sarah (2005) *Starter packs: a strategy to fight hunger in developing countries?*, Wallingford, UK; Cambridge, MA: CABI Pub. OCLC: 61751260.

Lukmanji, Zohra, Ellen Hertzmark, Nicolas Mlingi, Vincent Assey, Godwin Ndossi, and Wafaie Fawzi eds. (2008) *Tanzania Food Composition Tables*, Dar es Salaam Tanzania: MUHAS, TFNC, HSPH.

MacDonald, James M., Penni Korb, and Robert A. Hoppe (2013) "Farm Size and the Organization of U.S. Crop Farming," Economic Research Report 152, U.S. Department of Agriculture, Economic Research Service.

Msyamboza, Kelias P., Damson Kathyola, and Titha Dzowela (2013) "Anthropometric measurements and prevalence of underweight, overweight and obesity in adult Malawians: nationwide population based NCD STEPS survey," *Pan African Medical Journal*, Vol. 15.

Pingali, Prabhu and Naveen Sunder (2017) "Transitioning Toward Nutrition-Sensitive Food Systems in Developing Countries," *Annual Review of Resource Economics*, Vol. 9, No. 1, pp. 439–459.

Rivera-Padilla, Alberto (2020) "Crop choice, trade costs, and agricultural productivity," *Journal of Development Economics*, Vol. 146, p. 19.

Sibhatu, Kibrom T., Vijesh V. Krishna, and Matin Qaim (2015) "Production diversity and dietary diversity in smallholder farm households," *Proceedings of the National Academy of Sciences*, Vol. 112, No. 34, pp. 10657–10662.

Sotelo, Sebastian (2020) "Domestic Trade Frictions and Agriculture," *Journal of Political Economy*, Vol. 128, No. 7, pp. 2690–2738.

USDA and HHS (2020) *Dietary Guidelines for Americans, 2020-2025*: U.S. Department of Agriculture and U.S. Department of Health and Human Services, 9th edition.

Appendix

A Nutrient Richness in the Data

To evaluate the nutrient intakes of sample households, I use the Nutrient Rich Foods index (NRF), developed to assess the nutrient density of individual foods (Fulgoni et al., 2009; Drewnowski, 2010), but applied since then to assessing the quality of the overall diet, for example by Drewnowski et al. (2018).

The most utilized version of the Nutrient Rich Foods index is NRF9.3, which is based on an individual's intake of 9 qualifying nutrients (protein, fiber, vitamins A, C, and D, calcium, iron, potassium, and magnesium) and 3 disqualifying nutrients (added sugar[31], saturated fat, and sodium). For each of the qualifying nutrients, the index is based on the intake of the nutrient relative to the recommended daily allowance (RDA) for the individual, capped at 1 (so that ample consumption of one nutrient does not compensate for deficiencies in other nutrients). For each of the disqualifying nutrients, the index is based on the relative intake in excess of the maximum recommended value (MRV).

$$\text{NRF9.3} = \left(\sum_{q \in Q} \min \left\{ \frac{Intake_q}{RDA_q},\ 1 \right\} - \sum_{d \in D} \max \left\{ \frac{Intake_d}{MRV_d} - 1,\ 0 \right\} \right) \times 100$$

where Q is the set of qualifying nutrients and D is the set of disqualifying nutrients. The maximum possible value of the index is 900.

I also use a simpler version of the index: NRF9, which excludes the disqualifying nutrients instead of penalizing for them:

$$\text{NRF9} = \left(\sum_{q \in Q} \min \left\{ \frac{Intake_q}{RDA_q},\ 1 \right\} \right) \times 100$$

Table A.1 shows the correlations between NRF9 and NRF9.3, as well as the two measures of dietary variety (# foods and food diversity). NRF9 is reasonably correlated with the measures of dietary variety, but NRF9.3, which penalizes for

[31] While the Malawian food composition table reports the added sugar of foods, the Tanzanian food composition table does not. For a small fraction of foods whose nutritional composition I take from the Tanzanian FCT, I use total sugar in lieu of added sugar.

"bad" nutrients, is only weakly correlated.

TABLE A.1: Dietary variety and nutrient richness measures correlation matrix

	# foods consumed	food diversity	NRF9	NRF9.3
# foods consumed	1	.	.	.
food diversity	.82	1	.	.
NRF9	.39	.34	1	.
NRF9.3	.06	.08	.18	1

NOTE. # foods consumed is a count of distinct foods consumed by the household. Food diversity is a diversity index applied to the market values (physical quantity × median price) of distinct foods consumed by the household. NRF9 is the Nutrient-Rich Food Index, a sum of the ratios of daily values of 9 qualifying nutrients. NRF9.3 additionally subtracts the relative consumption in excess of maximum recommended daily values of 3 disqualifying nutrients.

In Table A.2, I regress the two NRF variants on farm size and non-farm income, optionally controlling for energy intake to remove the mechanical correlation between the overall quantity of food consumed and the incidental quantity of nutrients in that food. Like dietary variety, nutrient richness is increasing in farm size and non-farm income, but far more of the variation in nutrient richness is left unexplained, especially for NRF9.3. These results suggest that wealthier households indeed consume a more diverse and nutrient-rich diet (even when controlling for their energy consumption), but they don't necessarily try to limit their consumption of undesirable nutrients.

TABLE A.2: Nutrient richness measures vs size and income

	NRF9		NRF9.3	
	(1)	(2)	(3)	(4)
log output	17.046***	5.695***	−13.296***	−13.400***
	(0.964)	(0.724)	(3.326)	(3.358)
log non-farm income	10.285***	2.441***	−7.257**	−7.305**
	(0.792)	(0.603)	(3.898)	(3.548)
log kcal intake		124.025***		0.550
		(2.282)		(26.234)
N	8,675	8,674	8,675	8,674
Adj. R^2	0.054	0.451	0.002	0.002

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

NOTE. Output, non-farm income, and kcal intake are relative to the caloric requirement of the household. NRF9 is the Nutrient-Rich Food Index, a sum of the ratios of daily values of 9 qualifying nutrients. NRF9.3 additionally subtracts the relative consumption in excess of maximum recommended daily values of 3 disqualifying nutrients.

B PROOFS

Common Notation. Let λ be the Lagrange multiplier associated with the land constraint (2), μ be the Lagrange multiplier associated with the budget constraint (3), η_i be the multiplier associated with good i's resource constraint (4), and $\theta(y)$ be the multiplier associated with the nonnegativity constraint for any variable y within (5). All these multipliers are nonnegative.

Proof of Proposition 1.

The equality of marginal costs between goods i and j for household h means $\eta_{h,i} = \eta_{h,j}$.

First consider the case of $\sum_i c_{h,i}k_i < K_{req,h}$. Because both intake and requirement are positive,

$$\frac{K_{req,h}}{\left(\sum_i c_{h,i}k_i\right)^2} > \frac{K_{req,h}}{K_{req,h}^2} = \frac{1}{K_{req,h}}.$$

Hence, the derivative of f w.r.t the caloric intake is negative:

$$f_1\left(\sum_i c_{h,i}k_i, K_{req,h}\right) = \psi\left(\frac{1}{K_{req,h}} - \frac{K_{req,h}}{\left(\sum_i c_{h,i}k_i\right)^2}\right) < 0$$

Since we assumed $\varphi_i = \varphi_j$, $\eta_{h,i} = \eta_{h,j}$, and $k_i > k_j$,

$$\varphi_i\left(k_j f_1\left(\sum_i c_{h,i}k_i, K_{req,h}\right) + \eta_j\right) > \varphi_j\left(k_i f_1\left(\sum_i c_{h,i}k_i, K_{req,h}\right) + \eta_i\right).$$

Hence,

$$\frac{c_{h,i}}{c_{h,j}} = \left(\frac{\varphi_i}{\varphi_j}\frac{k_j f_1\left(\sum_i c_{h,i}k_i, K_{req,h}\right) + \eta_{h,j}}{k_i f_1\left(\sum_i c_{h,i}k_i, K_{req,h}\right) + \eta_{h,i}}\right)^{\sigma} > 1$$

Considering instead the case of $\sum_i c_{h,i}k_i > K_{req,h}$, all inequalities flip and ultimately $\frac{c_{h,i}}{c_{h,j}} < 1$.

Proof of Proposition 2.

Because good i is assumed to be produced by household h, the first order condition for $c_{h,i}$ is

$$A\varphi_i c_{h,i}^{-\frac{1}{\sigma}} = k_i f_1\left(\sum_i c_{h,i}k_i, K_{req,h}\right) + \frac{\lambda_h}{z_{h,i}}$$

where

$$A = \left((1-\varphi_m)\left(\sum_{i=1}^{n}\varphi_i c_{h,i}^{\frac{\sigma-1}{\sigma}}\right)^{\frac{\sigma}{\sigma-1}\frac{\gamma-1}{\gamma}} + \varphi_m c_{h,m}^{\frac{\gamma-1}{\gamma}}\right)^{\frac{1}{\gamma-1}}(1-\varphi_m)\left(\varphi_i c_{h,i}^{\frac{\sigma-1}{\sigma}}\right)^{\frac{\sigma-\gamma}{(\sigma-1)\gamma}}$$

Combine the FOCs for $c_{h,i}$ and $c_{h,j}$ (which was also assumed to be produced) through λ_h:

$$A\varphi_i c_{h,i}^{-\frac{1}{\sigma}} z_{h,i} - z_{h,i} k_i f_1\left(\sum_i c_{h,i} k_i, K_{req,h}\right) = \lambda_h = A\varphi_j c_{h,j}^{-\frac{1}{\sigma}} z_{h,j} - z_{h,j} k_j f_1\left(\sum_i c_{h,i} k_i, K_{req,h}\right)$$

First consider the case of $\sum_i c_{h,i} k_i < K_{req,h}$. As shown in the Proof of Proposition 1, $f_1\left(\sum_i c_{h,i} k_i, K_{req,h}\right) < 0$. Combining this fact with the assumption that $k_i z_{h,i} > k_j z_{h,j}$,

$$k_i z_{h,i} f_1\left(\sum_i c_{h,i} k_i, K_{req,h}\right) < k_j z_{h,j} f_1\left(\sum_i c_{h,i} k_i, K_{req,h}\right)$$

This in turn implies

$$A\varphi_i c_{h,i}^{-\frac{1}{\sigma}} z_{h,i} < A\varphi_j c_{h,j}^{-\frac{1}{\sigma}} z_{h,j}$$

Finally,

$$\frac{c_{h,i}}{c_{h,j}} > \left(\frac{\varphi_i z_{h,i}}{\varphi_j z_{h,j}}\right)^{\sigma} = \frac{c_{h,i}^{CES}}{c_{h,j}^{CES}}$$

In the case of $\sum_i c_{h,i} k_i > K_{req,h}$, all inequalities are flipped.

Proof of Proposition 3.

Coming soon!

Proof of Proposition 4.

First of all, household h can never purchase and sell the same good: if we suppose that $x_{h,i}^p$ and $x_{h,i}^s$ are both positive, then $\theta(x_{h,i}^p) = \theta(x_{h,i}^s) = 0$ and combining the first order conditions with respect to $x_{h,i}^p$ and $x_{h,i}^s$ yields $\mu_h p_i d = \mu_h p_i \frac{1}{d}$, which is impossible since $d > 1$.

Therefore, since good j is sold ($x_{h,j}^s > 0$), it must be produced: $\theta(x_{h,j}^s) = \theta(x_{h,j}) = 0$. Merging the $x_{h,j}^s$ and $x_{h,j}$ first order conditions yields $\frac{\lambda_h}{\mu_h} = \frac{p_j z_{h,j}}{d}$.

Now, for any good i, merging the $x_{h,i}^s$ and $x_{h,i}$ first order conditions (without assuming that these two variables are positive) yields

$$\frac{\lambda_h}{\mu_h} = \frac{p_i z_{h,i}}{d} + \frac{\theta(x_{h,i}^s) z_{h,i}}{\mu_h} + \frac{\theta(x_{h,i}) z_{h,i}}{\mu_h}$$

where all terms on the right-hand side are non-negative.

Therefore, for any good i,

$$\frac{p_i z_{h,i}}{d} \leq \frac{\lambda_h}{\mu_h} = \frac{p_j z_{h,j}}{d}$$

Hence, the sold good j satisfies $p_j z_{h,j} \geq p_i z_{h,i}$ for any other good i.

Therefore, $j = \arg\max_i p_i z_{h,i}$.

Proof of Proposition 5.

Let j be the most revenue-productive good and let i be some other good.

Merging the $x_{h,i}$ and $x^p_{h,i}$ first order conditions, and using a result from the Proof of Proposition 4,

$$\frac{p_j z_{h,j}}{d} = \frac{\lambda_h}{\mu_h} = p_i z_{h,i} d - \frac{\theta(x^p_{h,i}) z_{h,i}}{\mu_h} + \frac{\theta(x_{h,i}) z_{h,i}}{\mu_h}$$

Suppose i is produced. Then $\theta(x_{h,i}) = 0$ and

$$\frac{p_j z_{h,j}}{d} = \frac{\lambda_h}{\mu_h} \leq p_i z_{h,i} d$$

This implies

$$d \geq \sqrt{\frac{p_j z_{h,j}}{p_i z_{h,i}}}$$

Therefore, if $d < \sqrt{\frac{p_j z_{h,j}}{p_i z_{h,i}}}$, then i must not be produced. If $\varphi_i > 0$, then $c_{h,i} > 0$ and i has to be purchased: $x^p_i > 0$.

If furthermore $d < \tilde{d}_h \equiv \sqrt{\frac{p_j z_{h,j}}{\max_{i \neq j} p_i z_{h,i}}}$, then no good i satisfies this condition: only j is produced.

C Computation

Coming soon!

D Additional Tables

Table D.3: Dietary diversity measures vs size and income

	# foods consumed		food diversity	
	(1)	(2)	(3)	(4)
log output	0.651***	0.381***	0.395***	0.343***
	(0.062)	(0.060)	(0.034)	(0.034)
log non-farm income	1.422***	1.235***	0.857***	0.821***
	(0.059)	(0.056)	(0.033)	(0.032)
log kcal intake		2.964***		0.571***
		(0.119)		(0.065)
N	8,675	8,674	8,675	8,674
Adj. R^2	0.109	0.171	0.131	0.138

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note. Output, non-farm income, and kcal intake are relative to the caloric requirement of the household. # foods consumed is a count of distinct foods consumed by the household. Food diversity is a diversity index applied to the market values (physical quantity × median price) of distinct foods consumed by the household.

E Additional Figures

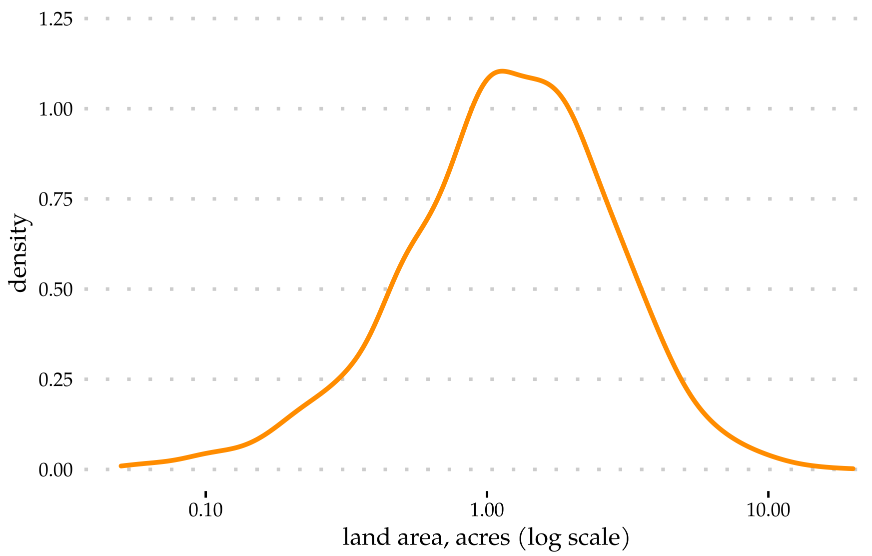
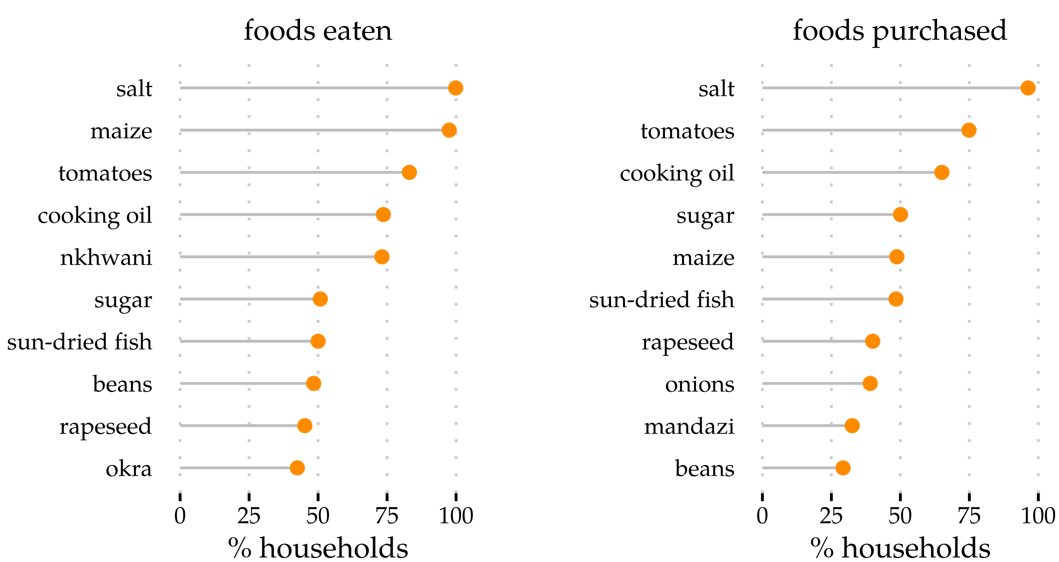
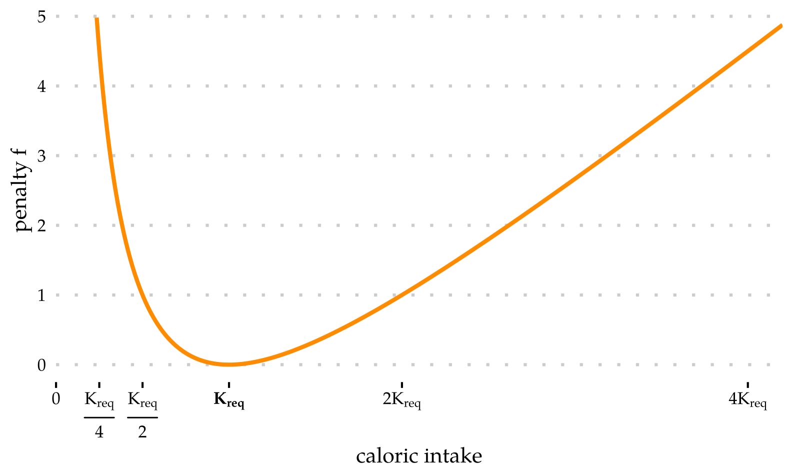


Figure E.1: Farm area distribution



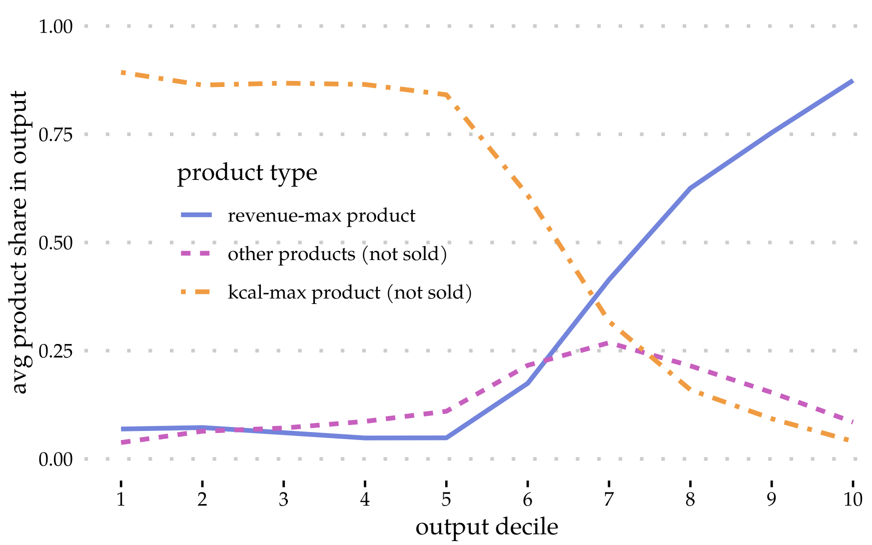
Note. Foods are ranked by the % of households consuming each food (left) or purchasing each food (right), and only the top 10 are displayed.

Figure E.2: Foods ranked by the % of households consuming and purchasing them



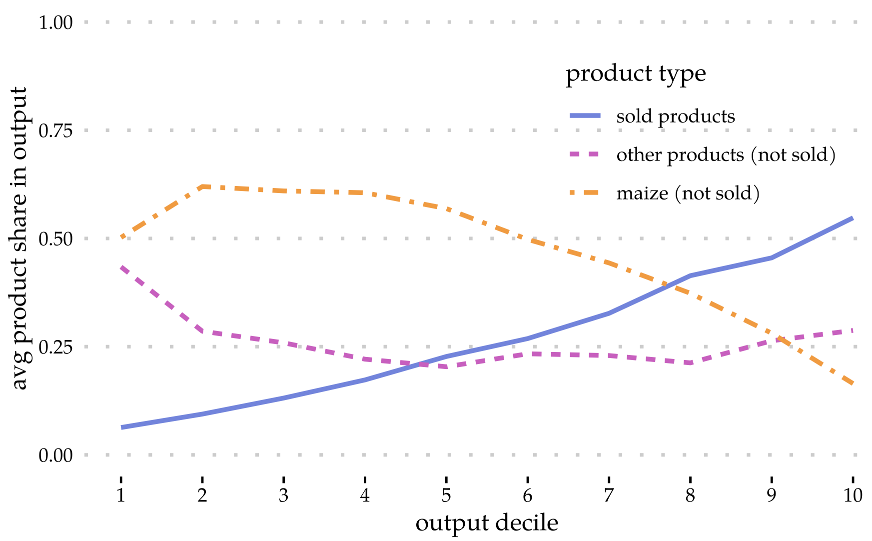
NOTE. $f\left(\sum_i c_i k_i, K_{req}\right) = \psi\left(\frac{\sum_i c_i k_i - K_{req}}{K_{req}}\right)^2 \frac{K_{req}}{\sum_i c_i k_i}$. Value $\psi = 2$ used for illustration. The x-axis is caloric intake expressed as ratios of caloric requirement K_{req}.

FIGURE E.3: Penalty function f for deviations of caloric intake from caloric requirement



NOTE. Products are split into groups at farm level. "kcal-max product (not sold)" is the $\arg\max_i k_i z_{h,i}$ good, unless it's the same as the $\arg\max_i p_i z_{h,i}$ and is sold. "revenue-max product" is the $\arg\max_i p_i z_{h,i}$ good, unless it's the same as the $\arg\max_i k_i z_{h,i}$ and is not sold. "other products (not sold)" are all other goods. Avg product shares are the average output value shares of each product group among farms in each output value decile.

FIGURE E.4: Farm size and product choice (average shares): model



Note. Products are split into groups at farm level. Avg product shares are the average output value shares of each product group among farms in each output value decile.

Figure E.5: Farm size and product choice (average shares): data