Nutrition Demand, Subsistence Farming, and Agricultural Productivity

Stepan Gordeev*

December 16, 2024

click for latest version

Abstract

In many of the poorest countries, agriculture is unproductive and subsistence farming is widespread. 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 47% 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, Francesco Cecchi, Gaston Chaumont, Paulo Lins, Kristina Manysheva, Jeffrey Michler, Roman Merga, Walter Steingress, and the participants of AFES21, NCDE21, YES, SEA, STEG PhD Workshop, ASSA 2023, CSAE, MWIEDC, SED 2024 Winter Meeting, the international reading group and the lunch seminar at the University of Rochester for their comments and suggestions. I thank STEG, a program funded by FCDO, for funding support. The views expressed are not necessarily those of FCDO.

*Texas Christian University, s.gordeev@tcu.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 for understanding the enormous income differences across countries (Gollin, Lagakos, and Waugh 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 over three quarters of all households in Malawi 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 farmers' demand for nutrition, driven by dietary energy needs, as a force that can explain much of the consumption, production, and selling behavior of farmers when trade is costly. I argue that the subsistence-driven product choice of individual farmers ultimately contributes to keeping the productivity of the whole agricultural sector low.

In an environment where domestic trading frictions make subsistence behavior at the level of individual farms common, farmers may choose not to specialize in their profit-maximizing comparative advantage product, but rather to grow products that directly satisfy their own family's demand for food. The latter 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, 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 farms. I find that there is significant subsistence behavior even at the level of in-

1. World Development Indicators.

dividual households: most sell almost none of their agricultural output and rely on their farm as an important source of food. This self-oriented production behavior suggests that many farmers are unable to specialize in their comparative advantage and to commercialize the farm. Moreover, I document that the consumption, production, and selling behavior of farmers is dependent on the scale of their farms: smallest farmers specialize consumption and subsistence production in calories, medium farmers diversify both, and large farmers commercialize their production.

Motivated by this self-reliance and scale-dependence in the data, I develop a model of farm-operating households facing caloric needs and trading costs. 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. This penalty is the main novelty of the model.

The combination of CES preferences with the caloric deviation penalty makes the household's consumption allocation across goods depend on the size of its farm and income relative to its caloric requirement. Furthermore, in the presence of a proportional trading cost the nutritional situation of the household 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, permitting them to increasingly shift their consumption to the purchased non-food good, which requires growing and selling their comparative advantage farm 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, trade cost, and good-specific productivity draws. They consume multiple agricultural goods whose caloric densities, taste weights, and productivity distri-

butions 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 they do consume significantly more diverse (and, in the data, nutrient-rich) diets.

Next, I test the predictions of the model on selling behavior. In the model, households sell either no goods or just the single good they have comparative advantage in. However, unless trade costs are low enough to permit full specialization, they also engage in subsistence production of 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, overall production by survey households with better market access is more specialized and less correlated with their consumption choices.

Both in the model and in the data, larger farmers are more likely to be sellers and sell more on average. The main reason for this behavior in the model is that larger farmers are less calorically constrained and want to shift consumption to the manufactured good (offered on the market), which requires 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 are subsistence farmers who target their own dietary needs with their farm product choice.

I use the model to test the importance of subsistence-driven product choice by individual farmers for keeping the aggregate agricultural productivity of Malawi

low. The model suggests that a reduction in trade costs sufficient to allow Malawian farmers to increase 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 47%, realizing over half of the productivity gains promised by completely costless domestic trade. Just over half of this 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 demand for dietary energy or dietary diversity. The productivity, consumption value, and calorie intake of the smallest farmers, who are originally the most calorically constrained, respond the most to this reduction. 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.

To disentangle the role of subsistence itself from its interaction with nutrition demand in depressing the aggregate agricultural productivity, I repeat the quantitative exercises in two versions of the model that replace the caloric deviation penalty with alternative commonly used preferences specifications: CES-only or Stone-Geary. I find that the implications of the two alternative models on aggregate productivity are comparable to those of the proposed caloric needs model. This implies that while considering nutrition is powerful for understanding the micro-behavior of individual farmers and their heterogeneous response to counterfactual policy, it is the fact of subsistence—not its interaction with nutrition demand—that matters the most for the macro-level aggregate productivity.

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 (2024) 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 (Kebede 2024) or regions (Gollin and Rogerson 2014; Rivera-Padilla 2020; Sotelo 2020). Using Malawian household-level data, I highlight significant subsistence behavior even at the level of individual households, whereas the literature above has modeled crop allocation and trade as being efficient within regions or at least villages.[2] Secondly, I present a novel set of facts documenting the scale-dependent product choice by Malawian farmers in both their consumption and production: smallest farmers specialize diets in calories and subsistence production in maize, medium farmers diversify subsistence production, and large farmers commercialize the farm. Finally, I build a model in which heterogeneous farm-operating households[3] allocate their diet and production between heterogeneous products while facing explicit caloric needs. This novel nutritional channel generates a tradeoff between dietary energy, dietary diversity, and non-food consumption: this tradeoff is absent from existing models with CES (Sotelo 2020; Kebede 2024) or Stone-Geary (Gollin and Rogerson 2014; Rivera-Padilla 2020) preferences and allows the model to make endoge-

2. Household-level subsistence has been empirically documented outside of the aggregate productivity literature: Kebede (2022) is a recent example.

3. For modeling purposes, household-level subsistence being common in the data implies that agents may be unable to split production among multiple producers according to their comparative advantage: each farm-operating household has to tailor production to its *own* demand, not to the demand of a representative consumer at a higher level of aggregation.

nous predictions on how the composition of the household's diet depends on its nutritional situation, caloric densities and costs of agricultural products. When coupled with costly domestic trade, the nutritional channel allows the model to explain the salient empirical patterns in the consumption, production, and selling behavior of farmers that I document and that existing models are silent on.

I also contribute to the literature on forces that can affect farm product choice in low-income countries. Blanco and Raurich (2022) explore the reallocation of farmland toward capital-intensive crops along the path of development. Allen and Atkin (2022) 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), which shows that trade costs induce farmers to reallocate production toward a staple crop, which is needed to satisfy a Stone-Geary subsistence constraint and also has lower trade costs compared to a cash crop. I contribute to this literature by showing that explicitly modeling the caloric needs of farmers choosing between crops with objectively measured nutritional values in the presence of costly domestic trade can endogenously (without having to impose subsistence requirements for specific crops) explain the allocation between any number of agricultural products (not just two) and reproduces several empirical patterns in product choice and selling behavior of individual farmers (which Stone-Geary preferences are largely unsuccessful at, as I discuss in Section 5.1).

Farmers' incentives to shift from specializing in a single cash crop to diversifying production between a cash crop and a staple crop grown for own consumption in response to trade costs has also been explored in agricultural economics, most notably in Omamo (1998a) and Omamo (1998b). The finding that smaller farmers devote a greater share of their farm to growing a staple crop rather than a cash crop has also recently been documented in Li (2023) and Haque (2022), which they explain with a combination of trade costs and non-homothetic preferences between staple and cash crops. These mechanisms are also present and important in my model. But by examining the choice among many heterogeneous crops and explicitly modeling the caloric channel, I am able to explore and explain the nature of diversification not only between cash production and subsistence production overall, but also across different goods *within* subsistence production. Moreover, the caloric channel can simultaneously explain several facts in the scale

dependence of the consumption behavior of households as well, and to make endogenous predictions on which farmers choose to produce and consume which agricultural goods, without having to impose the staple-cash distinction on crops. Finally, the model allows me to explore the implications of this farm-level behavior for the productivity of the whole sector.

Subsistence farmers' nutritional outcomes have been the subject of several studies in nutritional science. In particular, Sibhatu, Krishna, and Qaim (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 subsistence farming for aggregate agricultural productivity. Section 7 concludes.

2 Data

Around two billion people around the world are smallholder farmers. Many of them are food-insecure and engage in subsistence farming to directly provide food for the family's table (Rapsomanikis 2015). Malawi is one country in sub-Saharan Africa that exemplifies this problem: most of its population is occupied in smallholder farming and suffers from food insecurity; the agricultural sector is unproductive and agricultural markets fail to provide adequate access to food (Benson 2021). I explore the interplay between household food consumption, farm production, and selling behavior in Malawian household-level 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 and Land. Survey enumerators manually measured the total area of land cultivated by each household using GPS. I use this as a measure of farm area.[4]

Households report total agricultural output by product in the last rainy season and the last dry season[5] 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 a sizeable fraction does not.[6] 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.[7] This lets me construct household-product output and sales observations. Most agricultural products reported in the survey are crops, but some are animal products (such as meat, milk, and eggs), which I also include in the empirical analysis. In total, 45 products appear in the data: these are listed in Appendix Table D.1, ranked by the share of households growing each.[8]

As will be discussed in Section 3, there is very little selling of agricultural

4. For some, GPS measurements are unavailable, so I take the farmers' self-reported land area.
5. Together these amount to roughly one year.
6. For example, maize is often measured in pails, tobacco in bales, and bananas in bunches.
7. Ultimately, I am able to obtain a weight estimate for 95% of household-product output observations.
8. 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*).

products in this sample. Revenue is therefore not a useful 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 sale price[9] 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.

I use output value instead of land area to measure farm size (and product choice) in the data, as the latter would ignore land quality differences across farms and would require ignoring many commonly produced and consumed products that do not use land as an explicitly measurable input (like chicken eggs).

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, Haytowitz, and Stadlmayr 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[10]) of each observation using the nutritional facts about each food.

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

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

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

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, Kathyola, and Dzowela (2013). Summing the physiological energy needs of each person in a family gives the total *caloric requirement* of the household.[11]

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. I construct weekly intakes of several nutrients analogously to caloric intakes. 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.

11. This procedure ignores two sources of potentially predictable variation in total daily energy expenditure (TDEE) across individuals that are not captured in the data: body weight and physical activity. Any differences in these two variables that are systematically related to the intensity of farming done by the household would affect the exact shape of the relationship between calorie intake and farm size discussed below in Section 5.1, but would not affect the model's ability to explain observed behavior. Note, however, that the amount of physical activity in active non-sedentary adults appears to have little to no relationship with TDEE due to the body endogenously adjusting its energy expenditure on non-physical activity to keep the total energy expenditure stable (Pontzer et al. 2016). This suggests that the systematic relationship between true unobserved TDEE and intensity of farming activities is likely to be limited within the sample of farm-operating households.

GAEZ Attainable Yields. The Global Agro-Ecological Zones (GAEZ) dataset produced by FAO and IIASA (2021) provides crop-location-specific estimates of potentially 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.[12] I use the crop-specific distribution of attainable physical yields across locations[13] in Malawi as a source of productivity distributions for the calibration of the model.

3 Subsistence in the Data

Malawian Farms are Small. The distribution of farm areas in Malawi is roughly log-normal, with the median farm spanning 1.22 acres.[14] Compare this to the area of a standard soccer pitch, 1.76 acres, or to the median US farm size, 45 acres (MacDonald, Korb, and Hoppe 2013).

Farms are an Important Source of Food for Their Owners. 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. 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.[15] 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.

12. I use estimates for rain-fed, low input usage agriculture.
13. E.g. for maize, GAEZ provides yield estimates for 1374 grid cells in Malawi.
14. See Appendix Figure D.1.
15. See Appendix Figure D.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.

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 employment and entrepreneurship. The farm's share is then $\frac{\text{farm output value}}{\text{farm output value + non-farm income}}$. The median of this measure 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.[16]

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?

16. The measure serves as an illustration but its interpretability beyond this role is limited since farm output value is not income from farming net of costs. A more fair measure of farm share would compare the income from farm sales to non-farm income, but this approach would be uninformative for most households since they tend to sell a tiny portion of their agricultural output.

I measure farm production diversity using the inverse of the Simpson index, which, in this setting, is the sum of squared product shares within a farm's output (measured at market value).[17] 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[18]), thus roughly corresponding to the scale of a 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. This suggests that not only do villages fail to specialize in one comparative advantage good, but even households within a single village are unable to specialize perfectly.[19]

---

17. For household h, $\text{Production Diversity}_h = \left(\sum_{i=1}^{n} \left(\frac{\text{output}_{h,i}}{\sum_{j=1}^{n} \text{output}_{h,j}} \right)^2 \right)^{-1}$, where $\text{output}_{h,i}$ is the market value of product i produced by h's farm. The Simpson index is the same as the Herfindahl index, with the former name being more common in ecological and agricultural settings. One simpler measure of diversity would be a count of different products that the farm produces. I use the diversity index as the primary measure 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.

18. 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.

19. Furthermore, relative diversity between farms and EAs being 0.54 suggests that Malawian

Coupled with the finding that most farmers sell very little of their output—even to their neighbors—this result suggests that the extent to which farmers can specialize in one good and trade with other farmers for other goods is limited, even within the same community. This leads to imperfect specialization not only across villages—as explored by Sotelo (2020)—but also across households within a single community, leaving room for household-level nutritional concerns to play a role in shaping farm production decisions.

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 ultimate 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.[20] Secondly, the model needs to feature heterogeneous agricultural products that farmers can choose from, so that it can make predictions on farm product choice and the tradeoff between specialization and diversification. 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

farmers tend to be in between the two extremes of producing the same set of products as their neighbors in the same village (relative diversity of 1) and producing unique sets of products within the same village (relative diversity close to 0).

20. See Kebede (2022) for an empirical exploration of (non-)separability of production and consumption in the setting of Ethiopian smallholder farmers.

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 physical 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 increase 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 a 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.[21] 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 household-specific trading cost $\tau_h > 0$. Define $d_h = 1 + \tau_h$ 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:

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

s.t.

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

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

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

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

21. Assuming a fixed land endowment per household is a reasonable assumption in this environment: land markets are practically non-existent in Malawi (Chen, Restuccia, and Santaeulàlia-Llopis 2023).

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 D.5 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.[22] It offers a natural way to embed dietary energy needs into the problem of a household. Theoretically, it makes the predictions of the model depart from those of standard CES preferences: some notable departures will be discussed in Section 4.4. Quantitatively, the caloric deviation penalty will be key in reproducing the empirical patterns in farm consumption and production behavior discussed in Section 5 below. Its advantages over several alternatives will also be discussed in Section 5.1.

22. The idea of introducing calorie considerations into utility has also been explored in the Giffen good literature, most notably as a calorie subsistence constraint in Gilley and Karels (1991). The caloric deviation penalty is more flexible and conducive to quantitative modeling: it is not undefined for households who cannot afford to satisfy the caloric requirement, and it does not collapse to a calorie-less utility function for households that can do so easily, which would lead to counterfactual predictions on rich households' food consumption as discussed in Section 5.

4.2 General Equilibrium & Calibration

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 manufactured 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 crops common in the data to be represented in the model:[23] 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

23. I restrict the set of goods represented in the model to a selection of *crops* because the model only considers the land input and is thus best suited to modeling production of crops rather than animal products. Choosing *six* crops allows for significant variation in consumed and produced diversity across households without making it prohibitively costly to solve for general equilibrium prices that clear all crop markets simultaneously.

households from the market, accounting for losses induced by the trade cost:

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

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.[24] 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 $\bar{p}_{\text{tobacco}}$, so its domestic market need not clear. Imposing that all other agricultural good markets clear, an application of Walras's Law implies

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

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's consumption of good i can be transformed and approximated (see Appendix A for derivation and details of the estimation) to yield the following regression specification:

$$\log c_{h,i} = \text{constant} + \delta^1_{h,source} + \underbrace{\sigma \log \varphi_i}_{\text{good FE}} + \sigma X_{1,h,i} - X_{2,h,i} \times \delta^2_{h,source} + \varepsilon_{h,i} \qquad (6)$$

For goods i that are *produced* by household h, two fixed effects $\delta^1_{h,source} = \delta^1_{h,produced}$ and $\delta^2_{h,source} = \delta^2_{h,produced}$ are shared by all such household-good combinations, $X_{1,h,i} = \log z_{h,i}$, and $X_{2,h,i} = k_i z_{h,i}$. For goods i that are *purchased* by household h, two fixed effects $\delta^1_{h,source} = \delta^1_{h,purchased}$ and $\delta^2_{h,source} = \delta^2_{h,purchased}$ are shared by all

24. UN Comtrade Database.

such household-good combinations, $X_{1,h,i} = \log \frac{1}{p_i d_h}$, and $X_{2,h,i} = \frac{k_i}{p_i d_h}$.

Thus the model produces an expression that can be estimated as a regression, providing estimates of the elasticity of substitution between foods σ (the coefficient on $X_{1,h,i}$) and taste weights $\{\varphi_i\}_i$ (which can be extracted from the good fixed effects 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 4.4, 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 (2024) 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. In Appendix Section C.3, I conduct an alternative calibration of the model with $\sigma = 1.3$ following Kebede (2024) and find that the key predictions of the model on farm behavior and the aggregate cost of subsistence remain similar to the baseline calibration with $\sigma = 0.75$.

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 four dimensions: land endowment L_h, non-farm income endowment wN_h, trade cost d_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. Because L_h is the only source of heterogeneity in the overall production capacity across farms in the model, this endowment does not map to measured land area in the data. Therefore, the parameters of the land endowment 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.[25] 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), which is more relevant for capturing the variation in production capacity across farms.[26]

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,[27] 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.

25. 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.

26. The calibrated L_h distribution should therefore be thought of as "effective" acres, capturing the differences in farm-level production capacity that could be due to differences in a multitude of factors: land area, farming ability, household size, etc.

27. 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 calibrated 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_h is what forces farms into subsistence: the higher the d_h, the more attractive it becomes to produce goods for own consumption, rather than sell farm output and buy food on the market. I assume that $d_h - 1$ is log-normally distributed: $d_h - 1 \sim \text{Lognormal}(\log(\bar{d} - 1), \sigma_d^2)$, making $\bar{d}$ the median of d_h. The parameters of this log-normal distribution are calibrated to match the mean and variance of the share of farm output sold. The calibrated values are $\bar{d} = 1.75$, $\sigma_d^2 = 0.01$.[28]

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. 1000 household types take independent draws from the N_h, d_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 80,000 households. Due to the structure of the household's problem, each household needs to be solved separately. The algorithm for solving the problem of a single household is described in Appendix E.

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

28. The median trade cost $\bar{d} = 1.75$ produced by this distribution implies a $1.75^2 = 3.1$x gap between the farm-gate price of the good's producer and the price paid by the final consumer. Although it appears large, it is in line with the difference in the distributions of farm-gate and consumption prices in Malawi shown in Fig. 6 of De Magalhães and Santaeulàlia-Llopis (2018).

TABLE 2: Model Calibration

parameter	value	moment/source	**data** moment	**model** moment
Distributions				
$\mathbb{E}(\log L_h)$	-14.9	avg $K_{in,h}/K_{req,h}$	1.036	0.904
$V(\log L_h)$	1.68	$V(\log \text{output}_h)$	1.528	1.546
$P(N_h = 0)$	0.112	$P(\text{non-farm income}_h = 0)$	0.112	0.113
$\mathbb{E}(\log N_h \mid N_h > 0)$	1	normalization	—	—
$V(\log N_h \mid N_h > 0)$	2.103	$V(\log \text{non-farm income}_h)$	2.103	1.940
$\bar{d}$	1.75	avg share sold	0.159	0.199
σ_d^2	0.01	variance of share sold	0.061	0.091
Parameters				
σ	0.75	estim. of Equation 6	—	—
γ	1	—	—	—
ψ	0.5	output elasticity of K_{in}	0.091	0.109
Good characteristics				
$\{\varphi_i\}_i$	...	estim. of Equation 6	—	—
φ_m	0.36	$\frac{\text{aggr. non-farm income}}{\text{aggr. farm output}}$	1.539	1.554
$\{k_i\}_i$	...	Food Comp. Tables	—	—
$\{z_i\}_i$	...	GAEZ attainable yields	—	—
z_m	1	normalization	—	—
$\bar{p}_{\text{tobacco}}/p_{\text{maize}}$	5.25	aggr. tobacco output share	0.091	0.092

4.3 AUXILIARY MODELS

To help understand the role of the nutritional mechanism, at times I will refer to two simpler versions of the model where the nutritional channel is absent.

The first auxiliary model is "CES-only": it sets $\psi = 0$, removing the caloric deviation penalty from the household's utility function, keeping just the CES term. The second auxiliary model is "Stone-Geary": it likewise drops the caloric deviation penalty, but also replaces the homothetic CES term with a non-homothetic Stone-Geary utility function, common in the structural transformation literature.

The definition and calibration of both models is discussed in Appendix C. The remainder of the paper will focus on using the baseline model with the nutrition

channel driven by the caloric deviation penalty: the two auxiliary models will only be employed on occasion for comparison and contrast.

4.4 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 later 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 by the household but have different calorie densities, the model makes a straightforward prediction on their relative consumption. If the household is failing to satisfy its caloric requirement, it will consume more of the relatively calorie-dense good at the cost of the calorie-empty one:

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, it will consume more of the calorically denser good:

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

2. if the household is overeating calories, it will consume more of the calorically emptier good:

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

a. See Appendix A for the proof of this proposition and 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 $z_{h,i} p_j = \lambda_h / (\mu_h d_h)$, where λ_h and μ_h 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. When the household fails to satisfy its caloric requirement, the caloric deviation penalty pushes the household's chosen consumption bundle away from the CES allocation toward the cheapest sources of calories:

Proposition 2 (Calories skew the CES solution)
Suppose there are two goods i and j that are both produced by household h (i.e. $x_{h,i} > 0$, $x_{h,j} > 0$), 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 the CES-only auxiliary model defined in Section 4.3 (i.e. with $\psi = 0$: 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 CES-only solution toward the calorically productive good:
$$\frac{c_{h,i}}{c_{h,j}} > \left(\frac{\varphi_i z_{h,i}}{\varphi_j z_{h,j}} \right)^{\sigma} = \frac{c^{CES}_{h,i}}{c^{CES}_{h,j}}$$

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

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 and

shows that the poorest households *only* care about maximizing their caloric intake, which is achieved by perfect specialization of their consumption and production:

Proposition 3 (Poorest households maximize calories, specialize)
a)

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 calorie-maximizing 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).

b)

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 (assuming that the maximizers and the minimizer in the $\bar{d}_h$ condition below, as well as their optima, are unique).

Which good(s) are produced and which are consumed in this limit depends on the properties of the goods and the trade cost d_h. 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 $d_h < \bar{d}_h$, then the household produces only the most revenue-productive $\arg\max_i p_i z_{h,i}$ good and consumes only the cheapest per calorie $\arg\min_i p_i/k_i$ good.

If $d_h > \bar{d}_h$, then the household produces only the most calorie-productive $\arg\max_i k_i z_{h,i}$ 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 $d_h = \bar{d}_h$, 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 $K_{req,h}$ 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 $\arg\max_i k_i z_{h,i}$) 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_h, 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. How does food consumption differ across farms of different sizes in the calibrated 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.[29] 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 and energy consumption is quite flat in the model: the output elasticity of kcal intake is just 0.109.

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.[30] 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

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

30. This finding is robust to measuring farm size with area instead of output value: see column (1) of Appendix Table D.2.

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

	log kcal intake			food diversity		
	(1) **model**: CES-only	(2) **model**: baseline	(3) **data**	(4) **model**: CES-only	(5) **model**: baseline	(6) **data**
log output	0.810	0.109	0.091***	−0.058	0.445	0.395***
	(0.001)	(0.001)	(0.005)	(0.001)	(0.001)	(0.034)
log non-farm income	0.216	0.089	0.063***	0.012	0.425	0.857***
	(0.001)	(0.001)	(0.004)	(0.001)	(0.002)	(0.033)
N	71,040	70,750	8,674	71,040	70,750	8,675
Adj. R^2	0.951	0.395	0.063	0.036	0.758	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.

more than tenfold.[31] 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 calibration 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) of Table 3 shows that, just like the model predicts, Malawian households running bigger farms or earning more income do eat more diverse diets[32]. Another way to measure dietary diversity is to simply count the number of distinct foods consumed by the household: this measure is explored in Ap-

31. The fact that even the richest households do not consume far more calories than their estimated requirement is the reason the caloric deviation penalty was defined to be symmetric around $K_{req,h}$: if the function was instead monotonically decreasing in $K_{in,h}$ (effectively only penalizing undereating but not overeating), it would lead the model to predict that the richest households' utility is effectively homothetic and their caloric intake grows proportionally with their income, which is evidently counterfactual.

32. The facts that the income elasticity of caloric intake is far below one and that consumed food variety increases with income are well documented. See, for example, Behrman and Deolalikar (1989), Colen et al. (2018), Zhou and Yu (2015), and Li (2021). To facilitate the comparison to existing calorie-income elasticity estimates, I estimate the shadow income elasticity of kcal intake in column (2) of Appendix Table D.2. I proxy shadow income with the sum of output value and non-farm income. The estimated shadow income elasticity of caloric intake is 0.121. Using the distribution of existing calorie-income elasticities reported in the meta-analysis by Zhou and Yu (2015), 0.121 is significantly below the literature average of 0.35 but within one standard deviation of it.

pendix B and behaves similarly.

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 B.

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 "CES-only" model, intake would grow one-for-one with shadow farm income, resulting in large output and income elasticities of kcal intake: 0.810 and 0.216 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.

Can the behavior induced by the caloric mechanism be replicated with a simpler formulation of food preferences? Appendix Table C.2 shows the result of running the same regressions in a model with Stone-Geary preferences with a subsistence constraint in the composite agricultural good, defined in Appendix Equation C.2. Just as in the baseline caloric needs model, poor households in the Stone-Geary model devote a disproportionate share of their resources to obtaining food. This generates low output and income elasticities of kcal intake (0.233 and 0.203) compared to the CES-only model. However, these elasticities are still much higher than the empirical ones (0.091 and 0.063): despite targeting those elasticities with the $\bar{c}$ moment, the model with Stone-Geary preferences simply cannot get nearly as close to the observed elasticities as the baseline model with the caloric channel can. Moreover, Stone-Geary preferences with a subsistence level in food consumption only create income dependence in the allocation between food and the manufactured good but not in the allocation between different foods. Therefore, the Stone-Geary model counterfactually predicts close to

no relationship between farm size and dietary diversity (compare columns (4) and (6) of Appendix Table C.2), performing no better than the CES-only model. Compared to the baseline model, the auxiliary model with Stone-Geary preferences still has crop choice, costly domestic trade, and non-homothetic preferences: it lacks only the caloric channel. As a result, although the simpler Stone-Geary preferences get closer to the baseline model in reproducing the observed calorie consumption behavior than the CES-only model does, they still fall short of the empirical moment, and fail in reproducing the food diversity behavior even qualitatively.

Two generalizations of the conventional Stone-Geary and CES preferences exist which I do not explore in this paper (Matsuyama (2023) reviews and classifies these and other non-homothetic generalizations of CES). One is to allow for good-specific subsistence requirements in Stone-Geary preferences. The other is to allow for good-specific income elasticities in CES preferences. Both of these, if appropriately calibrated, could potentiallly replicate the scale-dependent behavior of the baseline model with a caloric deviation penalty. They would do so, however, by introducing multiple additional parameters (one per simulated agricultural good). This would place greater demands on the data and the calibration procedure.[33] It would also reduce the external validity of the model: applying it to a different country with a different typical diet would require estimating a completely new set of subsistence requirements or income elasticities. In contrast, the caloric deviation penalty is governed by a single parameter ψ that is enough to generate all of the scale-dependent behavior discussed in this section. It makes the model equally applicable to countries other than Malawi: if placed in a region where staples other than maize are most productive, poor households in the model would endogenously switch to growing those staples as the most efficient source of calories. Finally, the caloric deviation penalty offers an intuitive

33. Stone-Geary preferences with good-specific subsistence requirements have two additional disadvantages not shared by non-homothetic CES with good-specific income elasticities. First, the domain of Stone-Geary preferences is restricted to consumption bundles that satisfy all subsistence requirements. This complicates quantitative analysis with heterogeneous farmers: distributions must be simulated in a way that ensures that all farmers can afford to satisfy the calibrated requirements in all goods at once, without the ability to substitute away from unproductive goods offered by the caloric deviation penalty. Second, Stone-Geary preferences are asymptotically homothetic: this leads to the counterfactual prediction that, among rich households, increases in farm size or income lead to proportional increases in caloric intake and no change in dietary diversity.

and simple mechanism endogenously explaining *why* households with different resources consume different diets and prefer different goods, without requiring such heterogeneous preferences to be specified/estimated exogenously.

Explicitly modeling dietary energy needs allows the baseline 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 do not 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.[34] 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

34. 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.

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 same for sellers) accounts for 67% of output.[35]

5.3 LARGER FARMS SELL MORE

Model. As long as trade costs d_h are 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_h, 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 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

35. 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 reflects perfect specialization) when only their sales are considered.

very little because the need to obtain calories eclipses all else for them, and the d_h drawn from the calibrated distribution 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 dynamics and risk. Having to allocate land between crops in the planting period before stochastic yields and prices are realized in the harvest period would make both small and large farmers prefer a less one-sided split between staple and cash crops. Small farmers would specialize less in the calorie-maximizing crop, choosing to devote some land to a marketable crop in case their staple crop harvest fails and they must purchase calories at the market to avoid suffering from an extreme caloric deviation penalty. Larger farmers would specialize less in the cash crop, choosing to devote some land to subsistence crops in case the cash crop yield (or market price) drops and makes it difficult to afford their caloric requirement with revenue alone.[36]

Still, the overprediction of the size-selling relationship by the present 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.

36. See Ashraf, Giné, and Karlan (2009) for an empirical exploration of one example of this mechanism: specialization in cash crops exposes farmers to additional risk that leads some to revert to growing staples.

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.

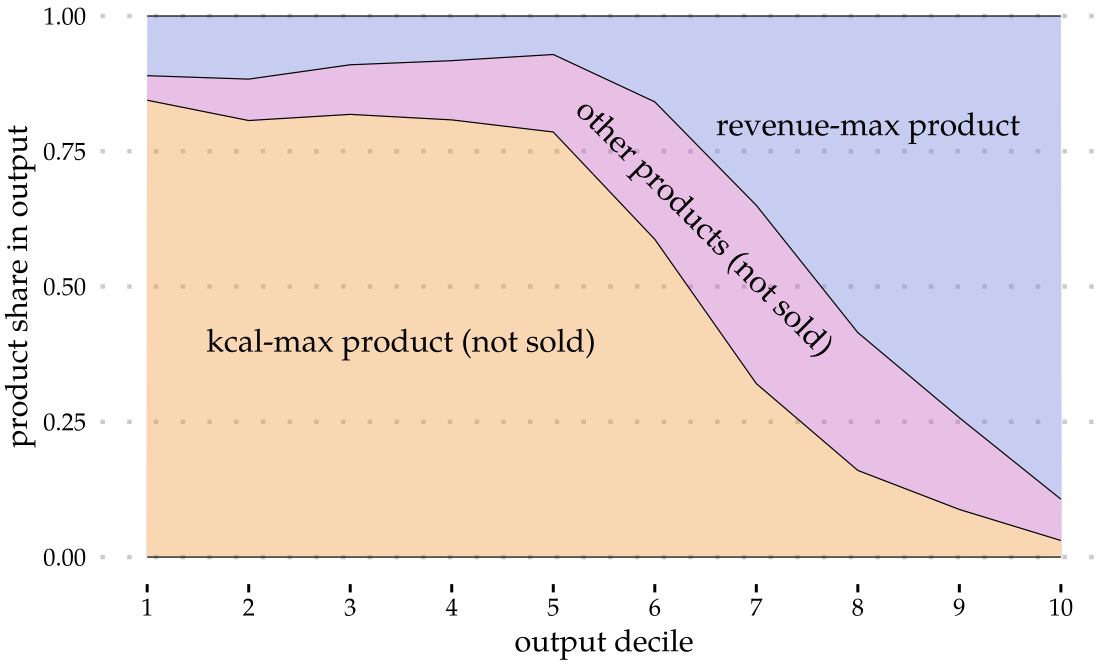
The smallest farmers cannot afford to reach their caloric requirement at all. Facing a high caloric deviation penalty, they do the best they can to get as close to their caloric requirement as possible, which is approximately following the strategy described in Proposition 3.[37] As Proposition 3 suggested should be the case for small farms facing sufficiently costly trade (most simulated farms indeed face $d_h > \bar{d}_h$), 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.

Households operating larger farms (those that can produce more output relative to their caloric requirement) are able to meet their caloric needs even without perfectly specializing their diet in the cheapest source of calories, allowing them to begin diversifying their diet.[38] 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-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

37. The allocation solving the problem in Proposition 3 gives the highest possible caloric intake, which comes out to only 64% of the caloric requirement for the median farmer in the smallest output decile and 99% in the second-smallest decile.

38. E.g. the median farmer in the sixth output decile can afford 182% of their caloric requirement by following the calorie-maximizing strategy from Proposition 3, implying that they no longer have to follow that strategy. Thus, farmers above roughly median size can easily satisfy their caloric needs while including more diverse foods and the manufactured good into their consumption bundle.



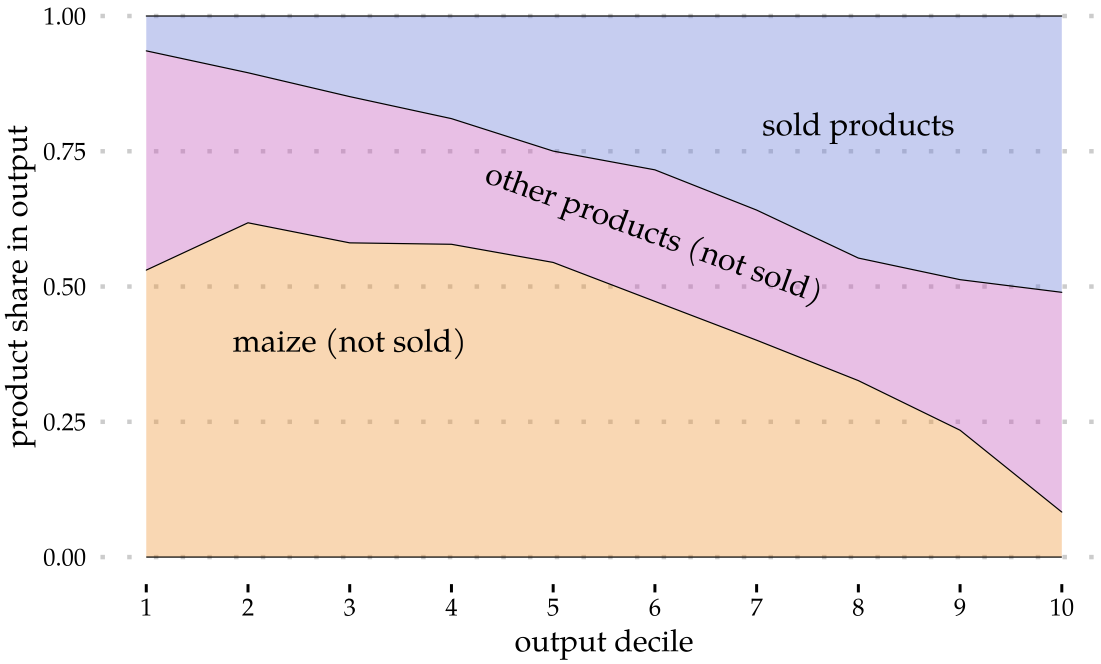
NOTE. Products are split into groups at farm level. "kcal-max product (not sold)" is the $\arg\max_i k_i z_{h,i}$ good, unless it's the same as the $\arg\max_i p_i z_{h,i}$ and is sold. "revenue-max product" is the $\arg\max_i p_i z_{h,i}$ good, unless it's the same as the $\arg\max_i k_i z_{h,i}$ and 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

in farm size, dominating farm output for households at the top of the farm size distribution, for whom satisfying their caloric needs is trivial.[39]

Data. Figure 2 is the conceptually similar data analogue of Figure 1. It shows the relationship between farm size and output shares of three types of products: maize (as long as it's not sold by the farm), other unsold products, and products that are at least partly sold. 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 predicts, 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

39. Appendix Figure D.6 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

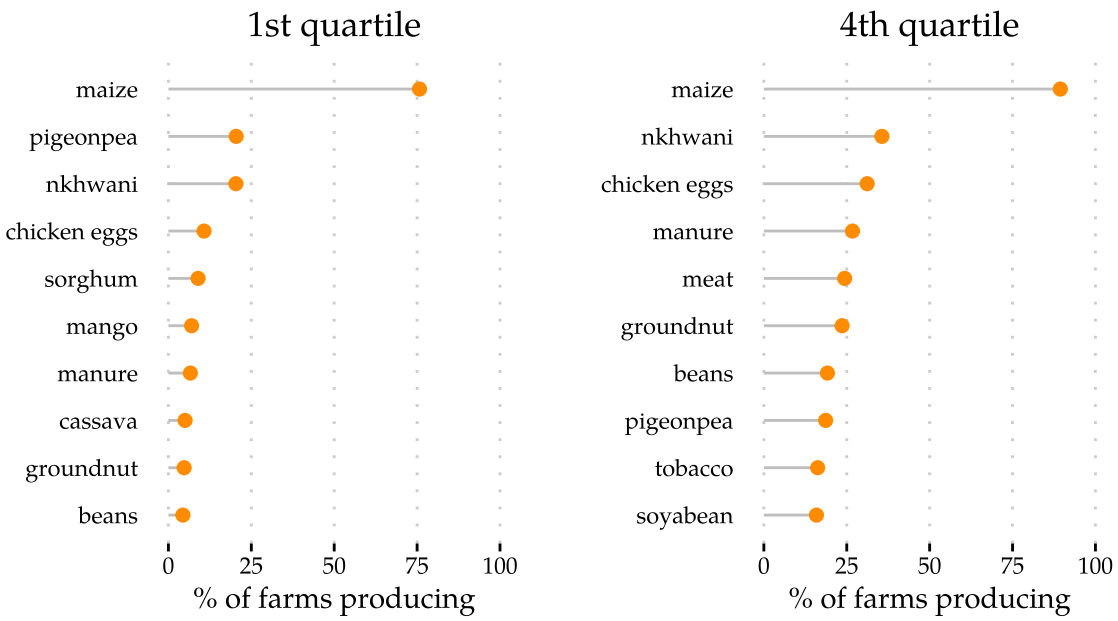
are at least partially sold.[40]

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 and price volatility, which would raise the incentives to diversify production for farmers across the entire size distribution.[41]

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

40. Appendix Figure D.7 shows a similar pattern, but using average product shares across farms within a decile, rather than product shares within decile-level output.

41. Another mechanism absent from the model is within-farm land heterogeneity akin to that modeled within regions by Sotelo (2020) and within villages by Kebede (2024). If productivities are drawn at plot-crop level rather than farm-crop level, larger farmers—who operate more plots—are likely to have heterogeneous comparative advantage crops at different points of the farm. This channel would provide another reason for higher production diversity observed at larger farms. However, it would fail to explain the changing *selection* of products with farm size displayed in Figure 2—or any of the results on farmers' food consumption behavior.



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

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, 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.[42]

42. One potential reason why larger farms may have more diverse production is to restore soil depleted by maize, which is demanding in soil nutrients. While the incentive to rotate crops is the same for everyone, smaller farmers may be constrained in their ability to sacrifice short-term output (and caloric intake) for long-term yields. Thoroughly exploring this channel is impossible with a single cross-section of the survey used in this paper. Still, Appendix Figure D.4 presents some limited evidence that larger farms have more diverse production even when maize-growing farms are excluded from the sample. The figure shows the 10 most common product combinations (unlike Figure 3, which looked at individual products) by farm size quartile, excluding farms that

The data on households' consumption of various goods adds nuance to this picture. Figure D.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 Leave Subsistence, Specialize

Model. The complex scale-dependent farm product choices described above only arise in the model when domestic trade costs are sufficiently high to make (semi-)subsistence attractive to farmers, which is the case for almost all farmers in the calibrated model. If the trade costs were much lower, farmers' production behavior would align perfectly with their individual comparative advantages and become decoupled from their nutritional situations:

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 $d_h < \sqrt{\frac{p_j z_{h,j}}{p_i z_{h,i}}}$ 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 this cutoff is satisfied for all goods $\left(d_h < \tilde{d_h} \equiv \sqrt{\frac{p_j z_{h,j}}{\max_{i \neq j} p_i z_{h,i}}}\right)$, 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 $d_h \geq \sqrt{\frac{p_j z_{h,j}}{p_i z_{h,i}}}$ (flipping the inequality), whether i is produced or purchased cannot be determined analytically and becomes a numerical issue.

If d_h is low enough (below $\tilde{d_h}$), the household behaves in a canonical Ricardian

grew any maize (unlike Appendix Figure D.3, which includes all farms). Larger non-maize farms have more unique combination choices (reflected by the smaller shares of all top-10 combinations) and are more likely to grow multiple products at once (reflected by the presence of multi-product combinations).

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 household.

The higher the trade cost d_h, the greater is the subsistence that the household exhibits. One way to operationalize subsistence is to measure the household-level correlation between the fraction of each product in the household's output and its fraction in the household's consumption. A low correlation indicates that a household's choice of which products to grow is not driven by its preferences for consuming those products, and that the choice of which products to consume is not driven by which products can be cheaply grown.[43] Column (1) of Table regresses the correlation between each product's share in the market values of the household's consumption and production on farm size, income, and trade cost in the model. As predicted, simulated households facing a higher trade cost exhibit greater subsistence (greater correlation between consumption and production shares). Reiterating the previous results, households operating bigger farms or having more non-farm income exhibit less subsistence. Column (3) repeats this exercise while measuring the consumption share of each product using its caloric content rather than its market value: results are similar.

Data. Repeating this exercise in the data requires establishing a proxy of trade costs. I construct two alternative proxies.

The first proxy is based on the market access measure developed in Donaldson and Hornbeck (2016):

$$\text{trade cost}_f = -\log\left(\sum_c \tau_{f,c}^{-\theta} N_c\right) \tag{7}$$

where $\tau_{f,c}$ is the distance from farm f to city c, and N_c is the population of city c. I use all major cities in Malawi. θ is the trade elasticity: I borrow the value of 1.5, which was used by Allen and Atkin (2022) in a similar context of internal trade in a developing country.

The first measure requires calculating the distance between each farm and each city: these rely on knowing the location of each farm. While the IHS survey provides the coordinates of each farm, these have random noise added to them to

43. This measure is inspired by the large literature on the separability hypothesis, primarily by Kebede (2022).

TABLE 5: Trade cost and the correlation between product shares in consumption and production: model and data

	product share correlation			
	cons. value, prod. value		**cons. kcal, prod. value**	
	(1) **model**	(2) **data**	(3) **model**	(4) **data**
log output	−0.081	−0.020***	−0.082	−0.011***
	(0.001)	(0.003)	(0.001)	(0.003)
log non-farm income	−0.057	−0.046***	−0.075	−0.027***
	(0.001)	(0.002)	(0.001)	(0.003)
trade cost	0.490	0.025***	0.514	0.007*
	(0.032)	(0.003)	(0.035)	(0.004)
N	70,750	8,100	70,750	8,099
Adj. R^2	0.108	0.072	0.112	0.018

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

NOTE. Kcal intake, output, and non-farm income are relative to the caloric requirement of the household. Product share correlation is calculated across products within each household. "cons. value, cons. kcal, prod. value" indicate how each share was measured: in market value or in kcal intake. "trade cost" denotes $\log d_h$ in the model and an inverse of the market access index in the data.

preserve anonymity. However, the survey provides the exact distances between each farm and various infrastructure objects. I use these to construct the second proxy: the principal component of the distances between each farm and the nearest agricultural market, population center, and road. This measure condenses the information contained in these three distances into a single dimension.

Columns (2) and (4) of Table 5 repeat the correlation exercise in the data, using the first (market access) trade cost proxy. Appendix Table D.3 uses the second (PCA) trade cost proxy. Both proxies match the predictions of the model: households facing a higher trade cost exhibit a higher correlation between product shares in their consumption and production. Note that while the signs of the coefficients are indicative of the mechanism, the magnitudes of the coefficients are not directly comparable between the model and the data, as well as between the two trade cost proxies.

Another 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 6, 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 higher trade costs should diversify their production. Column (2) of Table 6 confirms this in the data, using the trade cost proxy based on the market access index.

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 predictions on the consumption behavior of households and on their farm prod-

TABLE 6: Commercialization, trade cost, and production diversity

	production diversity	
	(1)	(2)
sold output share	−0.044***	
	(0.016)	
trade cost		0.059***
		(0.008)
log(output value)	0.102***	0.180***
	(0.013)	(0.006)
log(non-farm income)	−0.051***	−0.041***
	(0.009)	(0.006)
N	4,042	8,674
Adj. R^2	0.025	0.097

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

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). The trade cost is measured with the inverse of the market access proxy.

uct choice, many of which are borne out by the data.

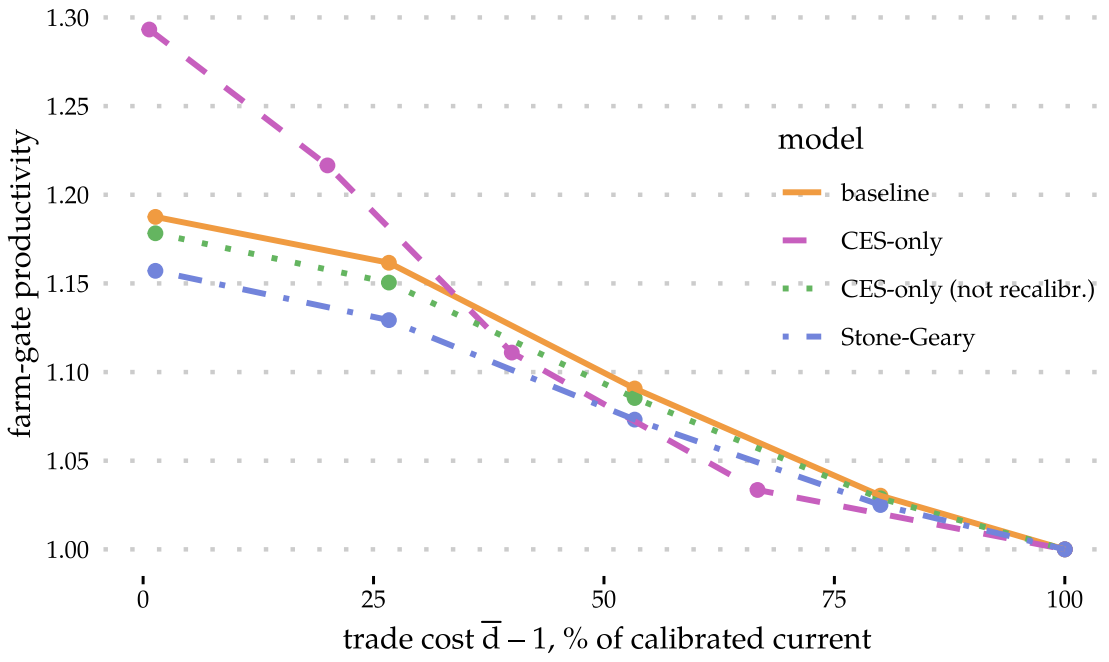
Subsistence farmers produce not what they are best at relative to others, but what they want to see on their family's table. If the two do not coincide, the agricultural output of the economy will not be maximal. The model defined in Section 4 and explored and tested in Section 5 provides a framework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

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

Each exercise is implemented by replacing the calibrated value of the median trade cost parameter $\bar{d}$ with (lower) counterfactual values while keeping all other parameters and distributions the same. The problems of all 80,000 model households are then re-solved with the new trade cost level and a given set of prices, allowing households to adjust their product choice and consumption bundles to take advantage of easier trade. The process is repeated to find the new set of

market-clearing prices to ensure that the counterfactual solution is in a general equilibrium.

6.1 Falling Trade Costs Raise Farm Productivity



Note. Farm-gate productivity is the sum of farm outputs valued at market prices, divided by the sum of farm land endowments. Aggregate farm-gate output is chain-weighted between each pair of consecutive $\bar{d}$ levels to obtain real values. 100% of trade cost represents the median trade cost $\bar{d}$ calibrated for each model, 0% of trade cost represents a complete counterfactual removal of trade costs. Productivity at the calibrated current median trade cost $\bar{d}$ is normalized to 1.

Figure 4: Farm productivity and trade cost

Figure 4 shows the effect of gradual trade cost reductions from the calibrated median trade cost $\bar{d}$ (1.75 in the baseline model, mapped to 100% on the plot) on real farm output productivity: total agricultural output produced by all farmers in the economy relative to the land used.[44] The output is measured at "farm-gate": it does not account for any direct losses from $\bar{d}$ that sold goods are subject to.

Going to an allocation with costless trade in agricultural goods ($\bar{d} = 1$, mapped to 0% on the plot) would raise the productivity of farmers by 18.4% in the baseline

44. Prices are used for weighting different goods, but changes between consecutive levels of $\bar{d}$ are deflated by chain-weighting.

model. All of this improvement is driven by the increasing alignment between the products individual farmers choose to grow and their comparative advantages.

6.2 Subsistence Plays Key Role in Depressing Farm Productivity, Caloric Needs Are Secondary

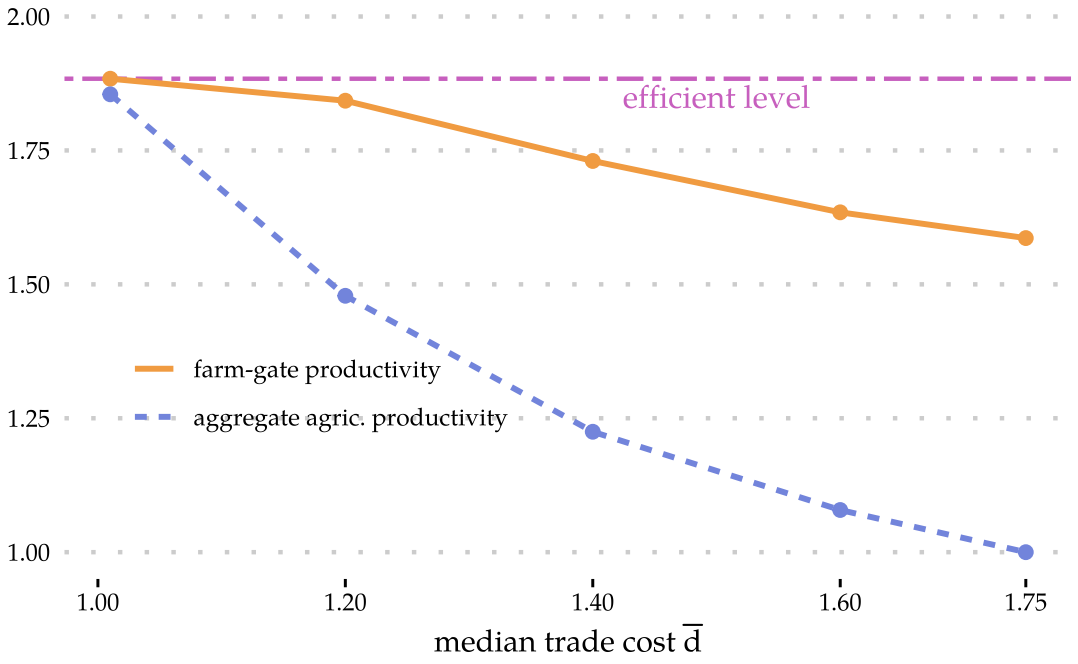
To separate the role of subsistence itself from its interaction with caloric needs in driving farm productivity, I repeat the trade cost reduction exercise in the two auxiliary models, CES-only and Stone-Geary. They simulate farm subsistence just like the baseline model, but lack the caloric mechanism. These are likewise shown in Figure 4.

For partial reductions in trade cost from the calibrated level, both the CES-only and the Stone-Geary model predict farm-gate productivity gains that are moderately smaller than those predicted by the baseline caloric needs model. Because the two alternative models do not push the poor households to prioritize the most calorie-productive goods above all else, these households have more room to align their production bundle with their productivity draws even when trade costs are high, reducing the drag of trade costs on their output—conversely, reducing the potential gain from alleviating those trade costs. For a complete trade liberalization, however, the CES-only model predicts a *higher* farm-gate productivity gain than the baseline model. This is driven by the higher median trade cost parameter $\bar{d} = 2.5$ required by the CES-only model to approximate the empirical share of output sold: if the CES-only model is made to use the parameter values calibrated for the baseline model (including $\bar{d} = 1.75$), the productivity gain it predicts is significantly lower (the "CES-only (not recalibr.)" series on the plot).

Farm productivity gains due to a counterfactual domestic trade cost reduction—or farm productivity losses due to the current level of domestic trade costs—are thus driven mainly by the fact of farm subsistence, not the caloric channel specifically. The two auxiliary models, CES-only and Stone-Geary, lack the caloric channel yet suggest that farm-level subsistence imposes a drag on farm-gate productivity of a similar magnitude to that suggested by the baseline caloric needs model. The nutritional channel does slightly amplify the productivity-dampening effect of a given trade cost magnitude, but this amplification loses relevance when the trade cost is recalibrated to reproduce empirical trade volumes. What matters

most for keeping the productivity of farmers depressed is that they have to grow what they want to consume—not the exact motive that determines which particular products they want (or need) to consume.

6.3 Mechanical Losses are Comparable to Product Choice Losses



Note. Farm-gate output (does not account for losses to trade cost) and aggregate agricultural GDP (accounts for losses to trade cost) of agricultural products are valued at market prices. Both series are chain-weighted between each pair of consecutive $\bar{d}$ levels to obtain real values.

Figure 5: Farm productivity and aggregate agricultural productivity

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 on the way from the farmer to the ultimate buyer when they are not the same household. The aggregate GDP of the entire agricultural sector captures both effects. Figure 5 compares 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 median trade cost $\bar{d} = 1.75$, aggregate agricultural

productivity in the baseline model is considerably more depressed than farm-gate productivity alone, relative to the efficient level. Likewise, the potential GDP gains from removing trade frictions entirely are significant: an 85% increase in agricultural GDP. Improved product choice is responsible for $\frac{1}{3}$ of this increase, with the remaining $\frac{2}{3}$ mechanically caused by smaller losses to the trade cost.

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 $\bar{d} = 1.75$). This intermediate stage is attained at $\bar{d} = 1.20$. The model predicts that lowering the trade costs to that level would raise aggregate agricultural productivity by 47%, with improved product choice causing over half—54%—of this increase. For even smaller reductions in the trade cost from the current level, the share of product choice in the productivity gain is even larger (see Appendix Figure D.8). Reductions in mechanical losses to trade costs play a smaller role in partial liberalizations simply because the volume of trade is not yet that large.

Finally, Appendix Figure D.9 compares the aggregate agricultural productivity response in the baseline model (already shown in Figure 5) to that in the two alternative models: CES-only and Stone-Geary. Similarly to farm-gate productivity, aggregate agricultural productivity gain from incomplete trade cost reductions is higher in the baseline caloric needs model.

6.4 Smallest Farmers are Most Affected

The product choice response of farms to a reduction in trade cost is heterogeneous among farms of different sizes in the baseline model. Figure 6 shows the heterogeneous responses to a trade cost reduction to an intermediate $\bar{d} = 1.2$ level, which raises the average output share sold from 16% to 50%.

The smallest farmers respond the most, with their yields increasing by 32% and their consumption value by 55%. 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. This results in a 33%

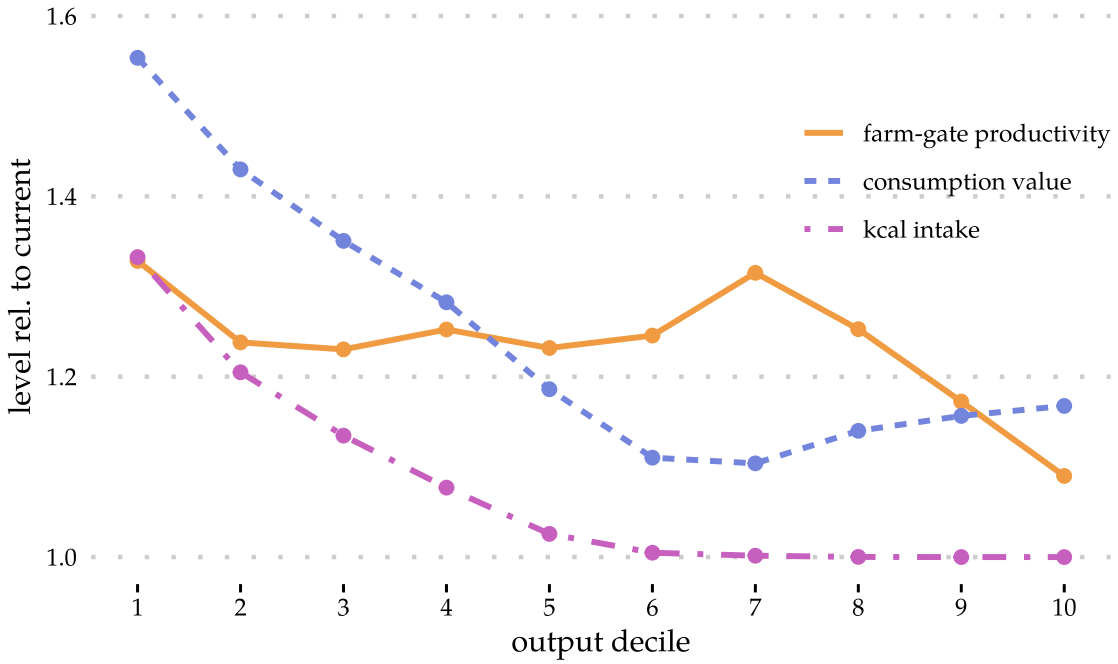
increase in calorie intakes within the smallest decile.

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 own 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.

Thus, the model suggests that improving market access benefits small nutritionally insecure farms even more than large commercialized ones: small farmers are the ones who most readily switch their specialization from staple crops to their comparative advantage crops if given a chance, and gain the most both in their consumption value and in their nutritional security. While the caloric needs channel was not the most important for the aggregate farm productivity results (Section 6.2), it is crucial for these heterogeneous effects. The logic outlined above is specific to the caloric needs model as it derives from the scale-dependent household behavior discussed in Section 5.

The heterogeneous response of farmers to trade cost reductions is also key for understanding the bias that is likely introduced into the model's aggregate predictions by the fact that it overpredicts the strength of the size-selling relationship (Section 5.3). This leads the product choice (and farm productivity) of large farmers to respond very little to trade cost reductions, since it is already close to perfect specialization in the revenue-maximizing crop. If the model was able to perfectly match the (more gradual) empirical relationship between farm size and commercialization, the product choice of large farmers would respond more strongly to trade cost reductions (but still less strongly than that of small or medium farmers), as large farmers would still have room to increase their specialization in the most revenue-productive goods. By understating the product choice response of large farmers due to the model's overprediction of the size-selling relationship, the model also understates the total farm productivity gains from trade cost reductions. At the same time, the overprediction of selling by large farms in their comparative advantage products likely overstates the mechanical



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

Figure 6: Response of farm-gate productivity and household consumption to a 1.75 → 1.2 reduction in median trade cost $\bar{d}$, by farm size decile

gains from trade cost reductions, since the current volume of trade by large farmers (who, again, contribute more to the aggregate) that the trade cost is applied to is overstated in the model. Thus, the model's overprediction of the size-selling behavior likely leads it to understate the farm productivity gains from trade cost reductions but overstate the mechanical gains from trade cost reductions. The net effect of these biases on the model's predicted aggregate productivity gains is ambiguous.

6.5 Subsistence Matters for Macro Behavior, Nutrition Matters for Micro Behavior and Heterogeneity

Domestic trade frictions push farmers into partial subsistence, requiring them to produce part of their consumption bundle on their own. This deviation between each farmer's comparative advantage and actual product choice depresses ag-

gregate agricultural productivity by leaving gains from specialization on the table (Sections 6.1, 6.3). As the comparison of alternative preferences formulations shows through the rough similarity of their aggregate counterfactual predictions, this aggregate result is driven by the fact of farm-level subsistence itself—not by the exact shape that farmers' food demand takes (Sections 6.2, 6.3).

The caloric needs specification introduced in this paper, the Stone-Geary preferences, and the CES-only preferences generate comparable aggregate predictions—at least in this application. Where the caloric needs model produces significantly different predictions is the heterogeneous behavior of individual farmers, making it a powerful tool for understanding their consumption, production, and selling behavior (Section 5) or predicting the heterogeneous effects of counterfactual policies (Section 6.4).

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 also determining the production decisions of these farmers.

I test the model's predictions 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 47% 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 food preferences. While the caloric channel in food preferences is crucial for the model's ability to reproduce and explain the heterogeneous behavior of individual farmers, the aggregate productivity cost of limited agricultural trade is driven by subsistence itself: the interaction of subsistence with nutrition plays a secondary role. What matters for keeping the aggregate agricultural productivity depressed is that farmers have to grow much of the food they want to eat—not how exactly their preferences for food are formed.

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. Such programs 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). 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. Because the model captures the trade-off between calories, dietary diversity, and commercial production, it would be able to say which farmers are likely to benefit from a given policy and which are likely to be harmed, covering crucially not only economic but also nutritional outcomes.

References

Allen, Treb, and David Atkin. 2022. "Volatility and the Gains From Trade." *Econometrica* 90 (5): 2053–2092.

Ashraf, Nava, Xavier Giné, and Dean Karlan. 2009. "Finding Missing Markets (and a Disturbing Epilogue): Evidence from an Export Crop Adoption and Marketing Intervention in Kenya." *American Journal of Agricultural Economics* 91 (4): 973–990.

Behrman, Jere R., and Anil Deolalikar. 1989. "Is Variety the Spice of Life? Implications for Calorie Intake." *The Review of Economics and Statistics* 71 (4): 666.

Benson, Todd. 2021. *Disentangling Food Security from Subsistence Agriculture in Malawi.* Washington, DC: International Food Policy Research Institute.

Blanco, Cesar, and Xavier Raurich. 2022. "Agricultural Composition and Labor Productivity." *Journal of Development Economics* 158:102934.

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

Chen, Chaoran, Diego Restuccia, and Raül Santaeulàlia-Llopis. 2023. "Land Misallocation and Productivity." *American Economic Journal: Macroeconomics* 15, no. 2 (1, 2023): 441–465.

Colen, L., P.C. Melo, Y. Abdul-Salam, D. Roberts, S. Mary, and S. Gomez Y Paloma. 2018. "Income Elasticities for Food, Calories and Nutrients across Africa: A Meta-Analysis." *Food Policy* 77:116–132.

De Magalhães, Leandro, and Raül Santaeulàlia-Llopis. 2018. "The Consumption, Income, and Wealth of the Poorest: An Empirical Analysis of Economic Inequality in Rural and Urban Sub-Saharan Africa for Macroeconomists." *Journal of Development Economics* 134:350–371.

Donaldson, Dave, and Richard Hornbeck. 2016. "Railroads and American Economic Growth: A "Market Access" Approach." *The Quarterly Journal of Economics* 131, no. 2 (1, 2016): 799–858.

Drewnowski, Adam. 2010. "The Nutrient Rich Foods Index Helps to Identify Healthy, Affordable Foods." *The American Journal of Clinical Nutrition* 91, no. 4 (1, 2010): 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* 10, no. 9 (1, 2018): 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* 139, no. 8 (1, 2009): 1549–1554.

Gilley, Otis W., and Gordon V. Karels. 1991. "In Search of Giffen Behavior." *Economic Inquiry* 29 (1): 182–189.

Gollin, Douglas, David Lagakos, and Michael E. Waugh. 2014. "Agricultural Productivity Differences across Countries." *American Economic Review* 104, no. 5 (1, 2014): 165–170.

Gollin, Douglas, and Richard Rogerson. 2014. "Productivity, Transport Costs and Subsistence Agriculture." *Journal of Development Economics* 107:38–48.

Haque, Zabin. 2022. *The Production Engel.* Working Paper.

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* 147 (1): 86–96.

Kebede, Hundanol A. 2022. "Market Integration and Separability of Production and Consumption Decisions in Farm Households." *Journal of Development Economics* 158:102939.

Kebede, Hundanol A. 2024. "Gains from Market Integration: Welfare Effects of New Rural Roads in Ethiopia." *Journal of Development Economics* 168:103252.

Levy, Sarah. 2005. *Starter Packs: A Strategy to Fight Hunger in Developing Countries?* Wallingford, UK; Cambridge, MA: CABI Pub.

Li, Nicholas. 2021. "An Engel Curve for Variety." *The Review of Economics and Statistics* 103 (1): 72–87.

———. 2023. "In-Kind Transfers, Marketization Costs and Household Specialization: Evidence from Indian Farmers." *Journal of Development Economics* 164.

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, August.

Matsuyama, Kiminori. 2023. "Non-CES Aggregators: A Guided Tour." *Annual Review of Economics* 15, no. 1 (13, 2023): 235–265.

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* 15.

Omamo, Steven Were. 1998a. "Farm-to-market Transaction Costs and Specialisation in Small-scale Agriculture: Explorations with a Non-separable Household Model." *Journal of Development Studies* 35 (2): 152–163.

———. 1998b. "Transport Costs and Smallholder Cropping Choices: An Application to Siaya District, Kenya." *American Journal of Agricultural Economics* 80 (1): 116–123.

Pingali, Prabhu, and Naveen Sunder. 2017. "Transitioning Toward Nutrition-Sensitive Food Systems in Developing Countries." *Annual Review of Resource Economics* 9, no. 1 (5, 2017): 439–459.

Pontzer, Herman, Ramon Durazo-Arvizu, Lara R. Dugas, Jacob Plange-Rhule, Pascal Bovet, Terrence E. Forrester, Estelle V. Lambert, Richard S. Cooper, Dale A. Schoeller, and Amy Luke. 2016. "Constrained Total Energy Expenditure and Metabolic Adaptation to Physical Activity in Adult Humans." *Current Biology* 26 (3): 410–417.

Rapsomanikis, George. 2015. *The Economic Lives of Smallholder Farmers.* Food and Agriculture Organization of the United Nations.

Rivera-Padilla, Alberto. 2020. "Crop Choice, Trade Costs, and Agricultural Productivity." *Journal of Development Economics* 146:19.

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* 112, no. 34 (25, 2015): 10657–10662.

Sotelo, Sebastian. 2020. "Domestic Trade Frictions and Agriculture." *Journal of Political Economy* 128 (7): 2690–2738.

USDA and HHS. 2020. *Dietary Guidelines for Americans, 2020-2025.* 9th ed. U.S. Department of Agriculture and U.S. Department of Health and Human Services, December.

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.* 1st Edition. Lilongwe, Malawi.

Zhou, De, and Xiaohua Yu. 2015. "Calorie Elasticities with Income Dynamics: Evidence from the Literature." *Applied Economic Perspectives and Policy* 37 (4): 575–601.

Appendix

A 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.

Notation.

Fix all parameters but L_h, N_h.

Define K_{max} to be the solution of the calorie-maximizing problem for given resources:

$$K_{max}(L_h, N_h) = \max_{\substack{\{c_{h,i}, x_{h,i}, x^p_{h,i}, \\ x^s_{h,i}\}_{i=1}^n, c_{h,m}}} \sum_{i=1}^{n} c_{h,i} k_i$$

s.t. constraints (2)-(5) with provided L_h, N_h.

Define K_{in} to be the optimal calorie intake in the solution of the original problem for given resources:

$$K_{in}(L_h, N_h) = \sum_{i=1}^{n} c_{h,i} k_i$$

in the solution of the full problem in (1)-(5), with provided L_h, N_h.

Define U to be the first (CES) term in the utility function of the original problem:

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

Define U_{max} to be the maximum attainable CES utility (disregarding the caloric deviation penalty) for given resources:

$$U_{max}(L_h, N_h) = \max_{\substack{\{c_{h,i}, x_{h,i}, x^p_{h,i}, \\ x^s_{h,i}\}_{i=1}^n, c_{h,m}}} U\left(\{c_{h,i}\}_i, c_{h,m}\right)$$

s.t. constraints (2)-(5) with provided L_h, N_h.

Define $\Delta f\ (a, b, L_h, N_h)$ to be the difference in caloric deviation penalty achieved by moving the caloric intake from fraction a of K_{max} to fraction b of K_{max}:

$$\Delta f\ (a, b, L_h, N_h) = f\left(bK_{max}(L_h, N_h), K_{req,h}\right) - f\left(aK_{max}(L_h, N_h), K_{req,h}\right)$$

Proof of part a).

Need to show that

$$\lim_{L_h,N_h\to 0} \frac{K_{in}(L_h,N_h)}{K_{max}(L_h,N_h)} = 1.$$

Note that $\frac{K_{in}(L_h,N_h)}{K_{max}(L_h,N_h)} \leq 1$ by definition. Let $\varepsilon > 0$. Define $a = \max\{1-\varepsilon, 0\}$. Need to show that $\exists \delta > 0$ s.t. if $L_h, N_h < \delta$, then $\frac{K_{in}(L_h,N_h)}{K_{max}(L_h,N_h)} > a$. Let $b = \frac{a+1}{2}$.

$$\Delta f(a,b,L_h,N_h) = \psi\left(\frac{K_{max}(L_h,N_h)}{K_{req,h}}\frac{1-a}{2} - \frac{K_{req,h}}{K_{max}(L_h,N_h)}\frac{1-a}{a(a+1)}\right)$$

Note that $K_{max}(L_h,N_h)$ is increasing in L_h, N_h, and hence so is Δf. Furthermore, $\lim_{L_h,N_h\to 0} \Delta f(a,b,L_h,N_h) = -\infty$. Hence, $\exists L_{h,1}, N_{h,1}$ small enough s.t. $\Delta f(a,b,L_h,N_h) < -1 \ \forall L_h < L_{h,1}, N_h < N_{h,1}$. The choice of negative constant -1 is arbitrary here. Furthermore, $\lim_{L_h,N_h\to 0} U_{max}(L_h,N_h) = 0$. Hence, $\exists L_{h,2}, N_{h,2}$ s.t. $U_{max}(L_h,N_h) < 1 \ \forall L_h < L_{h,2}, N_h < N_{h,2}$.

Let $L_{h,3}, N_{h,3}$ be s.t. $K_{max}(L_{h,3},N_{h,3}) < K_{req,h}$. Let $\tilde{L}_h = \min_i L_{h,i}$, $\tilde{N}_h = \min_i N_{h,i}$, $\delta = \min\{\tilde{L}_h, \tilde{N}_h\}$. Finally, let $L_h, N_h < \delta$. Will show that $\frac{K_{in}(L_h,N_h)}{K_{max}(L_h,N_h)} > a$.

Suppose not: $\frac{K_{in}(L_h,N_h)}{K_{max}(L_h,N_h)} = c \leq a$.

$$\Delta f(c,b,L_h,N_h) = \Delta f(c,a,L_h,N_h) + \Delta f(a,b,L_h,N_h)$$

It has already been shown that $\Delta f(a,b,L_h,N_h) < -1$. It can also be shown that $\Delta f(c,a,L_h,N_h) < 0$ as long as $a, c \in (0,1)$ and $K_{max}(L_h,N_h) < K_{req,h}$, which are satisfied in this case. Therefore, $\Delta f(c,b,L_h,N_h) < -1$.

$bK_{max}(L_h,N_h) < K_{max}(L_h,N_h)$, so deviating to an alternative allocation "+" yielding a caloric intake of $\sum_{i=1}^{n} c_{h,i}^{+} k_i = bK_{max}(L_h,N_h)$ is feasible.

Furthermore, deviating to allocation "+" would increase the 2nd term of the utility function by $-\Delta f(c,b,L_h,N_h) > 1$. It would simultaneously reduce the 1st term of the utility function by at most $U\left(\left\{c_{h,i}^{*}\right\}_i, c_{h,m}^{*}\right) < U_{max}(L_h,N_h) < 1$, where "*" denotes the original allocation.

Therefore, overall utility is increased in this alternative allocation "+". Therefore, the original allocation could not have been optimal, implying $\frac{K_{in}(L_h,N_h)}{K_{max}(L_h,N_h)} > a$.

Proof of part b).

Let $j = \arg\min_i p_i/k_i$.

Suppose good $l \neq j$ is purchased. Reduce $x^p_{h,l}$ by any $y \in \left(0, x^p_{h,l}\right]$, raise $x^p_{h,j}$ by $y\frac{p_l}{p_j}$, which still satisfies the budget constraint. The household loses yk_l calories but gains $y\frac{p_l}{p_j}k_j$ calories. By assumption, $k_j/p_j > k_l/p_l$, implying $y\frac{p_l}{p_j}k_j > yk_l$. Therefore, this deviation increases $\sum_i c_{h,i}k_i$, and purchasing l could not have been optimal. Since this holds for any y up to $x^p_{h,l}$, it must be that $x^p_{h,l} = 0 \; \forall \; l \neq j$ in the optimum, hence only j can be purchased.

Now let $q = \arg\max_i \left\{\max\left\{k_i z_{h,i}, \; \frac{p_i z_{h,i}}{d_h} \cdot \frac{k_j}{p_j d_h}\right\}\right\}$.

Suppose that quantity x of good $l \neq q$ is produced and consumed. Reduce x by $y \in (0, x]$, raise $x_{h,q}$, $c_{h,q}$ by $y\frac{z_{h,q}}{z_{h,l}}$. The household loses yk_l calories but gains $y\frac{z_{h,q}}{z_{h,l}}k_q$ calories. By the same argument as before, this deviation increases $\sum_i c_{h,i}k_i$, so producing and consuming x of l could not have been optimal.

Suppose that quantity x of good $l \neq q$ is produced and sold. Reduce x by $y \in (0, x]$, raise $x_{h,q}$, $x^s_{h,q}$ by $y\frac{z_{h,q}}{z_{h,l}}$. Because only j can be purchased, the household loses $y\frac{p_l}{p_j d_h^2}k_j$ calories but gains $y\frac{z_{h,q}}{z_{h,l}}\frac{p_q}{p_j d_h^2}k_j$ calories. By the same argument as before, this deviation increases $\sum_i c_{h,i}k_i$, so producing and selling x of l could not have been optimal.

As l can neither be produced and consumed nor produced and sold, it is not produced. Therefore, only q can be produced.

Therefore, only good q is produced, only good j (if any) is purchased, and either one or two of these are consumed.

Which good is q depends on d_h. If $d_h < \sqrt{\frac{\max_i p_i z_{h,i}}{\min_i p_i/k_i \cdot \max_i k_i z_{h,i}}}$, then $\max_i k_i z_{h,i} > \max_i \frac{p_i z_{h,i}}{d_h} \cdot \frac{k_j}{p_j d_h}$, and q is the maximizer of the former term. If the inequality is reversed, q is the maximizer of the latter. If the d_h condition is satisfied with equality, the household is indifferent between the two solutions.

Proof of Proposition 4.

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

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

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

$$\frac{\lambda_h}{\mu_h} = \frac{p_i z_{h,i}}{d_h} + \frac{\theta(x^s_{h,i}) z_{h,i}}{\mu_h} + \frac{\theta(x_{h,i}) z_{h,i}}{\mu_h}$$

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

Therefore, for any good i,

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

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,

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

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

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

This implies

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

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

Derivation of Equation 6.

The first order condition for the consumption of good i by household h 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

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

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

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

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

for HH-good combinations where the good is purchased.

The equations being estimated for produced and purchased goods respectively are then:

$$\log c_{h,i} = \text{constant} + \delta^1_{h,produced} + \delta_i + \sigma X_{1,h,i} - X_{2,h,i} \times \delta^2_{h,produced} + \varepsilon_{h,i} \quad \text{(A.3)}$$

and

$$\log c_{h,i} = \text{constant} + \delta^1_{h,purchased} + \delta_i + \sigma X_{1,h,i} - X_{2,h,i} \times \delta^2_{h,purchased} + \varepsilon_{h,i} \quad \text{(A.4)}$$

Here, $X_{1,h,i}$ is the log of reported physical land yield for produced goods and the negative log of reported purchased price for purchased goods, $X_{2,h,i}$ is the caloric land yield for produced goods and the inverse of purchased price per calorie for purchased goods, and δ terms are fixed effects.

Appropriately defining the variables for household-good combinations belonging to each of the two groups (produced and purchased), Equations A.3 and A.4 can be combined into a single Equation 6, and estimated as such.

B Alternative Dietary Variety Measures: Nutrient Richness and Food Count

The measure of dietary variety explored in Section 5 is a diversity index. One alternative measure is a simple count of distinct foods consumed by the household.[45] As Tables B.1 and B.2 show, this measure is highly correlated with food diversity and has a similar relationship with farm size and non-farm income: 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. Table B.2 also shows that dietary variety measures retain their positive association with farm size and income even when calorie intake is controlled for.

45. The left panel of Figure D.2 shows the most widely consumed foods.

TABLE B.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.

TABLE B.2: 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.

One physiological rather than hedonic reason to prefer a diverse diet is to ensure a sufficient intake of essential macro- and micronutrients. 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, Keast, and Drewnowski 2009; Drewnowski 2010), but applied since then to assessing the quality of the overall diet, for example by Drewnowski, Rehm, and Vieux (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,[46] 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 nutri-

46. 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.

ents). 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 B.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 "bad" nutrients, is only weakly correlated.

In Table B.3, 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 only if "bad" nutrients are not penalized. This result suggests 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 B.3: 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.

C AUXILIARY MODELS

C.1 DEFINITIONS

CES-only. This version of the model sets $\psi = 0$, effectively leaving just the homothetic CES component of the preferences. Thus, the objective function 1 is replaced by

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

The rest of the model remains the same. The caloric contents k_i of foods and the household's caloric requirement $K_{req,h}$ become irrelevant in this model. This CES-only model is calibrated to the same moments except for the output elasticity of K_{in}, which is not targeted as the preferences lack the variable income elasticity of

food consumption needed to match this moment.

Stone-Geary. The second auxiliary model likewise drops the caloric deviation penalty, but also replaces the homothetic CES term with a non-homothetic Stone-Geary utility function, common in the structural transformation literature:

$$\max_{\substack{\{c_{h,i},x_{h,i},x^{p}_{h,i},\\ x^{s}_{h,i}\}_{i=1}^{n},c_{h,m}}} \left((1-\varphi_m) \left(\left(\sum_{i=1}^{n} \varphi_i c_{h,i}^{\frac{\sigma-1}{\sigma}} \right)^{\frac{\sigma}{\sigma-1}} - \bar{c} \right)^{\frac{\gamma-1}{\gamma}} + \varphi_m c_{h,m}^{\frac{\gamma-1}{\gamma}} \right)^{\frac{\gamma}{\gamma-1}} \tag{C.2}$$

where $\bar{c}$ represents the subsistence level of food consumption. The caloric contents k_i of foods and the caloric requirement $K_{req,h}$ are also irrelevant in this model. However, the preferences are non-homothetic (unlike in the CES-only model) and food will occupy a higher fraction of the poor households' resources, like in the baseline model with the caloric channel. This auxiliary model is calibrated to the same parameters as the baseline model, with $\bar{c}$ taking the role of ψ as the main parameter targeting the output elasticity of K_{in}.

C.2 Calibration of Auxiliary Models

Table C.1 compares calibrated parameters that are different in the CES-only and Stone-Geary calibrations compared to the baseline model's calibration. The CES-only model needs a much higher median trade cost $\bar{d}$ to get close to the empirical sold share of 0.16: $\bar{d} = 2.5$ vs $\bar{d} = 1.75$ for the baseline and Stone-Geary models. This is driven by the fact that the poor households in the CES-only model do not treat food as a necessity (as they largely do in the baseline and the Stone-Geary models) and seek to consume significant quantities of both foods and the manufactured good, which has to be purchased on the market with proceeds from farm output sales if the household lacks sufficient non-farm income. The Stone-Geary model, in turn, generates an overly high output elasticity of K_{in} at all values of $\bar{c}$, with the calibrated $\bar{c}$ simply getting the closest with an elasticity of 0.23 vs 0.09 in the data.

TABLE C.1: Calibrated parameters different across models

parameter	**baseline**	**CES-only**	**Stone-Geary**
$\mathbb{E}(\log L_h)$	-14.9	-14.75	-14.85
$\bar{d}$	1.75	2.5	1.75
ψ	0.5	0	0
$\bar{c}$	0	0	0.5×10^{-8}
φ_m	0.36	0.34	0.94
$\bar{p}_{\text{tobacco}}/p_{\text{maize}}$	5.25	6.26	5.36

C.3 ALTERNATIVE $\sigma > 1$ CALIBRATION

The baseline model with the caloric deviation penalty used throughout the paper relies on the elasticity of substitution $\sigma = 0.75$ estimated in Section 4.2. To verify that the predictions of the model are not sensitive to the assumption of $\sigma < 1$, I also conduct an alternative calibration of the model with $\sigma = 1.3$, using the value estimated by Kebede (2024), and recalibrating the other parameters to hit the same target moments as in the baseline calibration.

The behavior of farmers in the alternative calibration is largely similar to that in the baseline calibration: see Table C.2 below. Once calibrated, the model with $\sigma > 1$ gets close to the low empirical output elasticity of caloric intake. It also successfully reproduces the strong positive relationship between farm size (or income) and food diversity.

The aggregate predictions of the alternative calibration are likewise similar to the baseline one: removing the domestic trade frictions entirely promises an 88% increase in agricultural GDP (compared to 84% in the baseline calibration), while a partial removal of trade frictions to raise the average share of output sold to 50% promises a 47% increase in agricultural GDP (same as in the baseline calibration).

Thus, the qualitative micro and macro behavior of the model does not hinge on the estimated value of σ being below 1.

C.4 Alternative Model Farm Behavior

Table C.2 shows the dependency of household food consumption on farm size in the Stone-Geary version of the model, defined in Eqn. C.2, and in the *sigma* > 1 calibration of the baseline model.

Table C.2: Household food consumption vs farm size: Stone-Geary model, $\sigma > 1$ calibration, and data

	log kcal intake			**food diversity**		
	(1) **model**: Stone-Geary	(2) **model**: $\sigma > 1$	(3) **data**	(4) **model**: Stone-Geary	(5) **model**: $\sigma > 1$	(6) **data**
log output	0.233	0.137	0.091***	−0.100	0.343	0.395***
	(0.001)	(0.001)	(0.005)	(0.001)	(0.001)	(0.034)
log non-farm income	0.203	0.112	0.063***	0.012	0.330	0.857***
	(0.001)	(0.001)	(0.004)	(0.001)	(0.002)	(0.033)
N	70,937	62,923	8,674	70,937	62,923	8,675
Adj. R^2	0.762	0.509	0.063	0.134	0.642	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.

Just as in the baseline caloric needs model, poor households in the Stone-Geary model devote a disproportionate share of their resources to obtaining food. This generates low output and income elasticities of kcal intake (0.233 and 0.203) compared to the CES-only model. However, these elasticities are still much higher than the empirical ones (0.091 and 0.063): despite targeting those elasticities with the $\bar{c}$ moment, the model with Stone-Geary preferences simply cannot get nearly as close to the observed elasticities as the baseline model with the caloric channel can. Moreover, Stone-Geary preferences with a subsistence level in food consumption only create income dependence in the allocation between food and the manufactured good but not in the allocation between different foods. Therefore, the Stone-Geary model counterfactually predicts close to no relationship between farm size and dietary diversity (compare columns (4) and (6) of Appendix Table

C.2), performing no better than the CES-only model. Compared to the baseline model, the auxiliary model with Stone-Geary preferences lacks only the explicit caloric channel. As a result, although the simpler Stone-Geary preferences get closer to the baseline model in reproducing the observed calorie consumption behavior than the CES-only model does, they still fall short of the empirical moment, and fail in reproducing the food diversity behavior even qualitatively.

In contrast, the alternative $\sigma > 1$ calibration of the baseline caloric needs model successfully produces low output and income elasticities of kcal intake and high output and income elasticities of food diversity that are close to empirical values and to those in the baseline calibration (Table 3).

D Additional Figures and Tables

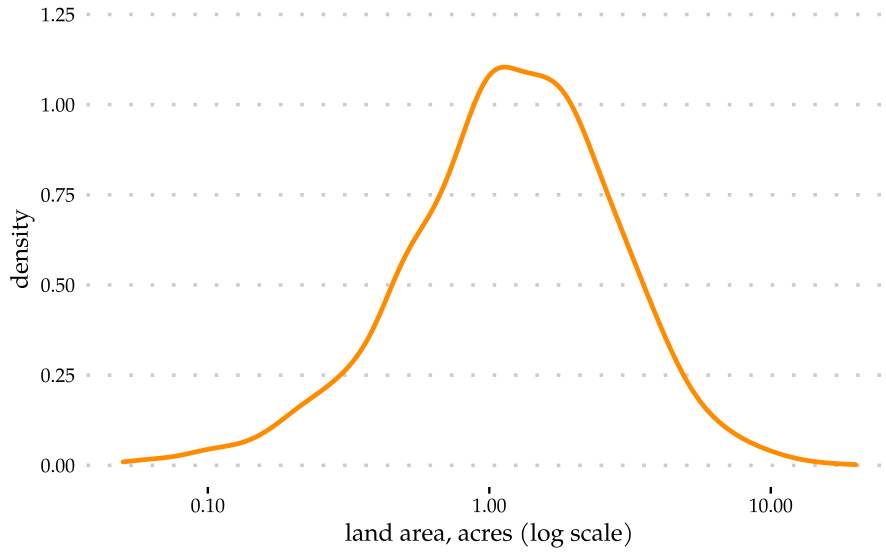
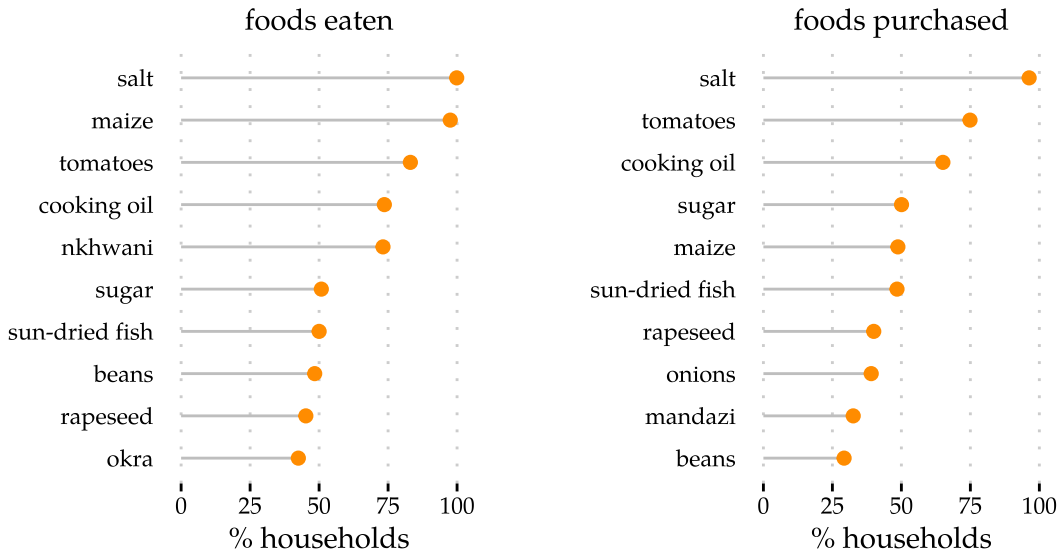
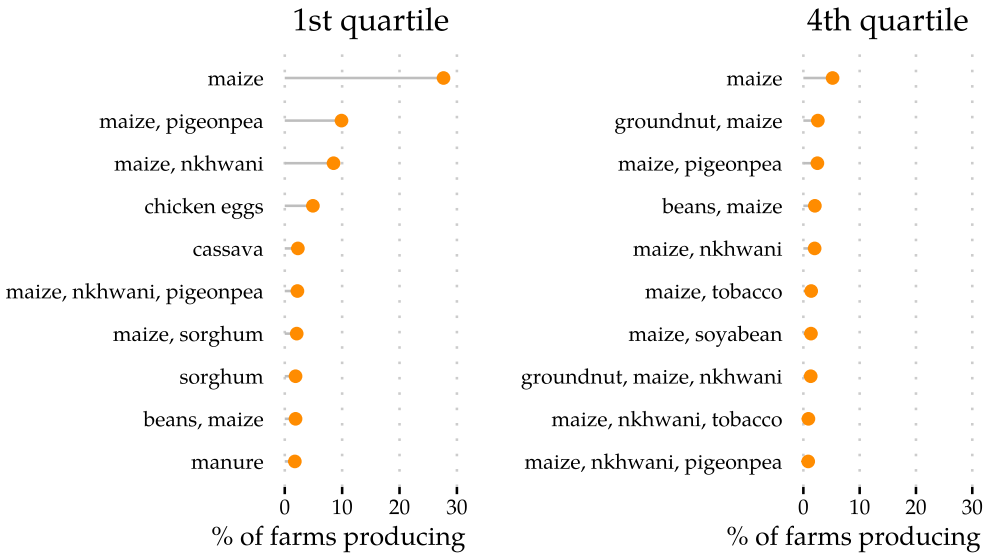


Figure D.1: Farm area distribution



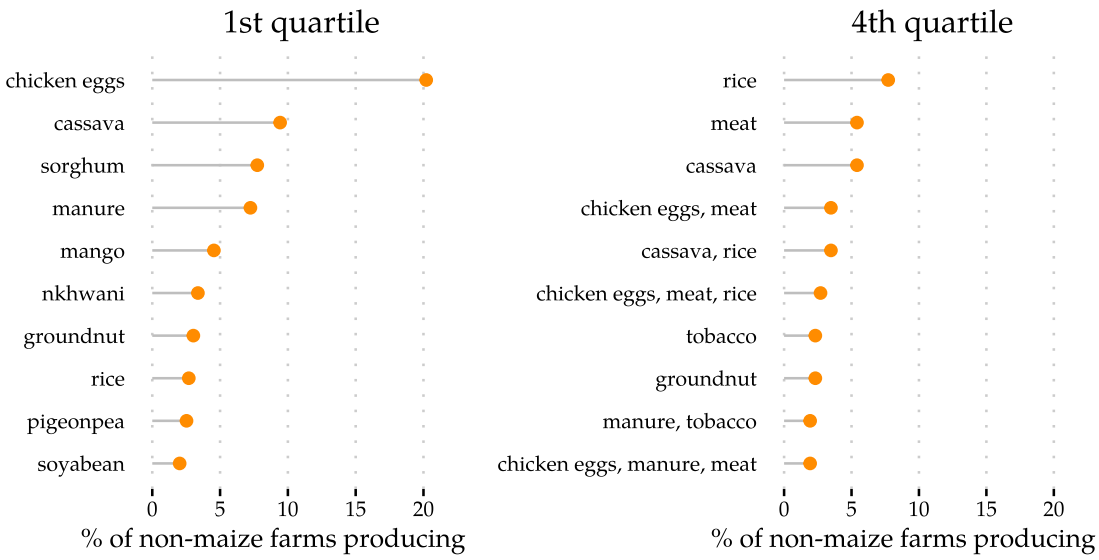
Note. Foods are ranked by the % of households consuming each food (left) or purchasing each food (right), and only the top 10 are displayed.

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



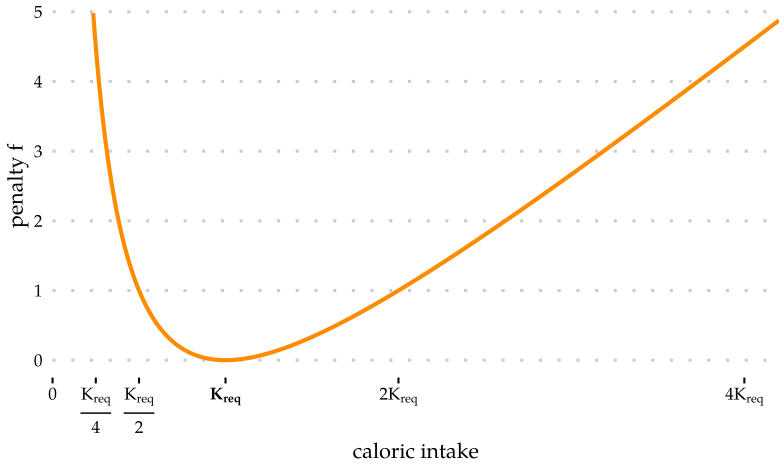
Note. Farms are grouped by output quartile (output market value relative to the caloric requirement). For each group and each combination of products, the percentage of farms in that group that produce that combination is computed. Product combinations are ranked by this measure within each group, and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

Figure D.3: Product combinations ranked by the % farms within each output quartile producing each combination



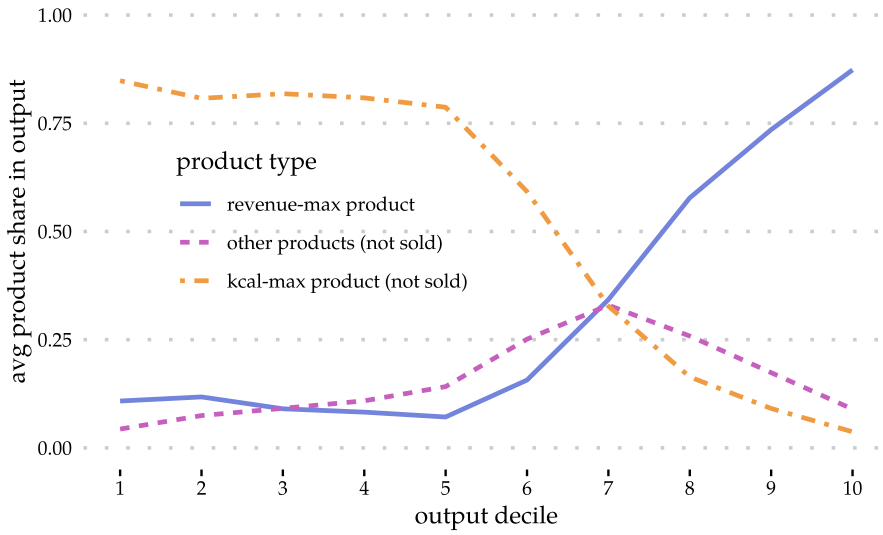
Note. Farms are grouped by output quartile (output market value relative to the caloric requirement). For each group and each combination of products, the percentage of farms in that group that produce that combination is computed. Product combinations are ranked by this measure within each group, and only the top 10 are displayed. Unlike Appendix Figure 3, only the farms not producing maize are considered.

Figure D.4: Product combinations ranked by the % of non-maize farms within each output quartile producing each combination



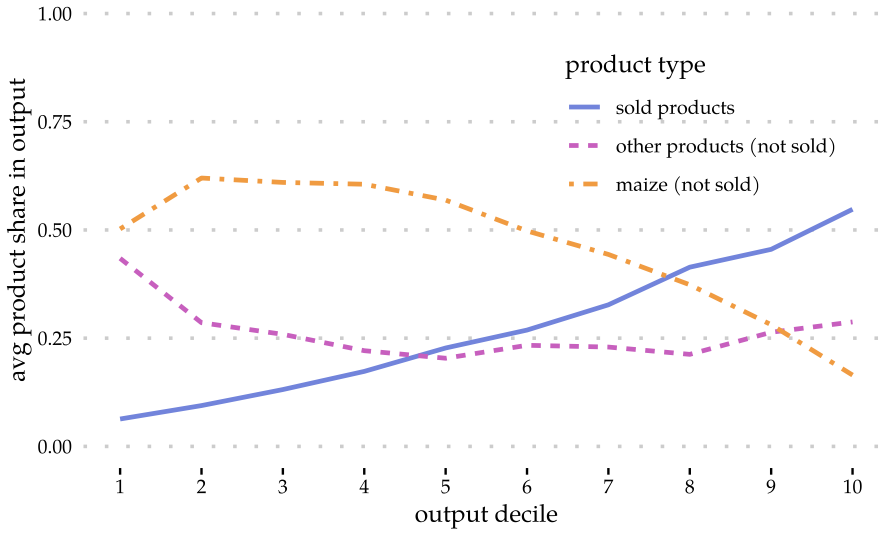
NOTE. $f\left(\sum_i c_i k_i, K_{req}\right) = \psi\left(\frac{\sum_i c_i k_i - K_{req}}{K_{req}}\right)^2 \frac{K_{req}}{\sum_i c_i k_i}$. Value $\psi = 2$ used for illustration. The x-axis is caloric intake expressed as ratios of caloric requirement K_{req}.

FIGURE D.5: Penalty function f for deviations of caloric intake from caloric requirement



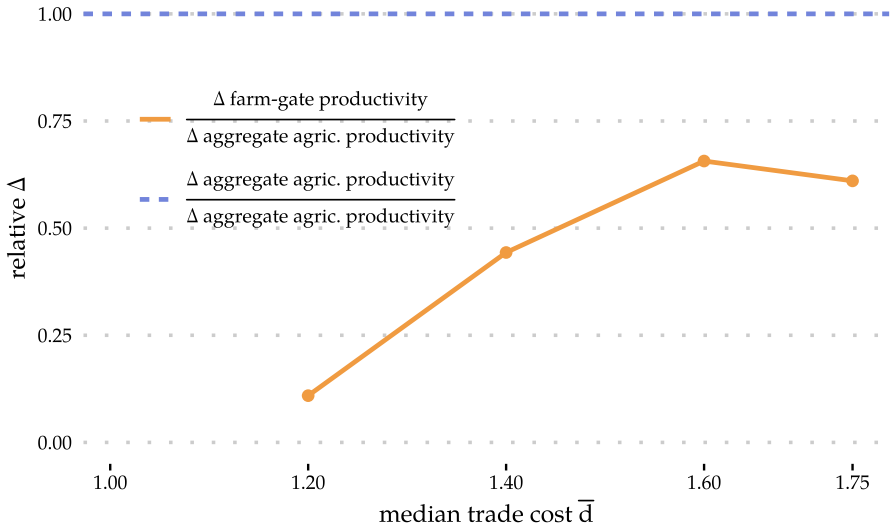
NOTE. Products are split into groups at farm level. "kcal-max product (not sold)" is the $\arg\max_i k_i z_{h,i}$ good, unless it's the same as the $\arg\max_i p_i z_{h,i}$ and is sold. "revenue-max product" is the $\arg\max_i p_i z_{h,i}$ good, unless it's the same as the $\arg\max_i k_i z_{h,i}$ and 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 D.6: Farm size and product choice (average shares): model



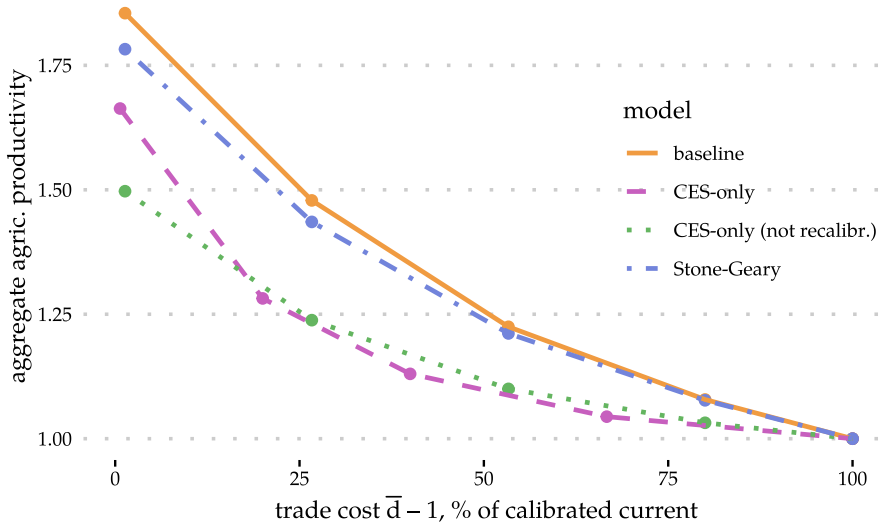
Note. Products are split into groups at farm level. Avg product shares are the average output value shares of each product group among farms in each output value decile.

Figure D.7: Farm size and product choice (average shares): data



Note. Productivity changes are between consecutive levels of $\bar{d}$ (marginal trade cost reductions).

Figure D.8: Share of product choice changes in marginal productivity gains



NOTE. Aggregate agricultural output is valued at market prices. Each series is chain-weighted between each pair of consecutive $\bar{d}$ levels to obtain real values. 100% of trade cost represents the median trade cost $\bar{d}$ calibrated for each model, 0% of trade cost represents a complete counterfactual removal of trade costs. Productivity at the calibrated level of $\bar{d}$ is normalized to 1.

FIGURE D.9: Aggregate agricultural productivity and trade cost across models

TABLE D.1: List of products produced by households

product	% farms	product	% farms
maize	86.7%	pearl millet	1.0%
nkhwani	28.6%	tanaposi	1.0%
pigeonpea	23.4%	mexican apple	0.5%
chicken eggs	18.7%	orange	0.5%
manure	15.1%	finger millet	0.4%
groundnut	13.5%	okra	0.4%
beans	11.8%	masau	0.3%
mango	11.4%	onion	0.3%
meat	9.6%	tangerine	0.2%
soyabean	8.6%	sugar cane	0.2%
cassava	7.4%	skins and hides	0.2%
sorghum	7.3%	cabbage	0.2%
tobacco	5.8%	lemon	0.2%
rice	5.2%	guinea fowl eggs	0.1%
sweet potato	3.5%	coffee	0.1%
peas	2.0%	paprika	0.1%
banana	1.7%	peach	0.1%
cotton	1.6%	tea	0.1%
sunflower	1.6%	pineapple	0.1%
tomato	1.5%	custard apple	0.0%
avocado	1.3%		
guava	1.3%		
cow milk	1.2%		
papaya	1.1%		
irish potato	1.0%		

TABLE D.2: Elasticity of kcal intake, alternative specifications: data

	log kcal intake	
	(1)	(2)
log land area	0.120***	
	(0.008)	
log non-farm income	0.063***	
	(0.004)	
log shadow income		0.121***
		(0.006)
N	8,417	9,762
Adj. R^2	0.062	0.051

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

NOTE. Kcal intake, land area, non-farm income, and shadow income are relative to the caloric requirement of the household. Shadow income is the sum of output value and non-farm income.

TABLE D.3: Trade cost and the correlation between product shares in consumption and production: model and data, PCA proxy

	product share correlation			
	cons. value, prod. value		**cons. kcal, prod. value**	
	(1) **model**	(2) **data**	(3) **model**	(4) **data**
log output	−0.081	−0.020***	−0.082	−0.012***
	(0.001)	(0.003)	(0.001)	(0.003)
log non-farm income	−0.057	−0.046***	−0.075	−0.026***
	(0.001)	(0.002)	(0.001)	(0.003)
trade cost	0.490	0.031***	0.514	0.020***
	(0.032)	(0.003)	(0.035)	(0.004)
N	70,750	8,100	70,750	8,099
Adj. R^2	0.108	0.075	0.112	0.022

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

NOTE. Kcal intake, output, and non-farm income are relative to the caloric requirement of the household. Product share correlation is calculated across products within each household. "cons. value, cons. kcal, prod. value" indicate how each share was measured: in market value or in kcal intake. "trade cost" denotes d_h in the model and the PCA proxy (the first principal component of distances between each farm and the nearest population center, agricultural market, and road) in the data.

E Computational Algorithm

This section describes the computational algorithm that solves the baseline caloric needs model for a given household. Household index h is omitted.

Notation. In addition to notation defined in Appendix A:

Let X be the set of agricultural goods the household chooses to produce ($x_i > 0$). Let P be the set of agricultural goods the household chooses to purchase ($x_i^p > 0$).

Let $\tilde{\tau}_i = \tilde{\eta}_i + \frac{f_1(K_{in}, K_{req})k_i}{\mu}$, where $\tilde{\eta}_i = \frac{\lambda}{\mu z_i}$ for $i \in X$, $\tilde{\eta}_i = p_p d$ for $i \in P$.

Let $C = \sum_i \varphi_i^\sigma \tilde{\tau}_i^{1-\sigma}$.

Conditional on a guess of which of the goods the household chooses to produce ($x_i > 0$), which to purchase ($x_i^p > 0$), and which to sell ($x_i^s > 0$), solving the problem requires solving a system of two or three non-linear equations, described further down. The main obstacle is that there are $\left(2^3\right)^n$ possible product choice guesses of whether each of the n goods ($n = 6$ in the calibration) is produced, purchased, and sold, making the problem's runtime exponential in n. However, it can be shown (the derivation is straightforward but lengthy and so omitted) that all but a handful of these guesses lead to contradictions and are certainly suboptimal, leaving at most $3n + 1$ product choice combinations to be checked. Hence, omitting all contradictory guesses before iterating through them makes the problem's runtime linear in n. Below I describe the algorithm for assembling the list of product choice combinations (guesses) that do need to be checked.

Algorithm for assembling a list of product choice guesses. Firstly, at most one good is both produced and sold ($x_i > 0, x_i^s > 0$) at the same time, call this good j. Likewise, at most one good is both produced and purchased ($x_k > 0, x_i^p > 0$), call this good k. In addition to guessing which (if any) good is j and which (if any) is k, also need to guess which goods are produced $x_i > 0$ (group X) and which are purchased $x_i^p > 0$ (group P).

1. Identify good $h \equiv \arg\max_i p_i z_i$.
2. For goods with $p_i z_i d \geq \frac{p_h z_h}{d}$, rank them by $p_i z_i$. These are the potentially produced goods, call this set E. For $e \in \{1, \ldots, |E|\}$:

(a) guess that top e goods in E are actually produced (fall in group X: $x_i > 0$ for $i \in X$), the rest are purchased (fall in group P: $x_i^p > 0$ for $i \in P$).

(b) construct several guesses to be solved:

e.1 goods j, k don't exist.

e.2 good h is j, good k doesn't exist (this case only needs to be added once, for $e = |E|$).

e.3 $\arg\min_{i \in X} p_i z_i$ is good k, and good j doesn't exist.

e.4 good h is j, $\arg\min_{i \in X} p_i z_i$ is good k (this requires $\frac{z_j p_j}{d} = z_k p_k d$ and does not occur in the calibrated model).

This procedure produces $3 \cdot |E| + 1$ product choice combinations that need to be checked, which is at most $3n + 1$. For each of these, execute the following algorithm.

Algorithm for solving the household's problem with a given product choice combination.

1. solve a system of equations for λ, μ, K_{in}

   – if this combination is of type e.1, the system is:

$$\frac{K_{in} \sum_{i \in X} \varphi_i^\sigma \tilde{\tau}_i^{-\sigma} \frac{1}{z_i}}{\sum_i \varphi_i^\sigma \tilde{\tau}_i^{-\sigma} k_i} = L$$

$$\frac{K_{in} \left(\sum_{i \in P} \varphi_i^\sigma \tilde{\tau}_i^{-\sigma} p_i d + \left(\frac{\varphi_m}{1-\varphi_m} \right)^\gamma p_m^{1-\gamma} C^{\frac{\sigma-\gamma}{\sigma-1}} \right)}{\sum_i \varphi_i^\sigma \tilde{\tau}_i^{-\sigma} k_i} = wn$$

$$\left((1-\varphi_m)^\gamma C^{\frac{\gamma-1}{\sigma-1}} + \varphi_m^\gamma p_m^{1-\gamma} \right)^{\frac{1}{\gamma-1}} = \mu$$

   – if this combination is of type e.2, the system is:

$$\frac{K_{in} \left(\sum_{i \in x} \varphi_i^\sigma \tilde{\tau}_i^{-\sigma} \frac{1}{z_i} + \frac{d}{p_j z_j} \left(\sum_{i \in P} \varphi_i^\sigma \tilde{\tau}_i^{-\sigma} p_i d + \left(\frac{\varphi_m}{1-\varphi_m} \right)^\gamma p_m^{1-\gamma} C^{\frac{\sigma-\gamma}{\sigma-1}} \right) \right)}{\sum_i \varphi_i^\sigma \tilde{\tau}_i^{-\sigma} k_i} = L + wn \frac{d}{p_j z_j}$$

$$\left((1-\varphi_m)^{\gamma}C^{\frac{\gamma-1}{\sigma-1}}+\varphi_m^{\gamma}p_m^{1-\gamma}\right)^{\frac{1}{\gamma-1}}=\mu$$

$$\frac{\lambda}{\mu}=\frac{p_j z_j}{d}$$

– if this combination is of type e.3, the system is:

$$\frac{K_{in}\left(\sum_{i\in X}\varphi_i^{\sigma}\tilde{\tau}_i^{-\sigma}\frac{1}{z_i}+\frac{1}{p_k z_k}\left(\sum_{i\in P}\varphi_i^{\sigma}\tilde{\tau}_i^{-\sigma}p_i d+\left(\frac{\varphi_m}{1-\varphi_m}\right)^{\gamma}p_m^{1-\gamma}C^{\frac{\sigma-\gamma}{\sigma-1}}\right)\right)}{\sum_i\varphi_i^{\sigma}\tilde{\tau}_i^{-\sigma}k_i}=L+wn\frac{1}{p_k z_k}$$

$$\left((1-\varphi_m)^{\gamma}C^{\frac{\gamma-1}{\sigma-1}}+\varphi_m^{\gamma}p_m^{1-\gamma}\right)^{\frac{1}{\gamma-1}}=\mu$$

$$\frac{\lambda}{\mu}=p_k z_k d$$

2. compute consumptions:

$$c_m=\frac{K_{in}\left(\frac{\varphi_m}{1-\varphi_m}\right)^{\gamma}p_m^{-\gamma}C^{\frac{\sigma-\gamma}{\sigma-1}}}{\sum_i\varphi_i^{\sigma}\tilde{\tau}_i^{-\sigma}k_i}$$

$$c_i=K_{in}\frac{\varphi_i^{\sigma}\tilde{\tau}_i^{-\sigma}}{\sum_i\varphi_i^{\sigma}\tilde{\tau}_i^{-\sigma}k_i}$$

3. compute productions, purchases, and sales:

– if the combination is of type e.2, calculate x_j^s from:

$$x_j^s=z_j\left(L-\frac{K_{in}\sum_{i\in X}\varphi_i^{\sigma}\tilde{\tau}_i^{-\sigma}\frac{1}{z_i}}{\sum_i\varphi_i^{\sigma}\tilde{\tau}_i^{-\sigma}k_i}\right)$$

– if the combination is of type e.3, calculate x_k^s from:

$$x_k^p=z_k\left(\frac{K_{in}\sum_{i\in X}\varphi_i^{\sigma}\tilde{\tau}_i^{-\sigma}\frac{1}{z_i}}{\sum_i\varphi_i^{\sigma}\tilde{\tau}_i^{-\sigma}k_i}-L\right)$$

– obtain the rest of $\{x_i, x_i^s, x_i^p\}$ directly from goods constraints.

4. verify that the solution satisfies all constraints.

After the household's problem has been solved for all allowed product choice combinations, pick the solution that maximizes the household's utility. In practice, there is always just one product choice combination whose solution satisfies all constraints of the maximization problem, making the step of comparing utilities across guesses redundant.

The algorithms for solving the two auxiliary models (CES-only and Stone-Geary) are omitted. They share the same algorithm for assembling product choice guesses. The structure of the algorithm solving the problem for each guess is the same, but the systems of equations needed to be solved are simpler.