Dataset Description

This data was collected to better understand the retail market for key food items in urban Tanzania, including maize flour, rice, mixed flour, and other flour. This data is split into multiple sections. Sections A and B include basic information about the retail outlet and owner/respondent, section B includes information about various retail practices, section C is a product inventory containing key information about the products sold, section D includes information about product sales and procurement, sections E contain information about discontinued products and possible new brands, sections G contain information about transaction characteristics and dynamics in maize flour and mixed flour products, section H is on employment, and section K is on perceptions on maize and mixed flour products. Section "CD" is also included, and it is a merge of sections C and D1 while excluding observations with missing data, and which took some cleaning to get a clean merge. This merged dataset excludes large supermarkets, where the data for section D is highly incomplete.

Sampling

Dar es Salaam

We conducted a survey of retail outlets in Dar es Salaam, April and May, 2016[1]. As depicted in figures 1 and 2, we used a multi-stage sampling strategy of four types of food retail outlets in the two cities. From most- to least traditional types of food retail outlets (see table 1), these types are: non-self-service open air municipal market stalls; small enclosed shops (dukas), located both within and outside of municipal markets, that are the most ubiquitous and appear to serve the majority of the population for most of their food needs; self-service grocery stores (or mini-supermarkets); and large chain and independent supermarkets.

[1] A few large supermarkets were covered in later months

Table 1. Typology of retail

Type of retail outlet	Market or Non-market	Traditional to modern	Identifying features
Open air stalls	Market	Traditional	Not self-service; not enclosed
Dukas	Both		Not self-service; enclosed
Mini supermarket	Mostly non-market	Modern	Self-service; relatively small
Large supermarket	Non-market		Self-service; relatively large; chain or independent

In each city we first conducted a census of all large supermarket outlets (in Dar es Salaam, 29 stores across 14 chains, and in Arusha, 4 stores across 3 chains[2]). Due to difficulties obtaining the full collaboration of managers in many of these large outlets, with few exceptions we conducted "product inventories" rather than full surveys. These inventories established the full range of processed products available among maize meal, other flour, rice, and lishe (mixed "healthy" flour), fruit juice products (including sodas that contained some juice content), and dairy products, and their prices, but did not obtain data on sales and procurement.

For all other retail outlet types, we conducted stratified, multi-stage random sampling[3] of outlets that contained at least one of the following items mentioned above. The sampling consisted of three general stages. Strata were districts (three in Dar es Salaam, and one in Arusha), outlets within markets versus those outside of markets, and outlet types. Sampling steps in Dar es Salaam were as follows (see figure 1):

1. We randomly selected 10 markets in each district.
2. We stratified on outlets within each market by listing all outlets and then conducting a separate random sample[4] of stalls, shops that carried lishe products, and shops that did not carry these

[2] Two of these supermarkets were part of the same local Arusha chain, and the other 2 were part of two separate chains that also have representation in Dar es Salaam

[3] All random sampling at each level was conducted by (1) listing the universe of sampling units, (2) assigning each unit a random number in Microsoft Excel, and (3) sorting and sampling the units with the lowest random numbers until the sample size was reached

[4] Details on sample sizes, etc., can be found in table 1.

products; this step was necessary in order to ensure sufficient lishe observations, as the number of outlets carrying them was relatively small.

3. For all outlets not located within markets, we randomly selected 10 wards in each district.
4. We conducted a census of all mini-supermarkets in each of these wards.
5. We randomly selected one Mtaa within each ward[5], listed all shops carrying and not carrying lishe products in these Mtaas, and randomly sampling each type of shop in each mtaa, respectively.

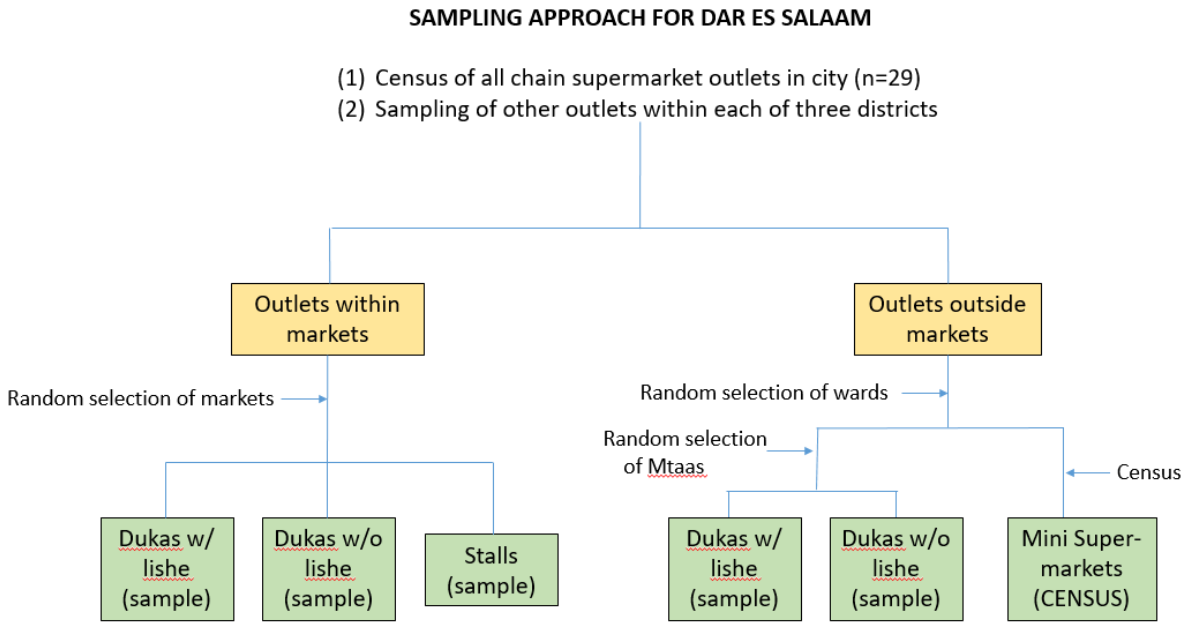

Figure 3. Sampling approach for Dar es Salaam

Arusha

We conducted a survey of retail outlets in Arusha, July and August, 2016. Sampling in Arusha followed a similar approach, with a few exceptions (see figure 2):

1. In all but one market (Kilombero), all outlet types were covered in a census, rather than a sample.

[5] In Dar es Salaam, we first took a sub-sample of 6 wards out of the 10, before taking a sub-sample of mtaas. We did this based on time and cost considerations.

2. In non-markets, traditional outlets were not stratified by store type, but instead by previously listed and newly discovered outlets.
3. The final sample size, and sample sizes per Mtaa, and the total number of sampled Mtaas, varied across cities, due to both the differing number of administrative units in each city, and cost considerations[6].

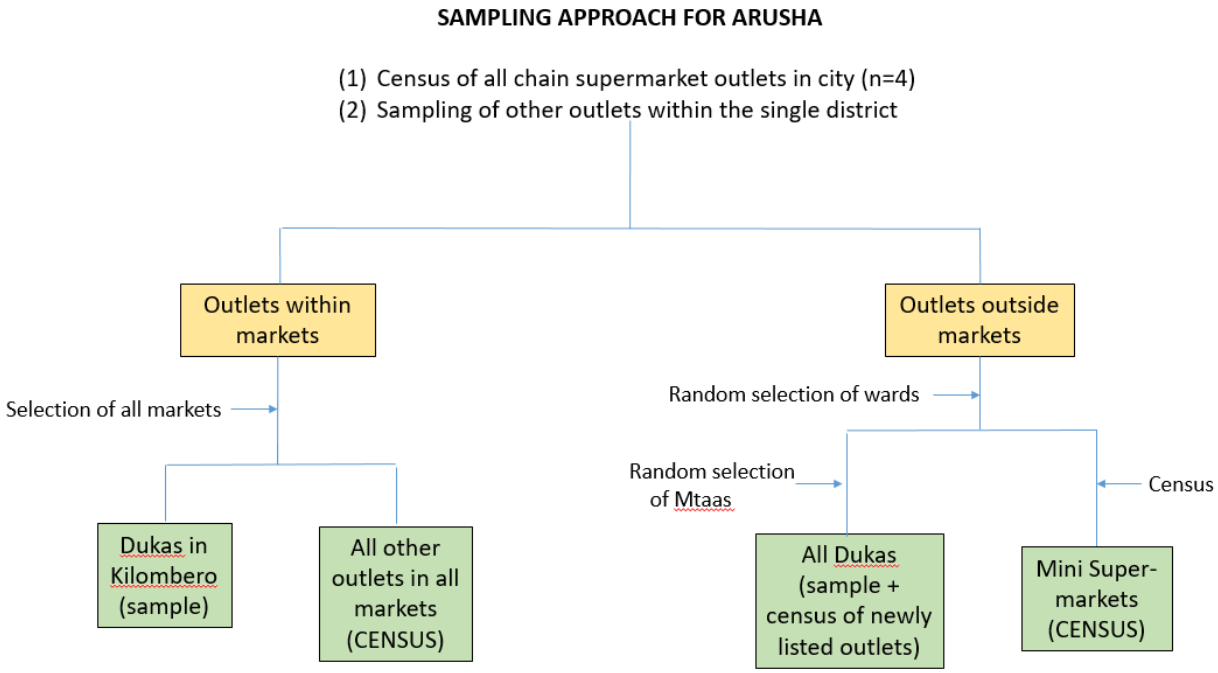

Figure 4. Sampling approach for Arusha

Details of the sampling strategy, original listing size, final sample size, and estimated population size in the retail survey for each city are shown in table 2 below[7].

[6] Arusha is a much smaller city with many fewer wards and mtaas. We decided to sample one mtaa each from 12 wards (which is 48% of all wards in Arusha) instead of from 18 wards like in Dar es Salaam (which is 18% of all wards in Dar es Salaam), and then take a larger maximum sample within each mtaa.
[7] Further information about population weighting for both the retail and miller surveys can be found in the appendix

Table 2. Details of retail sampling strategy

City	Market or Non-Market	Sampling cluster	Sampling strategy	Original Listing size	Final Sample size	Estimated population size
Dar es Salaam	Market	Shop - Lishe	Attempted full census in each of the 30 sampled sokos	34	30	57
		Shop - non-Lishe	Maximum sample size of 3 outlets in each of the 30 sampled sokos	154	68	224
		Stall	Maximum sample size of 8 outlets in each of the 30 sampled sokos	166	106	260
	Non-market	Shop - Lishe	Maximum sample size of 7 outlets in each of the 18 sampled mtaas	227	106	7349
		Shop - non-Lishe	Maximum sample size of 11 outlets in each of the 18 sampled mtaas	1071	196	37226
		Mini supermarket	Attempted full census in each of the 30 sampled wards	108	95	330
		Large supermarket	Full census, observational data and some partial/full surveys	29	29	29
	Total DES	**All Outlets**		**1789**	**630**	**45475**
Arusha	Market	Shops	Full census (see text for exception)	NA*	108	122
		Stall	Full census	NA	46	46
		Mini supermarket	Mini supermarkets found during survey in sokos	NA	7	7
	Non-market	Shop	Maximum sample size of 30 outlets in each of the 12 sampled mtaas. Census of new outlets discovered	330	278	3409
		Mini supermarket	Attempted full census in each of the 12 sampled wards	70	90**	190
		Large supermarket	Full census, observational data	4	4	4
	Total Arusha	**All Outlets**		**NA**	**533**	**3778**

*Due to small original sample size and large turnover from original listing, most sokos were relisted during survey and complete census was taken; **Sample size includes new mini-supermarkets found since the listing; “Maximum sample size” means that we took a sample of that size if there were more than that many outlets. If there were fewer outlets, then we took a full census.

Weighting

Dar es Salaam

We calculated the ward/district level weights (*DWgt*) as the inverse of the probability that each ward would be selected, i.e., *DWgt =* N_i/n_i where N_i is the total number of wards in district *i and* n_i is the total number of wards selected within that district.

The mtaa/ward level weights (*MWgt*) were calculated as the inverse of the probability that each duka would be selected, i.e., *MWgt = DWgt*(10/6)** (M_s/m_s), where (M_s/m_s) is the inverse probability that the mtaa would be selected within a given ward. M_s is the total number of mtaas in ward *s*, and m_s is the sampled mtaa within that ward.

Much[8] of the survey was conducted in the Spring of 2016 but the listing took place in the summer of 2015. We found that there had been a lot of turnover in shops, many new and closed or not found. We assumed that shops not found had been closed, and then we calculated the ratio of known shops that had been closed from the list. We then applied this ratio to the remaining replacement shops that we did not approach. Since we only sampled new outlets that were needed as a replacement, we made an assumption about the share of new shops to closed/not found shops based on the ratio found in Arusha (where we interviewed all new shops that were found) outside of sokos.

We calculated the final sample weight outside of markets as SWgt = MWgt*Dt/dt, where Dt is the estimated total population of outlets within a mtaa, and dt the number of sampled outlets. We estimated Dt by subtracting the estimated number of closed shops from the initial number that was listed, and then adding the estimated number of new shops to that number. That is, let L be the initial listing, C be the total number in the sample that were closed or not found (including replacements), R be the remaining replacement shops not approached, N be the number of new shops in Arusha, and C_a be the estimated number of closed shops in Arusha. Then (C + (C/L)*R) is the estimated number of closed shops, N/C_a is the estimated ratio of new to old shops (from Arusha), and N/C_a*(C + (C/L)*R) is the estimated number of new shops. Then, Dt = (L – (C + (C/L)*R) + N/C_a*(C + (C/L)*R)).

The final sample weight is calculated as SWgt = MWgt*(L – (C + (C/L)*R) + N/C_a*(C + (C/L)*R))/dt. Note that C_a is an estimate because there were 10 outlets in Arusha of which we did not know the status. For those outlets, we assumed that the share of them that are closed was equal to the share in the total

[8] Some of the large supermarkets were covered later

listing that were closed. There were 89 closed out of 307 known, 10 unknown, and 73 new outlets. Therefore, N = 73 and C_a = (89+(89/307)*10). Rewriting the formula above, we get SWgt = MWgt*(L – (C + (C/L)*R) + 73/(89+(89/307)*10)*(C + (C/L)*R))/dt

We calculated the soko weights using the following formula: SkWgt = (K_i/k_i), where K_i equals the number of sokos in the district *i*, and k_i is the number of sampled sokos in that district.

Since we only knew which sokos contained our food items from the ones that we tried to sample, we applied the share of known sokos containing our food items to the remaining list of unknown sokos, in order to estimate the final for each district. We calculated the final weights in sokos using a similar type of formula that was used in the mtaas: SWgt = SkWgt*Dt/dt, where Dt and dt are calculated the same way as above.

Arusha

We calculated the ward level weights (*DWgt*) as the inverse of the probability that each ward would be selected, i.e., *DWgt = N/*n where N is the total number of wards in the city and n is the total number of wards sampled. We adjusted the weight for the mini supermarkets in the wards that had a refusal or an outlet with an unsure status. For example, in Levolosi there were 15 initially sampled shops, 1 refusal, 1 closed shop, 4 new shops, and 1 with unsure status. This means there were 16 shops interviewed, 17 for sure in the listing population, and an estimated 1/9th of another shop in the population (that refused interview). The weighting factor then was DWgt*((17+1/9)/16).

The mtaa/ward level weights (*MWgt*) were calculated as the inverse of the probability that each duka would be selected, i.e., *MWgt = DWgt** (M_s/m_s), where (M_s/m_s) is the inverse probability that the mtaa would be selected within a given ward. M_s is the total number of mtaas in ward *s*, and m_s is the sampled mtaa within that ward.

We conducted the original listing around August 2015 and the survey in July and August of 2016. Given the lag, we found that there had been a lot of turnover in shops (i.e., many new, closed, and not found). We assumed that shops that we did not find again were closed, and then we calculated the ratio of known shops that had been closed to the entire original list. We applied this ratio to the remaining replacement shops that we did not approach in order to estimate the total number of closed shops. In order to produce the final weighting, we subtracted the estimated number of closed shops from the

initial listed number, and added the number of new shops. That is, let L be the initial listing, C be the total number of known shops closed or not found, R be the remaining replacement shops not approached, and N be the number of new shops found. Then, the final sampling weight is calculated as SWgt = MWgt*(L – (C + C/L)*R) + N.

We conducted a full census in all of the sokos and so we did not apply a weight. One exception to this was the Kilombero soko; there were many more shops than we anticipated and so we took a sample of 38 out of 52 listed dukas. In this case, SWgt equaled 52/38. We still took a full census of stalls.

Stratification and FPC variables

The primary-stage stratification categories are wards, markets, and large supermarkets, which we delineate by the variable called *Strata_Primary.* Large supermarkets are differentiated by city, and wards and markets are differentiated by district (and implicitly by city). The primary sampling units (PSUs) are also wards, markets, and large supermarkets, which we delineate by the variable *PSU*. We name the finite population correction (FPC) variable for the second stage *FPC1*

The secondary-stage stratification categories are mini supermarkets, the subset of sampled wards for non-mini supermarkets, and placeholder stratum for markets and large supermarkets. We delineate these by the variable called *Strata_MiniNMini*. The secondary sampling units (SSUs) are the same as the PSU (the only new selection is wards within wards for the non-mini supermarket strata), which we delineate using the variable called *SSU1*. We name the FPC variable for the second stage *FPC2*

The third stage does not involve stratification because the stratification from the previous stage applies. The secondary sampling units represent mini supermarkets and wards representing individually sampled mtaas (there is only 1 mtaa sampled per ward), along with the placeholder strata for markets and large supermarkets. The variable is called *SSU2*. The FPC variable for the third stage is *FPC3*

The fourth stage stratification categories differ across Dar es Salaam and Arusha. In Dar es Salaam there are three categories: (a) outlets with lishe, (b) outlets without lishe, and (c) stalls. In Arusha outside of markets, they are (a) new outlets and (b) a sample of listed outlets. In Arusha markets they are (a) all outlets grouped together in non-Kilombero markets, (b) stalls inside Kilombero market, and (c) dukas inside Kilombero market. There are also placeholder stratum for large supermarkets and non-market mini supermarkets. We delineate this variable as *Strata_NSType*. The SSUs are the units within each of

these categories (i.e., the outlet code), which we delineate using the variable called *SSU3.* Finally, we name the FPC variable for this stage FPC4.

Finally, in order to properly specify the sample design, we recommend the following code in STATA:

```
svyset PSU [pweight=SWgt], strata(Strata_Primary) fpc(FPC1) || SSU1, strata(Strata_MiniNMini) fpc(FPC2)  || SSU2,  fpc(FPC3) || SSU3, strata(Strata_NSType) fpc(FPC4) singleunit(scaled)
```