Dataset Description

This data was collected to better understand the maize and other flour miller/brand market in Dar es Salaam, Tanzania. Section A includes information on the background of the owner/manager. Section B includes information on the start-up, status, and finance of the firm, section C includes information on the products and innovations in products and activities. Section D contains information on trainings and association participation. Section E contains information on assets, access to infrastructure, and collaboration with other firms. Section F contains information on procurement of inputs. Section G contains information on labor, production, and technology. Section H contains information on sales. Finally, section I contains information based on qualitative questions about the business environment and expectations for the coming year.

Sampling

As depicted in figure 1 we used a multi-stage sampling strategy of small and medium sized maize flour firms within Dar es Salaam during September through November of 2016:

1. We selected known maize milling "clusters" (i.e., areas that contained many maize milling firms in close proximity) within the city by consulting with local key informants to identify all wards that fell within or intersected known or hypothesized maize milling clusters. Using this approach, we selected eight wards within Temeke district, three wards within Kinondoni district, and three wards within Ilala district.
2. We took a random sample of six wards within each of the three districts (totaling eighteen wards) from all of the remaining wards that we did not identify as falling within or intersecting with clusters.
3. We conducted a full listing of maize mills in each of the selected wards. Note that we did this before listing and sampling milling service buyers.
4. We then took a random sample of 12 mills in each of the wards (or a full census if there were fewer than 12 mills). The final listing had 312 mills and the final sample had 222 mills.
5. We sampled service buyers (SBs) (i.e., firms that don't own their own mill and instead purchase milling services).
    a. To identify these service buyers, we first identified the mill owners from our listing that claimed to provide milling services for other flour firms.

b. Then, we sampled up to three mills per ward in the wards where at least one of these mills met this criterion. This produced a list of 30 mills.
c. Next, we randomly assigned a day of the week (excluding Sunday) to each of these firms so that five mills were assigned to each day of the week, and also randomly chose replacements for each day of the week.
d. Enumerators visited each of these mills[1] for an entire day, listing any identified firms that generally purchased milling services from that mill (but not necessarily milling on that day)[2].
e. We found a total of 117 service buying firms through this method and were able to arrange and conduct interviews with 91 of them. These comprised the second stratum in the second stage.

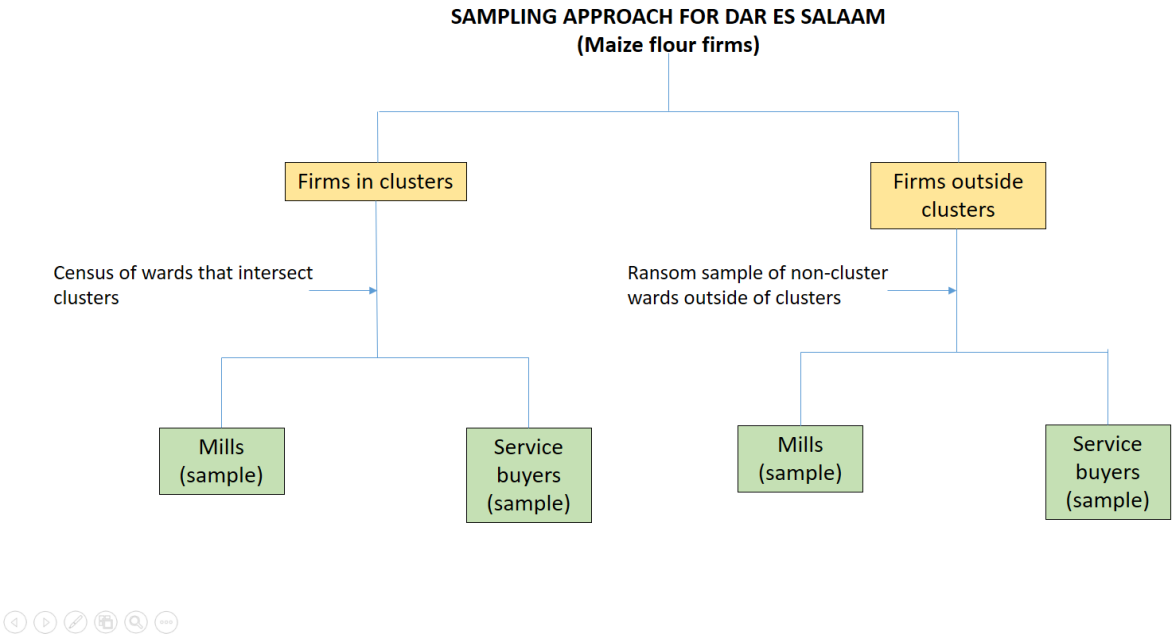

Figure 1. Diagram of multi-stage sampling strategy for maize flour firms in Dar es Salaam

[1] In total, three mills were replaced, and three mills only milled for consumers (contrary to the information that was listed) and were dropped from the sample.
[2] We wanted to know the probability of simply finding them on that day at that mill (sometimes they just come hang around). We asked these firms how many days in the month that they can be found at that mill (again not necessarily milling), in order to estimate the final population size of service buying firms. See the appendix for more information about weighting.

Weighting

The ward/district level weights (*DWgt*) were calculated as the inverse of the probability that each (non-cluster) ward would be selected, i.e., *DWgt* = N_i/n_i where N_i is the total number of (non-cluster) wards in district *i and* n_i is the total number of (non-cluster) wards selected within that district. In the Ilala district, we replaced three sampled wards that did not contain any millers in order to have sufficient representation from that district. However, those wards (with zero millers) were still accounted for in the final weighting (i.e. n_i was set equal to nine instead of six).

The within-ward weights (*WWgt*) were calculated as the inverse of the probability of a mill being selected within one of the wards, i.e., *WWgt* = B_j/b_j where B_j is the total number of mills in ward *j* and b_j is the number of sampled businesses in that ward. Note that in the wards where a full census was taken the weight equals one.

The final sampling weight (*OMM_SW*) for mills was calculated as the multiplication of ward/district level weights and the within ward weights: *OMM_SW = DWgt*WWgt.*

Also, in the second stage, we sampled service buyers (SBs) (i.e., firms that don't mill themselves and instead only purchase milling services). First, we identified the mills that operate milling machinery (OMMs) from our listing that claimed to provide milling services for other flour firms. Then, we sampled a maximum of three mills per ward in the wards where at least one OMM met this criterion. This produced a list of 30 mills. Next, we randomly assigned a day of the week (excluding Sunday) to each of these firms so that five mills were assigned to each day of the week, and also randomly chose replacements for each day of the week. Each of these mills[3] was visited for an entire day by an enumerator, who listed any identified firms that generally purchased milling services from that mill (not necessarily milling on that day). We found a total of 117 firms through this method and were able to arrange and conduct interviews with 91 of them. These comprised the second stratum in the second stage.

We calculated three additional weights for SBs. The first weight (*EWgt)* accounts for the sample of mills chosen to conduct daylong visits (i.e., *EWgt* = E_j/e_j where E_j is the number of mills in ward *j* that provide milling services to businesses, and e_j is the sample of those mills within that ward).

[3] In total, three mills were replaced and three mills only milled for consumers (contrary to the information that was listed) and were dropped from the sample.

The second weight (*NWgt*) accounts for the listed businesses that each mill provides services to that could not be interviewed (i.e., *NWgt =* N_m/n_m where N_m is the number of SBs that we found at the mill *m* and n_m is the number of SBs that were able to interview at mill *m).*

The third weight (BWgt) accounts for the fact that we only listed businesses found at that mill for one day. There might be variation over the days of the week, which we addressed by randomly assigning the different day-long visits to each day of the week. This assumed that on average the variation would average out across all of the businesses, and that there was no systematic weekly variation in that month.

The formula for this was *BWgt =* D_k/DPM_b,where D_k is the self-reported number of days per month that SBs are at mill *k* (again not necessarily milling on that day) and DPM_b is the estimated number of possible days in the month. If they reported working on Sunday, then this equaled 28, roughly assuming an average working week of 7 days and 4 weeks per month, and if they did not report working on Sunday, this equaled 24. Hence, the weight is the estimated inverse probability that they would have been found at that mill on any given day in the month, thus estimating the other SBs that might have been found on other days.

Note that there are some businesses that reported going to multiple mills. We did not explicitly ask about other mills they visit or include that information in our weighting. We instead reasoned that there was a non-zero probability of them being sampled at one of these other mills on a different day. If they had been sampled at one of these other mills, they would have reported the days that they spend there (as opposed to the original mill), and so these days (not at the original mill) would have also been accounted for in the weighting. This is relying on the simplifying assumptions that the business will only go to one mill per day (i.e., it is mutually exclusive) and that they will visit all their regular mills within the given week that we sampled.

The final sampling weight for SBs (*PMM_SW*) was calculated as the multiplication of the ward/district level weights from the first stage, with the three sets of weights from the second stage: *PMM_SW = DWgt*EWgt*NWgt*BWgt.* Note the sample of mills providing milling services to businesses was sampled independently from the sample of mills that were interviewed.

The final sampling weight (that should be used for analysis) is SWgt.

Stratification and FPC variables

The primary-stage stratification categories are milling clusters and the three districts if they were not in a milling clustering, which we delineated by the *Strata_Ward* variable. The primary sampling units (PSUs) are the wards themselves (*ward*), and the finite population correction (fpc) variable for the first stage is called *FPC1*.

The secondary-stage stratification categories are mills and non-mills (service buyers), which we delineated by the *Strata_Bus* variable. The secondary sampling units (SSUs) are the firm codes (*code*), and the fpc variable for the second stage is called *FPC2*.

Note that for application of stage level weights, the first stage weight is DWgt, and the 2nd stage weight was calculated as *codeWgt* = SWgt/DWgt. Note that *codeWgt* is not the final overall sample weight.

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

svyset ward [pweight=SWgt], strata(Strata_Ward) fpc(FPC1) || code, strata(Strata_Bus) fpc(FPC2) singleunit(centered)