Policy Research Working Paper 9905

Surveying Informal Businesses

Methodology and Applications

Gemechu Aga
David Francis
Filip Jolevski
Jorge Rodriguez Meza
Joshua Seth Wimpey

WORLD BANK GROUP
Development Economics
Global Indicators Group
January 2022

Policy Research Working Paper 9905

Abstract

Informal business activity is ubiquitous around the world, but it is nearly always uncaptured by administrative data, registries, or commercial sources. For this reason, there are rarely adequate sampling frames available for survey implementers wishing to measure the activity and characteristics of the sector. This paper presents a methodology to generate a representative sample of informal businesses using an adaptive, geographically based method called Adaptive Cluster Sampling. Developed for populations that are clustered and/or rare, this method helps with efficiently sampling Primary Sampling Units—blocks—that are fully enumerated, and from which Secondary Sampling Units—businesses—can be randomly sampled to conduct interviews. The paper shows how this methodology can be applied to surveying informal businesses, often reducing both the average variance of population estimates and fieldwork effort. Practical considerations and guidance for implementation and analysis are also provided.

This paper is a product of the Global Indicators Group, Development Economics. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://www.worldbank.org/prwp. The authors may be contacted at dfrancis@worldbank.org.

Surveying Informal Businesses: Methodology and Applications[1]

Gemechu Aga, David Francis, Filip Jolevski, Jorge Rodriguez Meza, Joshua Seth Wimpey[2]

JEL Codes: C83, O17, L22, R12
Keywords: Adaptive Cluster Sampling, Informality, Informal Businesses, Surveys

[1] The authors would like to thank Norman Loayza for the guidance and constructive comments, and the participants of the seminar held by the World Bank Development Economic Indicator Group, and the participants at the Sixth International Conference on Establishment Statistics (ICES-VI). All remaining errors are those of the authors. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent.
[2] The authors are with the Enterprise Analysis Unit, World Bank, Washington DC., emails: gayanaaga@worldbank.org; dfrancis@worldbank.org (corresponding author); fjolevski@worldbank.org; jrodriguezmeza@worldbank.org; and jwimpey@worldbank.org.

1. Introduction

In much of the world, a great deal of economic activity is informal (Schneider and Hassan, 2016; Loayza, 2016; Elgin et al., 2021). Estimates show that in the typical developing economy about 70 percent of employment is in the informal sector, though this share of labor only makes up about 30 percent of production (Loayza, 2018). This has led economists and policy makers to have broader concerns that such labor is mis-allocated by being informal, dampening overall productivity (Meghir et al., 2015; Ulyssea, 2018). While recent macro estimates point to a decline of the informal sector as a share of GDP, this downward movement is minor, and informality by and large remains persistently high, particularly in developing economies (Ohnsorge and Shu, 2021).

Informality as a whole has received a lot of attention, but recently a rising emphasis has been placed on one portion of the sector: businesses that operate informally. These businesses are the most analogous to their formal counterparts, yet they also are indicative of important distortions in those economies. Businesses that could formalize and be competitive are excluded from an economy's tax base and may lack opportunities to scale the size of their business; those informal businesses that could not survive after formalizing may employ labor that could be allocated more efficiently elsewhere, and in ways that provide more support (through regular work, income, and benefits) to those workers (Ulyssea, 2018). However, data on most such informal businesses is lacking. The limited data that is available has been used to examine the constraints to business formalization (De Mel et al., 2013; Campos et al., 2015; Benhassine et al., 2018), the characteristics of the informal businesses (Amin and Islam, 2015; Islam, 2019), and the interlinkages of the informal and formal sectors (Benjamin and Mbaye, 2012; Jolevski and Aga, 2019; Amin and Okun, 2020; Amin, 2021).[3] The size and persistence of the informal sector, and the wealth of research studying informal business activity reflects the need for high-quality, representative firm-level data collected in an efficient manner.

Businesses operating informally are frequently small and less productive, but widely abundant (see La Porta and Shleifer, 2014), implying that the typical (modal) business often does not appear in official registrar records, tax rolls, or most often, both. This very nature of informality

[3] Several of these analyses rely on data reported by formal firms on informal firms' activity.

implies that estimates of the population of informal businesses and their activity will be missing or incomplete; and sampling frames for developing high-quality, representative survey data will be unavailable. In turn, descriptions of the characteristics of many economies will be lacking information on a sector that is a ubiquitous source of jobs and business activity.

This paper lays out a methodology for conducting an enumeration exercise that generates a representative sample of informal businesses for a subsequent survey; it does so by utilizing an established geographic sampling method, Adaptive Cluster Sampling (ACS). ACS assumes that informal business activity is relatively clustered in certain areas and, in turn, it takes advantage of the information from the discovery of some informal businesses to alter the probability of selection of primary sampling units (PSUs) to capture concentrations of additional informal businesses more efficiently. ACS accounts for this altered probability of PSU selection and is able to provide estimates of the population totals of informal businesses, allowing any subsequent survey of those businesses to be considered representative. In the process, ACS frequently allows for reductions in fieldwork effort without an increase (and often a decrease) in the variance of these population estimates, if they were done repeatedly—a point which is later illustrated in this paper through simulation. The methodology is designed to be implemented in diverse situations, across different countries, while maintaining comparability.

Informal businesses may operate completely outside all or some legal channels (without a tax or business registration or license), or they may include businesses that operate informally by non-compliant behavior, such as employing workers 'off-the-books' (Ulyssea, 2018). Operationalizing data collection first requires a clear and functional definition of the underlying population of interest. A useful and basic typology of informality is the following: informality includes partial or full non-compliance with registration and licensing (legal informality), on employment (labor informality), or tax non- or under-payment (fiscal informality) (World Bank, 2020). This paper focuses on businesses' activities that are *legally* informal: that is, lacking one or all of the registration or licensing requirements to operate. This definition must be anchored to the local context of implementation but generally consists of those businesses that lack a business license, do not exist on a business registry, and/or are not registered with the relevant tax authority.

In addition to the operational aspect of this definition, there are important policy implications for surveying unregistered businesses. First, such businesses are prevalent and may

draw resources from and compete with the formalized business sector (this has been categorized as the 'parasitic view')[4]; by contrast, a competing view is that widespread informality represents a 'segmentation' between a formal labor market, the business environment they experience, and their informal analogs (Perry et al 2007; Maloney 2007; Kanbur, 2017; Ulyssea, 2020). Understanding the estimates and nature of informal sector businesses informs policy, particularly considering the absence of consistently effective policies to address informality, e.g., the limited success of formalization efforts (see Bruhn and McKenzie, 2014; Floridi et al 2020).

Existing methodologies, such as household-based sampling, are not designed to measure informal sector businesses directly. Such methods often do not capture informal sector businesses at their point of business or operation; when they do, businesses are reached via the places of employment of household members and, consequently, they are unable to estimate business density in a geographical area. Likewise, informal sector businesses are, in all but a few cases, excluded from economic censuses, which tend to be infrequent and expensive. Informal sector businesses, virtually by definition, are also not reached by official surveys proceeding from administrative sampling frames. By contrast, the proposed ACS method reaches units at their point of operation and is designed to measure informal sector businesses *per se*; this paper builds upon existing ACS literature and shows that this method not only provides a precise, unbiased population estimate but can also provide practitioners with large fieldwork efficiency gains when compared to other methods.

However, the ACS method does bring trade-offs and challenges. As noted above, to operationalize such a survey, a clearly delimited population of inference is needed. In this case, this consists of un-registered business activity, within a delineated geographic area (e.g., a set urban area). This implies that informality in non-covered locations, rural or excluded cities, will not fall within the purview of these surveys. The approach also implies that portions of labor and fiscal informality do not fall within the coverage of the survey, an argument to include potentially sensitive topics on the informal activities of formally registered firms in traditional firm-level surveys.[5] Practitioners should note that such a methodology, unlike household- or labor-survey-linked methods, is not designed to measure household activity; as such, additional care must be

[4] See La Porta and Shleifer (2014) for a discussion.
[5] The authors, for example, are also experimenting on different ways to ask for these sensitive topics in the Enterprise Surveys of the World Bank.

taken to avoid undercounting household-based informal sector businesses. Attention must also be paid to fluctuations of informal sector movement over different areas or times of the day or week.[6]

2. Collecting Data on Informal Businesses: Existing Methods and Trade-offs

Two main data collection methods are typically used for surveying informal businesses: i) household surveys, the prototypical examples being labor force surveys (LFS) or living standards surveys[7]; and ii) establishment surveys or censuses (ILO, 2013). For surveys it is useful to consider each method as a multi-stage sampling procedure, where with some probability of selection, units are selected in a first stage, with the selected units denoted as PSUs, followed by secondary sampling units (SSUs), and so on.[8] Define all individual PSUs $i \in N$, where N is the universe of PSUs. Likewise, SSUs are defined $j \in U$. Sampling units at each stage are well-defined, so N and U may exclude certain elements.[9] Each allows for the possibility of stratification.

The household survey approach is implemented through standard household surveys, such as Labor Force Surveys, and can capture informal (i.e., "off-the-books") employment as well as employment in the informal sector (i.e., employment in legally informal activities). PSUs are enumeration areas defined to capture households. As a result, N often excludes explicitly non-residential areas, such as markets, industrial centers, or commercial zones. Call this set of PSUs N_{LFS}.[10] By design, these surveys are representative of N_{LFS}—that is the geographical areas delimited to capture households. Measures of informal businesses only require a few additional questions in these surveys, exploiting economies of scale in a cost-effective way.

However, such approaches are not necessarily well suited for representative estimates of informal businesses. The number of businesses and counts of individuals are not always mapped one-to-one; an informal business can be jointly operated across individuals, and one individual may operate several businesses. More fundamentally, N_{LFS} rarely accounts for the concentration

[6] Mobility and timing issues are not unique to ACS but plague any attempt to sample informal businesses.
[7] The most widely known of these is the World Bank's Living Standards Measurement Survey (LSMS), which is run globally, in partnership with National Statistics Offices.
[8] Note that this embeds a single-stage sampling approach, where only PSUs are selected.
[9] An example of this is where N excludes certain inaccessible areas (like a government or military installations).
[10] Note the use of subscripts 'LFS' and 'ES' to denote, by shorthand, the Labor Force Survey and Establishment Survey methods. This notation is dropped from later discussions of the proposed method (Adaptive Cluster Sampling) for simplicity.

of informal businesses in certain areas. It is not designed to capture these informal businesses in their place of operation. As such, policy-relevant information such as the characteristics of these businesses, information on sales, expenditures, investment, competition, and the interaction with the business environment is de-linked from where these activities take place (with exceptions in household-based businesses). So-called 1-2 methods, where a second module following the household module is administered to an informal business/activity, implement a business-level survey module at the place of business, avoiding this last problem. However, such surveys are still limited to the representativeness of N_{LFS}.[11]

A second approach is the establishment survey or census, which may include informal businesses. In the case of an establishment (or economic) census, all units are selected with certainty, removing a multi-stage framework. By contrast, an establishment survey requires a suitable sampling frame, where each unit in that frame exists within the universe of applicable businesses; call this U_{ES}. In certain cases, they may come from an economic census, but they are infrequent and tend to exclude legally informal businesses.[12] Accurate projections of informal business activity also require credible estimates of $U_{ES,informal}$, which, virtually by definition, are missing from censuses and administrative counts. In the absence of such adequate frames, implementers usually construct a frame through a two-stage process, where an area (PSU, $i \in N_{ES}$) is first enumerated to generate a frame, and then a selection of informal businesses are selected in a second stage.[13] However, constructing such a frame is often intensely costly in terms of time and effort, particularly if implementers lack *a priori* information on the location of informal businesses.

This paper proposes a specific solution to address the implied time, effort, and cost developing such geographically based estimates and sampling frames: namely, we propose that survey implementers use a sampling method designed to adapt and more intensively sample areas where informal businesses are discovered. To do this, we use a sampling method known as Adaptive Cluster Sampling (ACS) (see Thompson, 1990).[14] ACS is a method of essentially guided sampling, where the discovery of informal businesses ($j \in U$) in a PSU ($i \in N$) triggers more

[11] Note that in these cases informal businesses are a tertiary sampling unit (TSU), where even if those businesses are selected with certainty, unit non-response will result in three-stage weighting.

[12] India's Economic Census and Rwanda's Establishment Census are two notable exceptions that include informal businesses.

[13] An example of this approach is the survey of informal businesses in townships in South Africa conducted by Sustainable Livelihood Foundation (http://livelihoods.org.za/).

[14] The use of ACS to measure informal business activity has been piloted by the World Bank's Enterprise Analysis Unit (DECEA) for several years, following an initial pilot of the method in Harare, Zimbabwe (2016/17).

intensive sampling in adjacent PSUs (Salehi and Seber, 1997). PSUs are selected randomly first, with or without stratification, and—based on a trigger threshold—adjacent PSUs are then enumerated. As such, the process is no less naïve in the initial selection of PSUs than other sampling approaches, but ACS exploits the information revealed in the process to efficiently target the population of interest when that population is clustered and/or rare (Thompson and Seber, 1996). The sampling process to survey informal businesses follows a two-stage method. In the first stage, ACS provides a probabilistic sample of PSUs which are fully enumerated to list all informal businesses. In the second stage, businesses are randomly selected from the list for an in-depth interview, discussed more in detail in Section 5.

3. Technical Foundations of Adaptive Cluster Sampling[15]

ACS begins with the definition of PSUs, $i \in N$. For the sake of generalization, let N be a well-defined geographic area, divided into a grid of equal-sized squares. For this reason, PSUs will also be described in this application of ACS as block areas, (henceforth BAs).[16] The first step draws an initial sample of BAs of size n; all informal businesses are fully enumerated in these BAs. Implementers set some condition C, which is an expansion threshold: if a BA contains sufficient units of the target population ($\sum j_{i \in n} \geq C$), additional BAs adjacent to all squares where the count of units that meets or exceeds that threshold are sampled. Implementers can set specific rules for this expansion, with the most straightforward being all eight adjacent neighbors. If these newly sampled BAs meet condition C, they trigger additional sampling of adjacent squares, and so on.

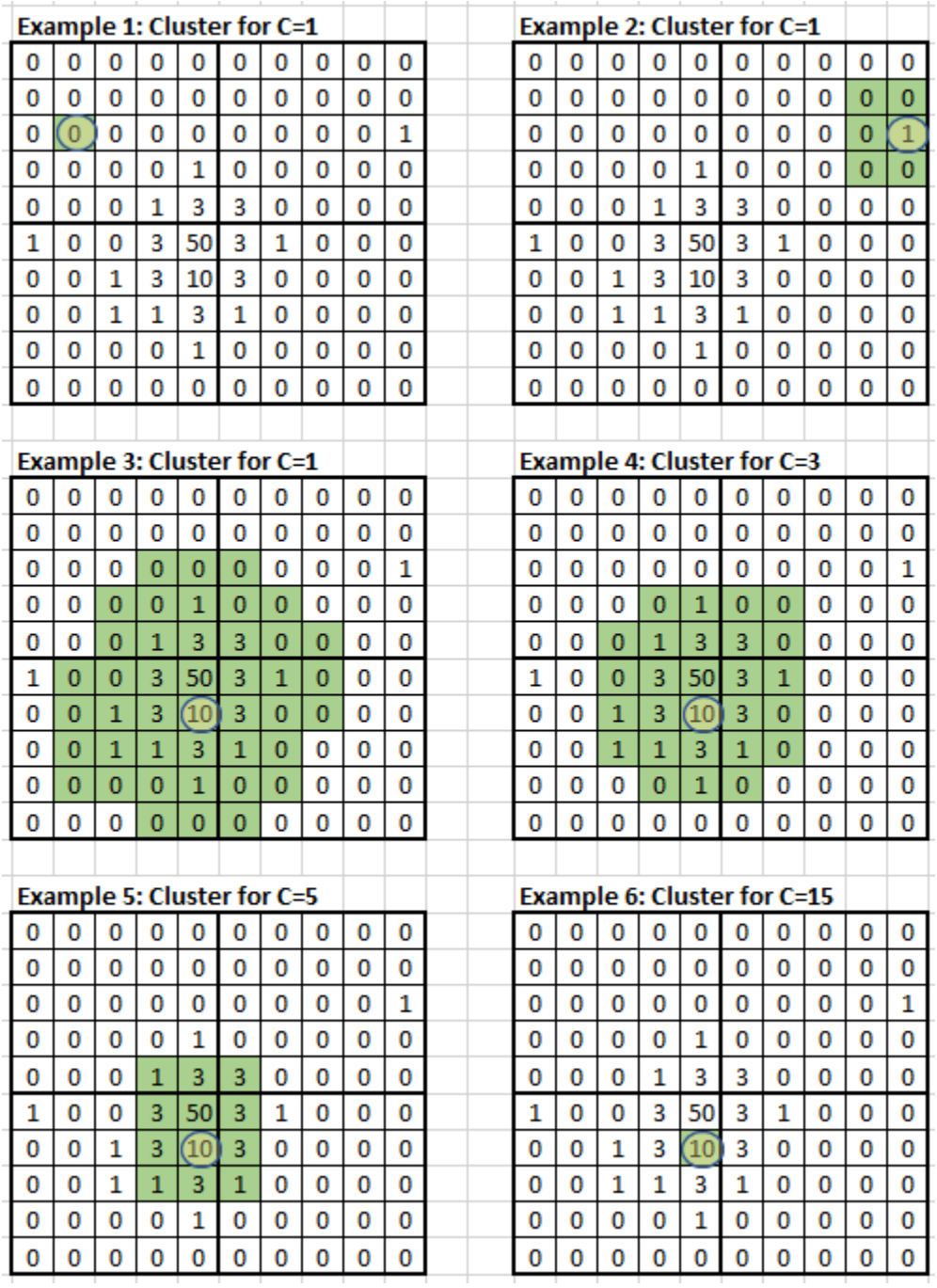
Define a ***neighborhood*** as a collection of BAs that includes both an initially selected BA, and in the case that it satisfies C, an expansion into additional, adjacent BAs. In this case, we use the pre-set rule of expanding into all eight, adjacent BAs; additional neighborhood expansions are shown in Appendix Figure A.1. A ***cluster*** is the collection of all connected neighborhoods that are enumerated because of the initial selection of BA i. Figure 1 below shows clusters highlighted in

[15] Note that the emphasis in this section is in the first stage of sampling informal businesses where BAs are selected using ACS and then fully enumerated. The second stage is just a simple random selection from the resulting enumeration list, which can be implemented via any number of random selection algorithms.

[16] This term is also to differentiate from city blocks or public squares, with a ready comparison of 'enumeration area.'

green, where the initially selected BA i is identified by a circle. The number in each square denotes $\sum j_{i \in n}$ within that BA. The figure shows that changing the condition C can change the extent of a cluster.[17]

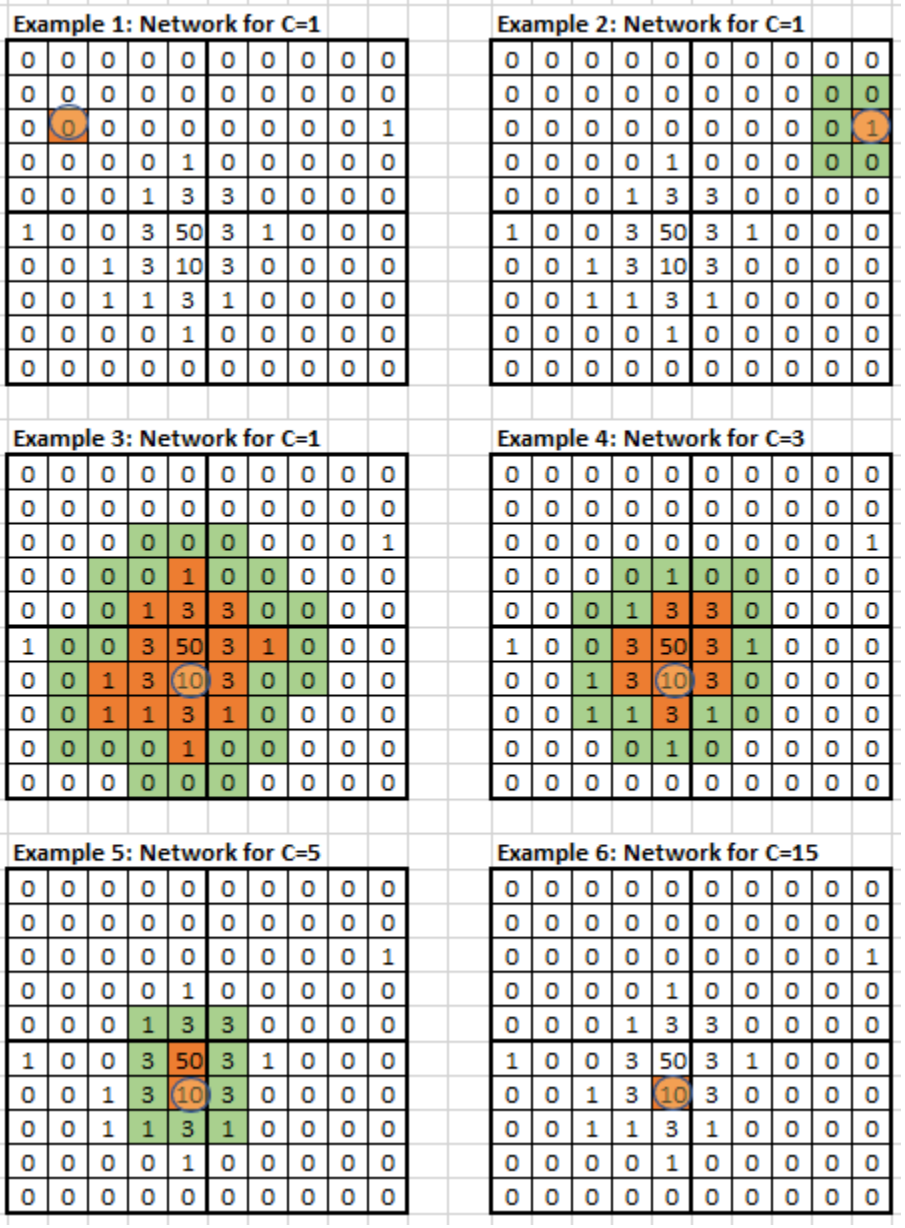
Figure 1: Examples of Clusters under Different Expansion Thresholds (C)



[17] While these examples show a single square included in the initial sample, when multiple squares are included in the original sample, it is possible for clusters to overlap partially or be fully duplicative.

Within a cluster there is a subcollection of BAs, termed a ***network***, with the property that selection of any square within the network would lead to inclusion in the sample of every other square in the network. By definition, then, a network consists of adjacent BAs that each meet the condition C. The examples in Figure 2 below highlight the networks in orange with the equal-sized or broader cluster highlighted in green. Note that in examples 1 and 6, the cluster and network are identical. An ***edge unit*** is one not satisfying C, but in an enumerated neighborhood with at least one BA meeting C. Green squares in Figure 1.b are all edge units. Note that in the case of example 1, where no expansion is triggered, the initial BA is a network size=1 with no edge units.

Figure 2: Examples of Networks under Different Expansion Thresholds (C)



The traditional Horvitz-Thompson (1952) (HT) estimator of the population's total and mean of the number of BAs are unbiased with ACS, when the HT estimator is modified to account for the specific BA selection probabilities that arise from adaptive expansion (Thompson, 1990).[18]

[18] Multiple unbiased population estimators are available using ACS. These include the HT estimator as well as the Hurwitz-Hansen (HH) estimator, which is also modified for ACS sample inclusion probabilities. Rao-Blackwell modifications of these estimators have been suggested by Thompson and Seber (1996), Salehi (1999) and Dryver and Thompson (2005). Simulation and practical application work done by others has generally shown that the modified

In ACS a BA can be selected either initially or as part of an adaptive expansion, where selection through expansion is a function of units (j) in all adjacent BAs.[19] This presents a challenge: we have information on all BAs in a neighborhood, but we do not have information on the count of units (j) in adjacent, unenumerated BAs (i.e., the white squares in Figure 1). As a result, one cannot know if a BA could have been selected through an alternative expansion pattern, via unenumerated BAs. This is an important difference with simple random sampling (SRS), where selection probabilities are known. With ACS, it is not possible to compute the probability of inclusion in the sample for all sampled BAs as not enough information is discovered for all sampled BAs to produce these calculations. However, the subset of BAs that make up the networks within a sample can have their inclusion probabilities computed. In other words, inclusion probabilities can be computed for the discovered cluster, minus any edge units.

A specific selection probability is required for the ACS HT estimator. Define the probability that a given BA, i, is included in the sample as π_{i}, where initial BAs are selected using SRS without replacement. Formally π_{i} is expressed as:

$$\pi_{\mathrm{i}} = 1 - \frac{\binom{N-m_i-a_i}{n_1}}{\binom{N}{n_1}} \qquad (Eq.1)$$

where N is the total number of BAs in the study area; n_1 is the number of initially selected BAs before any expansion takes place; m_i is the number of squares in the network to which i belongs; and a_i denotes the total number of squares found in networks to which square i is an edge unit. This probability can be used to write an HT estimator of the number of BAs for ACS as:

$$\hat{\mu}_{HT} = \frac{1}{N}\sum_{k=1}^{K}\frac{y_k^* I_k}{\pi_k} \qquad (Eq.2)$$

Where y_k^* is an aggregated indicator of y_i values in network k, where $i \in k \in K$. For example, for a population estimate, y_k^* is the sum of all informal businesses (j_i) in BA i, that form part of network (k), that is: $y_k^* = \sum_{i=1}^{I_k}\sum_{j=1} j_i$.[20] $I_k = 1$ with probability π_{i} if the sample intersects

HT estimator for ACS is more efficient than the HH estimators even when Rao-Blackwell adjustments are made (Tout 2009, Christman 1996, 1997, Thompson 1990, Salehi 2003). For that reason, the HT estimator is presented here.

[19] Technically of course, it is all adjacent squares that conform to the chosen expansion pattern defined for the neighborhood. See Appendix A.1 for some examples.

[20] Note for completeness I_k is the total number of BAs in network k.

network k and $I_k = 0$ otherwise. Following Thompson (1990), the population and unbiased, estimated variance of $\hat{\mu}_{HT}$ are shown in Appendix A.2.

4. ACS and Simple Random Sampling

ACS generally provides gains reducing what will be called here 'fieldwork effort', the number of blocks to be enumerated, particularly, when the underlying population of interest is rare and/or clustered (Borkowski and Turk, 2013).[21] ACS exploits data where concentrations of units of interest are in adjacent BAs (clustering) and where networks of these concentrations occur rarely: as such ACS can identify areas of notably high concentrations of these populations (Thompson and Seber, 1994). The definitions of relative clustering and rarity are specific to particular populations of interest; ACS has been shown to reduce fieldwork effort across a variety of fields, including rare populations of interest for the natural sciences.[22] The application of ACS to informal businesses presented in this paper is novel, however, and so the following sections discuss practical details for implementers interested in this application.

ACS implementers face a variety of design decisions that include the choice of expansion threshold C, the delineation of the total geographic area N, the size of BAs, the pre-set rule defining neighborhoods, and if stratification will be utilized. What follows is a simulation of different design choices and parameters to see how these choices perform against baselines based on SRS of BAs. The simulations use a unique data set from Eswatini (at the time in 2014, Swaziland) that provides a full census of all businesses operating in the country, including informal businesses. Uniquely, the data set includes GPS coordinates, and so allows for a real-life estimate of the

[21] Fieldwork effort has also been referred to as 'sampling effort'.

[22] ACS has been used in a variety of fields, including ecology, ethology, and environmental studies. More specifically, ACS techniques have been used to estimate the density of freshwater mussels (Smith et al., 2003) sea urchins (Skibo et al., 2008) black sea bass (Cullen et al. 2017), yellow perch (Yu et al., 2011) and thresher sharks landings along the California coast (Hariharan et al., 2013); to monitor the populations of endangered species (Davis et al, 2011, Salehi et al., 2015); to determine infestation extent and spread rates of forest insect pest (Coggins et al., 2010); to identify disease levels in plants (Gattone et al., 2013) as well as their restoration levels (Bried, 2013), or simply the abundance levels of specific plant species (Philippi, 2005, Bowering et al., 2017). The conclusions of most of the studies above suggest that ACS tends to overperform other sampling techniques when populations are rare and clustered.

clustering of informal-sector businesses.[23] As noted, these comparisons are drawn relative to SRS (which readers will note is equivalent to ACS with $C = \infty$). All simulations set BA size to 150 by 150 meters[24] and use a neighborhood definition of eight adjacent BAs; they vary by changing expansion thresholds (C) and the initial selection number of BAs, n_i. Later simulations build on these by also varying geographic scope, density, and by adding stratification.

The simulations focus on the largest metropolitan area in Eswatini—Manzini and the peri-urban neighboring city of Matsapha. Together, the area covers approximately 120 square kilometers and has a daytime population of over 150,000. An overlay of BAs yields 5,616 BAs of 150 meters by 150 meters, as shown in Figure 3. These squares contain 4,471 establishments, the true value of the population total. The population appears clustered and rare, with only about 13% of squares containing any establishment at all. Simulations were run using Stata, and every simulation contains 10,000 initial draws of BAs. For comparison, two sets of simulations were run: one set for SRS and one for ACS. SRS simulations include samples (n_1) of 2%, 3%, 5%, 10%, 20% and 30% of the 5,616 BAs. ACS simulations include samples where the initial sample of BAs, before any expansions, cover 2%, 3%, 5%, 10%, and 20% of the 5,616 squares and examine the following possible values for $C = \{1, 2, 5, 10\}$.

[23] The simulation results presented here incorporate all 4,471 businesses, 1,901 formal and 2,570 informal businesses. Simulations run using only the informal businesses show similar clustering and yield similar results and efficiency gains.

[24] The choice of a 150-meter-squared BA is as much practical as one driven by data: in an urban environment that can include large clusters of informality, interviewers must be able to enumerate a given BA in a reasonable amount of time. As described in later sections, the 150-meter-squared BA is an experience-based size that makes this fieldwork feasible.

Figure 3: Population of Businesses in Eswatini, with Grid Overlay

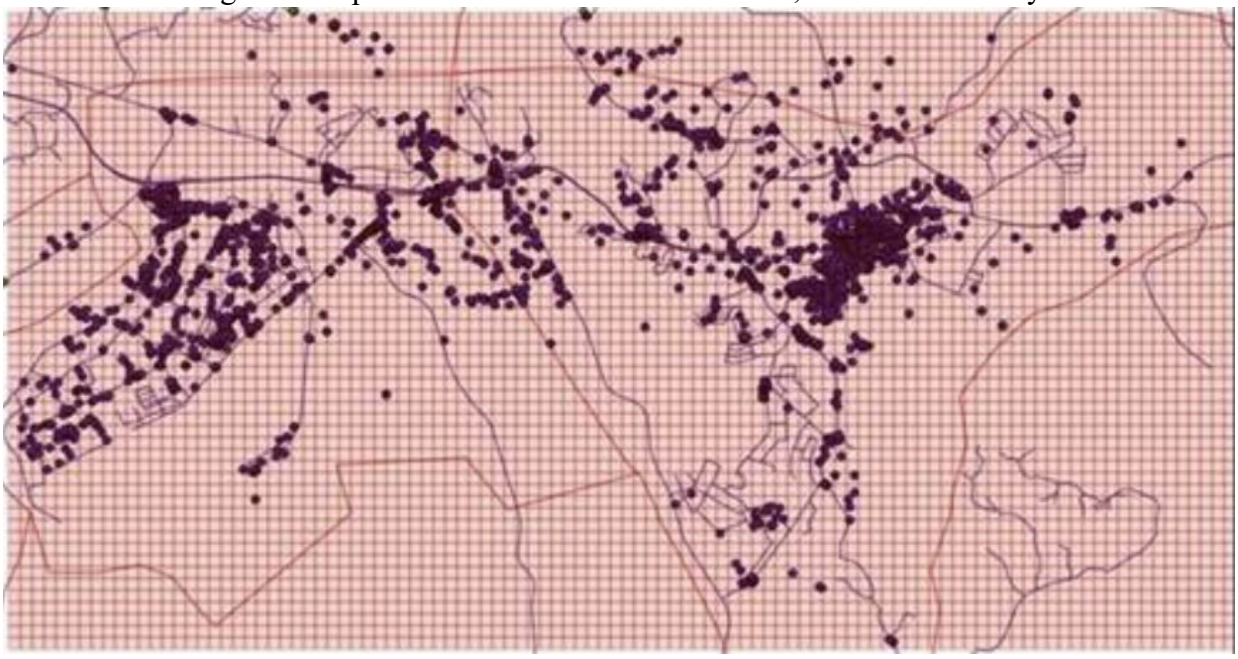


Table 1 shows the results of these simulations. The top panel of the table presents ACS results and comparable SRS simulation results are presented at the bottom (shaded portion). Column (1) provides the four expansion thresholds (C) used to simulate results with six different initial samples (n) presented as a percentage of the total, column (2) and in terms of raw numbers of BAs, column (3). For ACS, the total BAs will be greater than n, provided that at least one expansion occurs; as a result, the exact number of total BAs for ACS cannot be known with certainty. For SRS, all BAs sampled will always equal n (as shown at the bottom section for SRS results). For ACS the mean number of enumerated BAs, after 10,000 simulations, is presented in column (4). Similarly, columns (5), (6), and (7) present the simulation means of the population estimate —the total numbers of businesses, the mean of the standard error of this estimate, and the mean bias of the estimate, that is the difference between population-estimate and the true value, 4,471.[25] The last three columns of the table show the relative effort, that is the total number of BAs enumerated in ACS relative to set SRS baselines, namely samples of 10% (562 BAs), 20% (1,123 BAs), and 30% (1,685 BAs) of the total number of BAs, columns (8), (9) and (10). A relative effort

[25] As explained in footnote 13, we are ignoring the randomization procedure that may be used to select businesses within the enumerated blocks to simplify presentation as there is nothing innovative in this second stage.

below 100% indicates a fieldwork effort reduction of ACS relative to SRS; above 100% indicates an increase in fieldwork effort.

Both ACS and SRS produce unbiased population estimates; only SRS, n=2% has a mean bias with an absolute value greater than 1%. What is more, ACS estimates are consistently less variant compared to SRS. It is useful to compare rows with similar values for 'Total BAs'. The last two rows of the table show that SRS samples of 20% and 30%, with a range of mean enumerated BAs from 1,123 to 1,685, mean S.E. of 1,189 and 1,021, respectively. Illustrative of the gains from ACS, all ACS rows with a mean "Total BA" count in that range, indicated by bold italics, have demonstrably lower mean standard errors. Likewise, the relative effort gains of ACS are apparent in the last three columns. As expansion thresholds increase, the measures of relative effort decrease. The intuition being that with higher thresholds of C, a triggered expansion is more likely to efficiently discover clusters of informal business activity (i.e., there are fewer empty squares enumerated).

Table 1: Simulated Results of ACS Parameters compared to SRS

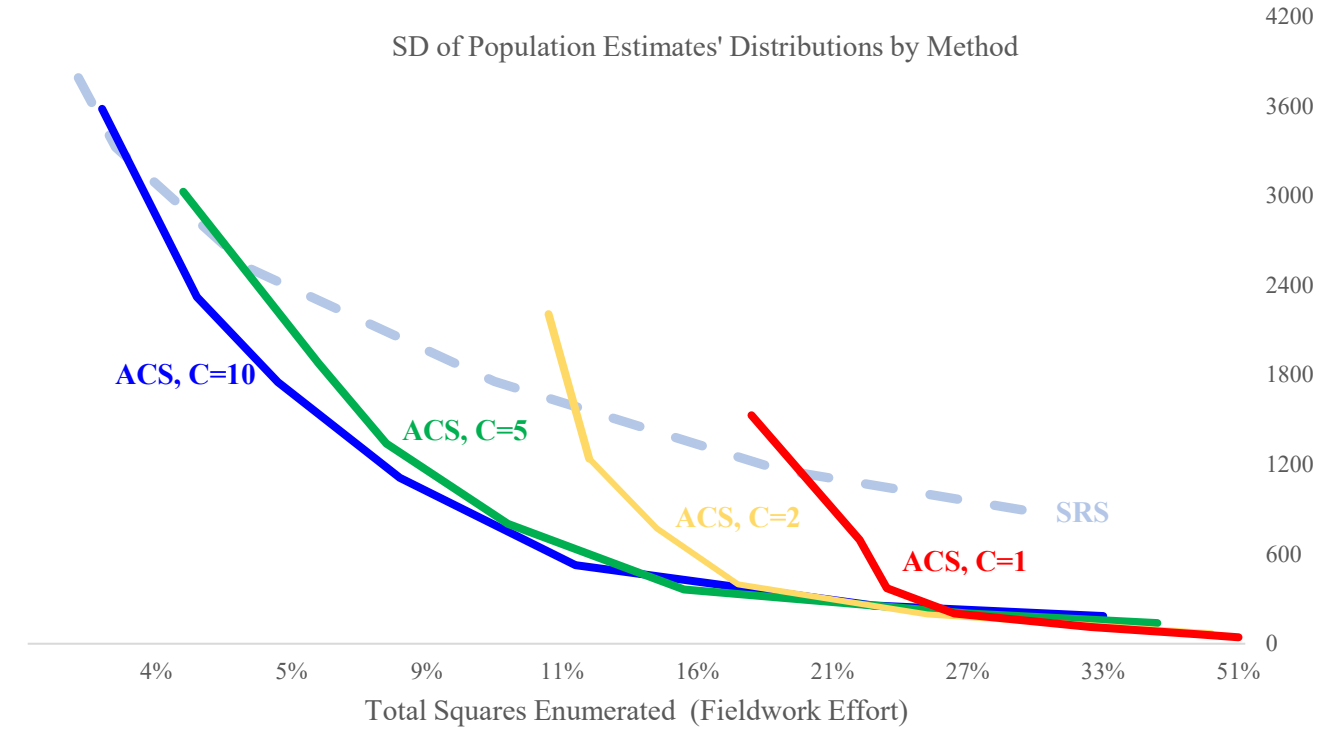
		Initial BAs (n_1)		Mean of 10,000 Simulations				Relative Effort ACS to SRS		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	C	%	#	Total BAs	Mean Pop. Estimate ($\hat{U}_{avg.}$)	Mean S.E.	Bias	SRS 10%	SRS 20%	SRS 30%
ACS	1	2	112	1,047	4,477	593	0.14%	186%	**93%**	**62%**
		3	168	1,209	4,474	580	0.07%	215%	108%	**72%**
		5	281	1,404	4,468	573	-0.07%	250%	125%	**83%**
		10	562	1,758	4,472	572	0.02%	313%	156%	104%
		20	1,123	2,314	4,473	574	0.04%	412%	206%	137%
		30	1,685	2,804	4,475	576	0.09%	499%	250%	166%
	2	2	112	580	4,471	648	-0.01%	103%	**52%**	**34%**
		3	168	721	4,469	612	-0.05%	128%	**64%**	**43%**
		5	281	917	4,471	586	0.01%	163%	**82%**	**54%**
		10	562	1,300	4,466	577	-0.11%	232%	116%	**77%**
		20	1,123	1,911	4,457	576	-0.31%	340%	170%	113%
		30	1,685	2,443	4,458	577	-0.30%	435%	218%	145%
	5	2	112	233	4,508	831	0.83%	**42%**	**21%**	**14%**
		3	168	328	4,447	742	-0.54%	**58%**	**29%**	**19%**
		5	281	490	4,472	661	0.03%	**87%**	**44%**	**29%**
		10	562	844	4,461	611	-0.23%	150%	**75%**	**50%**
		20	1,123	1,470	4,459	591	-0.26%	262%	131%	**87%**
		30	1,685	2,045	4,454	586	-0.39%	364%	182%	121%
	10	2	112	166	4,494	1,009	0.52%	**30%**	**15%**	**10%**
		3	168	240	4,460	884	-0.24%	**43%**	**21%**	**14%**
		5	281	375	4,491	753	0.44%	**67%**	**33%**	**22%**
		10	562	683	4,475	654	0.08%	122%	**61%**	**41%**
		20	1,123	1,265	4,473	611	0.04%	225%	113%	**75%**
		30	1,685	1,832	4,472	597	0.01%	326%	163%	109%
SRS		2	112		4,420	2,486	-1.14%			
		3	168		4,455	2,238	-0.37%			
		5	281		4,432	1,901	-0.87%			
		10	562		4,475	1,530	0.09%			
		20	1,123		4,488	1,188	0.37%			
		30	1,685		4,492	1,004	0.46%			

Note: Authors' calculations based on 10,000 simulation runs for each row's parameters. The true population value is U=4,471. Bias (Col. 7) is $\frac{\hat{U}_{avg.}-U}{U}$*, where* $\hat{U}_{avg.}$ *is the average (mean) population estimate of the total number of informal businesses, averaged over 10,000 simulations. The relative effort columns are based on the total enumerated BAs compared to the SRS samples of 10%, 20%, and 30% respectively. In the last three columns, figures in bold and blue indicate a lower relative fieldwork effort of ACS vs. SRS.*

Since the fieldwork effort of ACS cannot be known ahead of time, it is helpful to compare different sampling methods with approximately the same values of (mean) effort. Consider Figure 4. The figure shows the mean standard deviation of population estimates from 10,000

simulations on the y-axis; the x-axis shows the mean amount of the fieldwork effort, that is the percentage of BAs that are enumerated. The sampling methods that are shown include SRS, and ACS with expansion thresholds, $C = \{1, 2, 5, 10\}$. We leave the bias of these estimates aside, as Table 1 confirms that both ACS and SRS are unbiased. What is first clear is that, at a given level of mean standard deviation (that is, holding the value of the y-axis constant), ACS results in lower fieldwork effort by reducing the share of BAs that are enumerated. Intuitively, at lower expansion thresholds, fieldwork effort shifts rightward: as more expansions are triggered, naturally fieldwork effort increases. As C increases, fieldwork effort reduces, at generally no cost in terms of the variability of these estimates (recall that SRS is equivalent to an expansion threshold $C = \infty$). An illustration of how increasing C affects the rareness and clustering of a population is provided in Appendix A.3. Similarly, at roughly equivalent fieldwork efforts (on average)—that is, holding a value of the x-axis constant—ACS generally provides lower mean standard deviations. There is a noteworthy exception: at lower fieldwork efforts, the mean standard deviation for ACS can exceed that of SRS. This is because at lower fieldwork efforts, not enough expansions are triggered, resulting in more variable estimates of the total populations (this is further illustrated in the leftward mode in the distribution shown in Appendix A.5.) Under these conditions, then, ACS generally provides less variant population estimates on average, but does not strictly improve over simple random sampling under all conditions.

Figure 4: Comparison of Select ACS and SRS Parameters over 10,000 Simulations



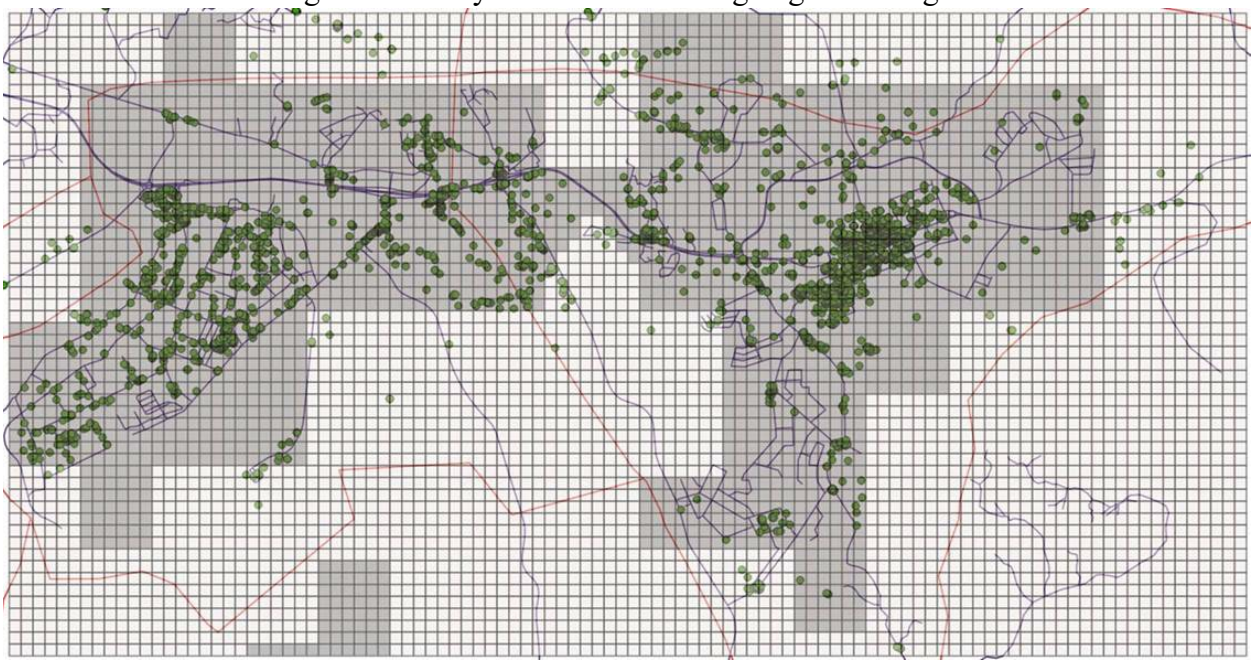
Note: Authors' calculations based on 10,000 simulation runs. Fieldwork effort (x-axis) represents the mean share of block areas enumerated over all runs; the standard deviation (SD, y-axis) is the standard deviation of all 10,000 of the population estimates for each case.

ACS and Stratified Random Sampling

In many cases additional information may be available that allows for stratifying a geographic area before sampling. With business activity (including informality), stratification variables that may be useful include land-use categorization, population density, road density, vegetative cover, nighttime lights, or any geographically mapped variable where one category is significantly more likely to intersect with business activity. In what follows, we use a nighttime lights satellite imagery to generate a stratification variable. Nighttime lights are readily available to the public and have been used widely as predictors of concentrated economic activity (Donaldson and Storeygard, 2016). Figure 5 shows the result of a binary stratification using nighttime lights overlaying the establishment locations, with higher concentrations shaded. Despite the relatively low resolution of the lights data at 500m resolution the strata are very

effective at discerning economic activity with 98% of establishments laying within the highest probability strata which also contains 43% of all squares.

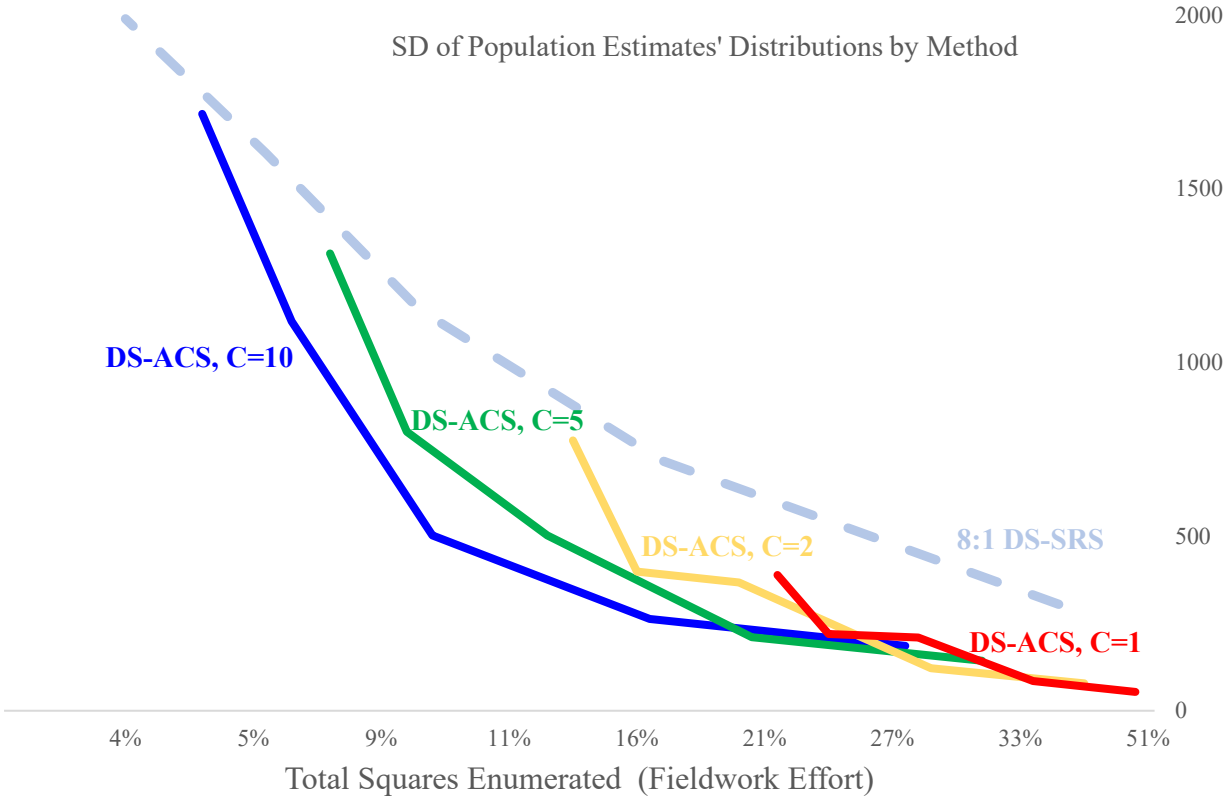
Figure 5: Binary Stratification Using Nighttime Lights



With well-defined strata in hand, survey implementers will likely want to use that information to disproportionately design their sample to capture more informal business activity while limiting the demands on fieldwork effort. In this case, we expect implementers will want to draw disproportionately higher samples in the high-density areas, which constitute a lower number of overall BAs but with higher concentrations of businesses. Specifically, we consider a case where using SRS, an implementer draws BAs in the high-density areas relative to the low-density areas at a ratio of 8:1. For ACS, since the final distribution of the sampled BAs is not known beforehand, for comparison purposes we need to choose a starting ratio that is disproportionate, but that will become more so as there are presumably more expansions in the high-density areas.[26] For the ACS simulations, we choose a ratio of high- to low-density BAs of 5:3.

[26] We note that the final sample ratio for C=1 is approximately 9:1, for C=2 it is approximately 6:1, and 3:1 for C=5.

Figure 6: Comparison of Select DS-ACS and DS-SRS Parameters over 10,000 Simulations



Note: Authors' calculations based on 10,000 simulation runs. Fieldwork effort (x-axis) represents the mean share of block areas enumerated over all runs; the standard deviation (SD, y-axis) is the standard deviation of all 10,000 of the population estimates. DS: disproportionate, stratified indicates a comparatively high initial starting share in high-density nightlight areas (Figure 4). For Simple Random Sampling (SRS) this ratio is 8:1. For ACS, this ratio is 5:3 for initial starting squares, as the expansions in the high-density areas are expected to make the ratio more disproportionate.

Figure 6 replicates Figure 4 but with all simulated samples using the two-category stratification. As without stratification, the ACS methods (now denoted with the prefix DS for disproportionate, stratified) require far less fieldwork effort for a given standard deviation (that is, holding the y-axis value constant). Very importantly, for a given level of sufficiently high fieldwork effort, ACS methods all provide less variable population estimates. This is partially due to the fact that, with well-defined strata, the underestimations that resulted before are minimized; compare Appendix A.5 to A.6. Note that since expansions will necessarily require that more BAs are enumerated, fieldwork effort (especially at lower expansion thresholds) will trend upward. Lastly, readers should note that even ACS without stratification generally reduces the standard deviation of population estimates and reduces fieldwork burden even over SRS with stratification. We provide an illustration of this in Appendix A.7; Appendix A.8 also replicates these exercises

with synthetic data that increases population density (given that Eswatini is a generally sparsely populated country) and shows that ACS's benefits largely hold throughout.

5. Implementing ACS in practice

Practically, implementers must first clearly define delineated geographical areas—that is N. While ACS can be applied in rural areas, this paper focuses on urban settings. When defining this geographical area, three broad sources of information are considered, often in combination: (1) administrative and natural boundaries; (2) supplementary information likely to be correlated with informal business activity; and (3) local knowledge. Implementers should balance the tradeoffs of relying on any single one or particular combination of these information sources. Administrative boundaries (or delineated statistical areas like a census tract) can correspond to official statistics, important for policy-relevant comparisons. However, administrative boundaries may be insufficient. This will be the case when political boundaries are outdated, or when informality occurs in the outskirts of administrative areas. Implementers will likely need to consult alternative sources of information—such as land-use or, even, night light intensity—as well as local knowledge.

The grid of block areas, BAs, can be easily generated using GIS[27] software, such as ArcGIS or QGIS. BAs should be small enough that they can be enumerated in a reasonable timeframe, but large enough to capture concentrations of informal businesses. Since 2016, the World Bank's Enterprise Analysis Unit has been piloting an ACS approach in a survey of informal businesses, known as the Informal Sector Business Survey (ISBS). From several initial rounds of the implementation of the ISBS, a practical suggestion is to begin with BAs sized at 150 by 150 meters. However, implementers will want to consider their own restrictions, such as budget and staffing, the relative density of informal businesses, and the appropriateness of other design decisions, such as the choice of neighborhood definitions. All of these choices will be interlinked. Stratification will require mutually exclusive categories and in the case of the ISBS, six stratification categories have proven quite useful: residential, commercial/industrial, mixed residential and commercial/industrial, market areas, open areas, and inaccessible.

[27] Geographic Information Systems.

Informal businesses or establishments ($j \in U$), as defined earlier, are those that are *legally* informal, that is they lack full or partial formal registration. The definition of legal informality must be uniquely tailored to each country based on relevant legislation and regulation for formal registration. In the case of the ISBS, an interview is administered to eligible informal business activities randomly selected from the list of informal businesses discovered as part of the enumeration exercise (j). Combining ACS and a random selection of businesses within the enumerated blocks is not only an appropriate method of providing unbiased population estimates, with comparatively low variance and fieldwork efficiency, but also a means—as in the case of the ISBS—of providing a framework for more detailed information, while preserving representativeness.

Implementers will also want to establish monitoring and quality control procedures for both data and fieldwork effort. Since the latter concern is the main focus of this paper, a few key details merit attention. Specifically, the integration of GPS-linked software is often important to ensure enumerators sufficiently apply the listing exercise to all potential units in selected BAs. Sufficiently doing so requires full coverage of paths without enumeration outside of a delineated BA; path tracking software can be used for this purpose.[28] Implementers may also want to integrate fieldwork quality controls to minimize known potential pitfalls, such as undercounting of household-based activities.

Implementing this methodology also requires close monitoring of fieldwork efforts, since under ACS, the total number of BAs to be enumerated is not known *a priori*. Practically, it may be recommended to enumerate the initial random sample of BAs to calibrate the value selected for the expansion threshold (C). Based on simulation exercises such as those presented in Section 4, it is recommended to initially select a somewhat high value of C, as a more conservative approach. That section showed that along the two dimensions of fieldwork effort and variance of population estimates, ACS is widely more efficient on both dimensions, holding the other constant: i.e., if one holds the level of variance constant, ACS is almost always more efficient in terms of fieldwork effort. However, practically, a survey implementer may want to have control over the expansions that occur through ACS, as they trigger additional fieldwork—which can be costly. That is, an implementer can start with a higher expansion threshold and then evaluate if the triggered expansions are within the needs and limits of the survey.

[28] For example, the ISBS is integrated with the open source OruxMaps.

While implementers will be interested in the population counts and characteristics available from the enumeration exercise, they will also want to use the much more extensive information collected through in-depth interviews with the randomly selected subset of informal businesses, which captures more detailed information on informal businesses' operations. In the case of the ISBS, second-stage selection for an interview happens in real-time, during the enumeration of BAs. This is made possible using readily available CAPI software.[29]

As explained above, the first-stage enumeration of BAs (i) is unbiased even when unequal probabilities of selection are considered. This requires the calculation of first-stage weights implied by the inverse of $Eq.1$, that is $wt_1 = 1/\pi_{\mathrm{i}}$ (that is, the first-stage weight). A sub-script for stratification is omitted, but readers should note that BA selection probabilities (π_{i}) are often within strata, which can be multiple urban areas and/or categories within these urban areas.[30] Specifically, re-write $Eq.1$ by substituting $G = N - m_i - a_i$, giving:

$$1 - \frac{G!\,(N-n)!}{N!\,(G-n)!} \qquad (Eq.3)$$

$Eq.3$ can be expressed in exponentiated natural logs (as shown in $Eq.4$) which can be computed using programs such as Stata and R.[31]

$$e^{\ln(1)} - \frac{\frac{e^{\ln(G!)}}{e^{\ln(n!)}e^{\ln((G-n)!)}}}{\frac{e^{\ln(N!)}}{e^{\ln(n!)}e^{\ln((N-n)!)}}} \qquad (Eq.4)$$

Implementers can also calculate a second-stage weight, as the inverse probability of selection for an interview within an enumerated BA. Analysis of the data of interviewed businesses, at the level of SSUs (j), can then be done using weighting given by $wt_{1(i)} * wt_{2(j)}$.

[29] The ISBS currently uses the World Bank's Survey Solutions. For more details, see https://mysurvey.solutions.

[30] Note that while stratification can help reduce fieldwork effort due to the grouping of similar BAs, it can present an issue when there is the possibility of networks crossing strata due to expansions (as initial PSU selection probabilities are a function of strata). The ISBS deals with this issue by not allowing expansions that cross strata. Also, note that where stratification is defined as separate urban areas (and not further stratification within an urban area), the issue is fully ignorable.

[31] For example, the Stata function ln factorial(n) can be applied to $Eq.4$. Code for this calculation is available from the authors upon request. Implementers should also note that it is important to set variable storage to 'double' to maximize storage precision. Note that storage precision may be machine and processor dependent.

Statistical packages—notably Stata—are equipped to handle such multi-stage sampling and weighting.[32] Since the probability of selection, $pr(sel_{LF})_j$, is set before enumeration, implementers will most likely want to apply weight adjustments based on unit refusals and the misclassification of units during fieldwork.[33] Implementers can adjust $wt_{2(j)}$ based on the total number of encountered informal businesses in a BA. The ISBS specifically applies three such adjustments, based on different assumptions, all of which remove confirmed formal businesses: i) counting all unit refusals as ineligible (strict assumption); ii) counting unit refusals with visible signage or permits as ineligible (median assumption); and iii) counting all unit refusals as eligible (weak assumption).

6. Discussion and Conclusions

While this paper lays out a proposed methodology for surveying informal businesses—along with several practical guidelines and informative results—several challenges remain. The methodology requires intensive enumeration, which must go hand-in-hand with monitoring. Some of these challenges can be minimized by technology, through data and fieldwork quality control, including path monitoring. There also remain risks to under-counting of informal businesses that are not readily visible, including those in households or that intentionally remain less visible, perhaps to avoid detection by authorities. This is borne out by the range of the share of businesses operating in the household. Examples from iterations of the ISBS include 59 percent of informal Somali businesses operate within the owners' household compared to just 14 percent in Lao PDR. Missing these types of businesses during enumeration is a potential form of design-based bias.

[32] In Stata, this can be done by correctly specifying the svy: prefix via svyset. For example, the following code snippet will declare survey data, with two-stage sampling, with stratification in the first stage (given by strata). The first stage weight (wt_1) is at the level of an identified cluster (id_cluster), while the second stage is identified at the level of id_j with a second-stage weight. For a discussion, see: https://www.stata.com/support/faqs/statistics/stratified-multiple-stage-designs/

```
svyset id_cluster [pw=wt_1], strata(strata) || id_j [pw=wt_2]
```

[33] In the case of ISBS, $pr(sel_{LF})_j$ is a function of a j's position in an enumerated roster. Enumerators may encounter either refusals or list an ineligible business (e.g., one that is registered), resulting in the need for a weight adjustment.

Some of the selected BAs are market centers, often the largest concentrations of informal business activity and those with higher selection probabilities, in turn. However, this raises logistical challenges due to the sheer number of these businesses, making it daunting for a single enumerator to cover the entire BA in one visit. This may require assigning more than one enumerator per square. Secondly, markets may be open some days of the week and at certain times of the day, which requires visiting the area at the right time. In open markets, operators may sit next to each other, which may result in greater refusal. Finally, it is important to note that the survey may require visiting areas with safety issues. In some cases, the area could be outright dangerous for enumerators; such areas may need to be excluded from the survey with implications in the extent of representativeness (via the definition of N).

One last, specific challenge is that policy makers and researchers may want to build panels of data for informal businesses (or even of BAs/PSUs), as the dynamics in and out of informality are an important area for analysis. Likewise, implementers may want to recontact businesses in the course of implementing an ISBS-type survey. Revisiting businesses and building such panels is notably difficult when informal businesses are mobile or have no fixed premises. Recontacting such businesses may require the collection of contact information, such as mobile numbers, requiring full compliance with any existing privacy protocols.

References

Amin, M. 2021. "Does Competition from Informal Firms Hurt Job Creation by Formal Firms? Evidence Using Firm-Level Survey Data". *World Bank Policy Research Working Paper, No. 9515*.

Amin, M. and A. Islam. 2015. "Are large informal firms more productive than the small informal firms? Evidence from firm-level surveys in Africa". *World Development*, *74*, pp.374-385.

Amin, M. and C. Okou. 2020. "Casting a shadow: Productivity of formal firms and informality". *Review of Development Economics*, *24*(4), pp.1610-1630.

Benhassine, N., D. McKenzie, V. Pouliquen, and M. Santini. 2018. "Does inducing informal firms to formalize make sense? Experimental evidence from Benin". *Journal of Public Economics*, *157*, pp.1-14.

Benjamin, N.C. and A.A. Mbaye. 2012. "The Informal Sector, Productivity, and Enforcement in West Africa: A Firm-level Analysis". *Review of Development Economics*, *16*(4), pp.664-680.

Bowering, R., R. Wigle, T. Padgett, B. Adams, D. Cote, and Y. Wiersma. (2017. "Searching for rare species: A comparison of Floristic Habitat Sampling and Adaptive Cluster Sampling for detecting and estimating abundance", *Forest Ecology and Management,* 407.

Bried, J. 2013. "Adaptive cluster sampling in the context of restoration", *Restoration Ecology,* 21: 585–591.

Bruhn, M. and D. McKenzie. 2014. "Entry regulation and the formalization of microenterprises in developing countries". *The World Bank Research Observer*, *29*(2), pp.186-201.

Burkowski, J. and P. Turk. 2013. "Adaptive Cluster Sampling: An Introduction", Conference paper*: International Conference on Applied Statistics*.

Campos, F., M. Goldstein, and D.J. McKenzie. 2015. "Short-term impacts of formalization assistance and a bank information session on business registration and access to finance in Malawi". *World Bank Policy Research Working Paper, No. 7183*.

Christman, M. C. 1996, "Comparison of efficiency of adaptive sampling in some spatially clustered populations," in ASA Proceedings of the Section on Statistics and the Environment, 122-126.

Christman, M. C. 1997, "Efficiency of some sampling designs for spatially clustered populations," *Environmetrics*, 8, 145-166.

Coggins, S.B., N. C. Coops and M. A. Wulder. 2010. "Estimates of bark beetle infestation expansion factors with adaptive cluster sampling*", International Journal of Pest Management*, 57(1): 11–21.

Cullen, D.W. and B.G. Stevens. 2017. “Erratum to: Application of systematic adaptive cluster sampling for the assessment of black sea bass Centropristis striata abundance” *Fisheries Science*, 83(683).

Davis, J.G., S. B. Cook and D. D. Smith. 2011. “Testing the Utility of an Adaptive Cluster Sampling Method for Monitoring a Rare and Imperiled Darter”, *North American Journal of Fisheries Management*, 31(6): 1123–1132.

De Mel, S., D. McKenzie, and C. Woodruff. 2013. “The Demand for, and Consequences of, Formalization among Informal Firms in Sri Lanka”. *American Economic Journal: Applied Economics*, *5*(2), pp.122-50.

Donaldson, D., and A. Storeygard. 2016. "The View from Above: Applications of Satellite Data in Economics." *Journal of Economic Perspectives*, 30 (4): 171-98.

Dryver, A. L. and S. K. Thompson. 1998, “Improving unbiased estimators in adaptive cluster sampling," in ASA Proceedings of the Section on Survey Research Methods, pp. 727-731.

Elgin, C., M.A. Kose, F. Ohnsorge, and S. Yu. 2021. “Understanding informality”. *CAMA Working Paper No. 76/2021.*

Floridi, A., B.A. Demena, and N. Wagner. 2020. Shedding light on the shadows of informality: A meta-analysis of formalization interventions targeted at informal firms. *Labour Economics*, *67*, p.101925.

Gattone, S.A., M. Esha, and J.W. Mwangi. 2013. “Application of Adaptive Cluster Sampling with a Data-Driven Stopping Rule to Plant Disease Incidence”, *J Phytopathol*, 161: 632–641.

Hariharan, A. V. Gallucci, and C. Heberer. 2013. “Estimation of relative efficiency of adaptive cluster vs traditional sampling designs applied to arrival of sharks” *arXiv:1304.2460.*

Horvitz, D. and D. Thompson. 1952. “A Generalization of Sampling Without Replacement from a Finite Universe”, *Journal of American Statistical Association*, 47(260): 663–685.

ILO. 2013. “Measuring informality: A Statistical Manual on the Informal Sector and Informal Employment.”

Islam, A., 2019. “The burden of water shortages on informal firms.” *Land Economics*, *95*(1), pp.91-107.

Jolevski, F. and G. Aga. 2019. “Shedding light on the informal economy: A different Methodology and new Data”, *Let’s Talk Development.*

Kanbur, R. 2017. “Informality: Causes, Consequences and Policy Responses.” *Review of Development Economics*, 21 (4), 939-961

La Porta, R., and A. Shleifer. 2014. "Informality and Development." *Journal of Economic Perspectives*, 28 (3): 109-26.

Loayza, N.V., 2016. "Informality in the Process of Development and Growth." The World Economy, 39(12), pp.1856-1916.

Loayza, N., 2018. "Informality: why is it so widespread and how can it be reduced?" *World Bank Research and Policy Briefs*, (133110).

Meghir, C, R. Narita, and J.-M. Robin. "Wages and Informality in Developing Countries." *American Economic Review* 105, no. 4 (2015): 1509-46.

Ohnsorge, F. and S. Yu. 2021. "The Long Shadow of Informality: Challenges and Policies". The World Bank.

Perry, G. E., O. Arias, P. Fajnzylber , W. F. Maloney , A. Mason, and J. Saavedra-Chanduvi. 2007. Informality: Exit and exclusion. The World Bank.

Philippi, T. 2005. "Adaptive Cluster Sampling for Estimation of Abundances within Local Populations of Low-abundance Plants", *Ecology,* 86: 1091–1100.

Salehi, M., M. Moradi, J. A. Al Khayat, J. Brown, and A. E. M. Yousif. 2015. "Inverse Adaptive Cluster Sampling with Unequal Selection Probabilities: Case Studies on Crab Holes and Arsenic Pollution", *Australian & New Zealand Journal of Statistics*, 57(2): 189–201.

Salehi, M., G. Seber. 1997. "Two-Stage Adaptive Cluster Sampling". *Biometrics*, 53(3): 959–970.

Salehi, M. M. 1999, "Rao-Blackwell versions of the Horvitz-Thompson and Hansen-Hurwitz in adaptive cluster sampling," *Environmental and Ecological Statistics*, 6, 183-195.

Salehi, M. M. 2003, "Comparison between Hansen-Hurwitz and Horvitz-Thompson estimators for adaptive cluster sampling," *Environmental and Ecological Statistics*, 10, 115-127.

Schneider F., Hassan M. 2016. "Size and Development of the Shadow Economies of 157 Worldwide Countries: Updated and New Measures from 1999 to 2013", *Journal of Global Economics*, 2016.

Skibo, K., C. J. Schwarz, and R. M. Peterman. 2008. "Evaluation of Sampling Designs for Red Sea Urchins Strongylocentrotus franciscanus in British Columbia", *North American Journal of Fisheries Management*, 28:1: 219–230.

Smith, D., R. Villella, and D. Lemarié. 2003. "Application of adaptive cluster sampling to low-density populations of freshwater mussels", *Environmental and Ecological Statistics*, 10: 7–15. DOI 10.1023/A:1021956617984.

Thompson, S. 1990. "Adaptive Cluster Sampling*", Journal of the American Statistical Association*, 85:412, 1050-1059.

Thompson, S. and G. Seber. 1994. “Detectability in Conventional and Adaptive Sampling”, *Biometrics*, 50(3): 712–724.

Thompson, S. and G. Seber. 1996. Adaptive Sampling. New York: Wiley

Tout, J. 2009, “An Analysis of Adaptive Cluster Sampling Design with Rare Plant Point Distributions”, Humbolt State University (Thesis).

Ulyssea, G. 2018. "Firms, Informality, and Development: Theory and Evidence from Brazil." *American Economic Review*, 108 (8): 2015-47.

Ulyssea, G. 2020. “Informality: Causes and Consequences for Development.” *Annual Review of Economics*, 12:525-546

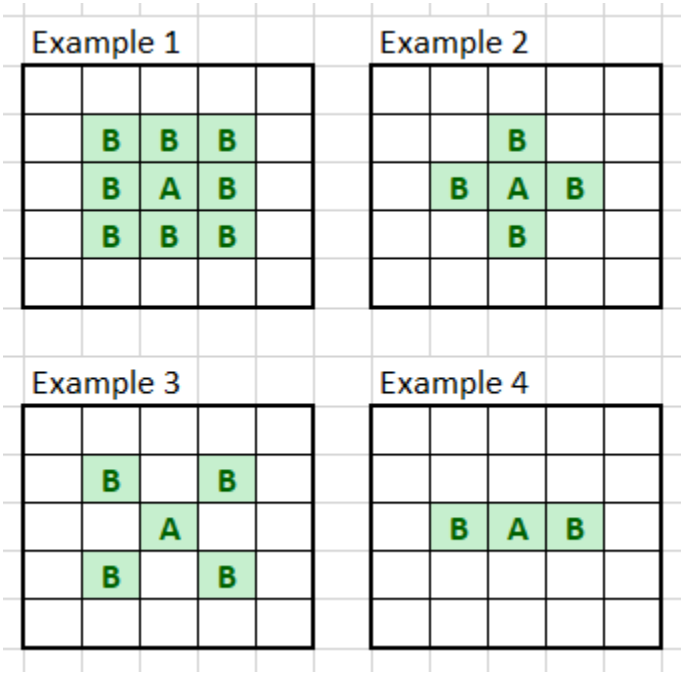
World Bank. 2020. Re-thinking the Approach to Informal Businesses: Typologies, Evidence and Future Exploration.

Yu, H., Y. Jiao, Z. Su, and K. Reid. 2012. “Performance comparison of traditional sampling designs and adaptive sampling designs for fishery-independent surveys: A simulation study”, *Fisheries Research*, 113(1): 73–181.

Appendix

A.1. Examples of Neighborhood Definitions

The figure below demonstrates some possible definitions of neighborhood. For this paper, we use what is known as a second-order neighborhood containing the original square and its 8 adjacent neighbors shown in Example 1. All ‘B’s’ are said to be in the neighborhood of ‘A’. Example 2 is a first-order neighborhood that is also commonly used in ACS. Example 3 is rarely used while Example 4 is a version of ‘strip-sampling’ that is sometimes used for certain ecological phenomena.



A.2. Population and Estimated Variance, HT Estimator Using ACS

In terms of variance, it is valuable to consider notation in terms of separate networks. Denote two separate networks, the k-th and l-th networks, respectively. Using SRS without replacement, the probability, π_{kl}, that an initial sample has an intersection of at least one BA is given by (Thompson, 1990):

$$\pi_{kl} = 1 - \frac{\left[\binom{N-m_{ik}-a_{ik}}{n} + \binom{N-m_{il}-a_{il}}{n} - \binom{N-(m_{ik}-a_{ik})-(m_{il}-a_{il})}{n}\right]}{\binom{N}{n}} \qquad (Eq.A1)$$

In turn, the variance of $\hat{\mu}_{HT}$ is given by:

$$Var[\hat{\mu}_{HT}] = \frac{1}{N^2}\left[\sum_{k=1}^{\zeta}\sum_{l=1}^{\zeta} y_k^* y_l^* \left(\frac{\pi_{kl}-\pi_k\pi_l}{\pi_k\pi_l}\right)\right] \qquad (Eq.A2)$$

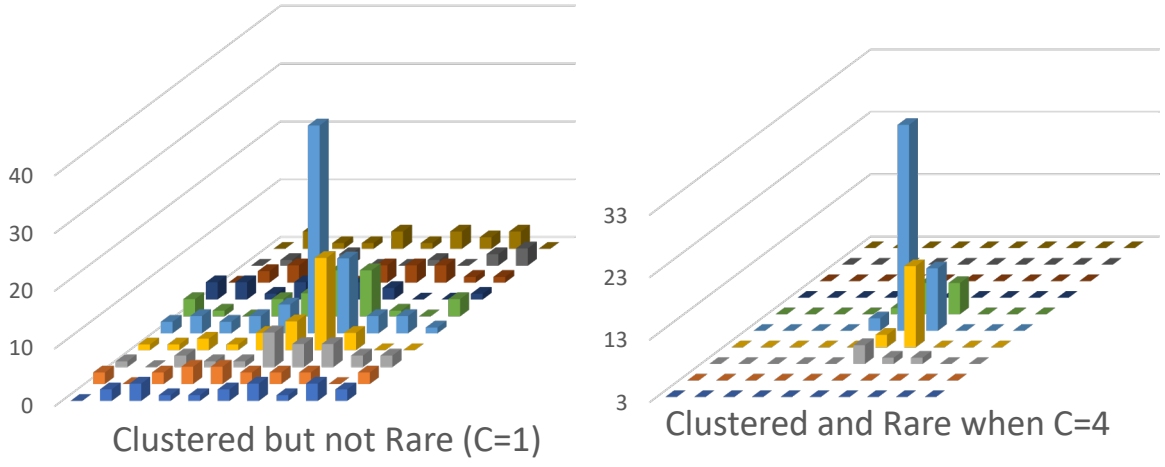
Where ζ is the total number of networks in the population. An unbiased estimator of this variance is then expressed by:

$$Var[\hat{\mu}_{HT}] = \frac{1}{N^2}\left[\sum_{k=1}^{\kappa}\sum_{l=1}^{\kappa} y_k^* y_l^* \left(\frac{\pi_{kl}-\pi_k\pi_l}{\pi_k\pi_l}\right)\right] \qquad (Eq.A3)$$

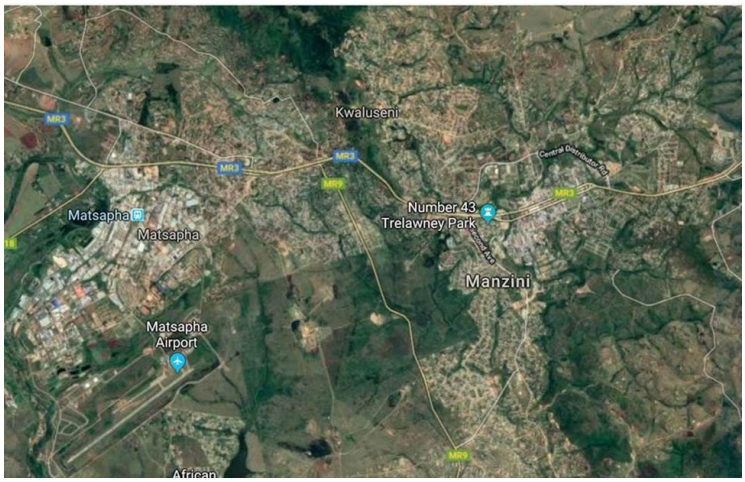
A.3. Illustrations of Clustering and Rareness Relative to Expansion Thresholds

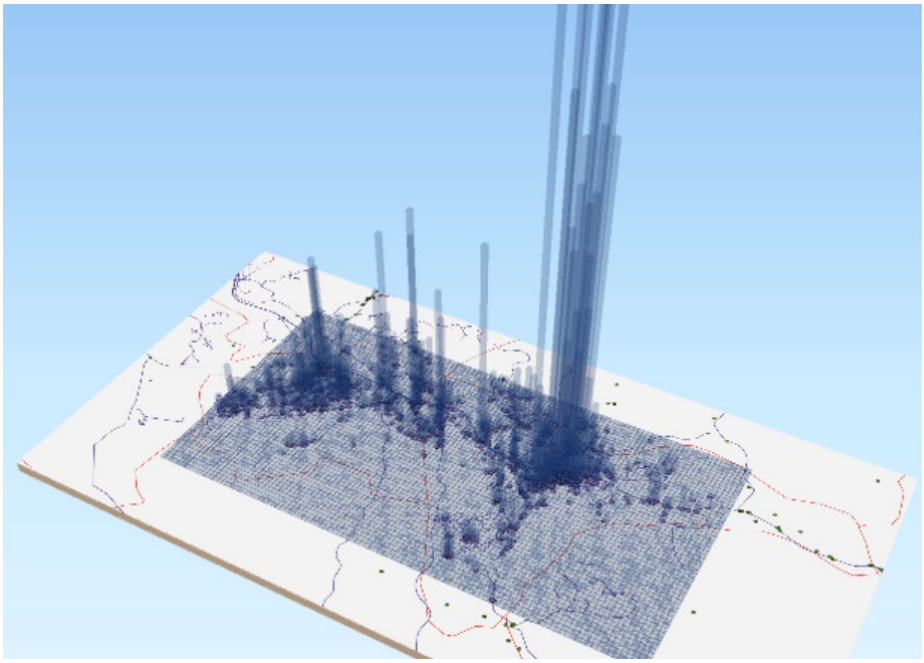
The realized reduction in fieldwork effort follows from both the underlying characteristics of a population of interest and the details implementers chose for the ACS design (see Borkowski and Turk [2013] for a discussion). Consider, for example, a population that is visibly clustered, but whose rareness depends on the expansion threshold parameter chosen (C). Figure A1.a shows a population that can be considered clustered but not rare under $C = 1$. Figure 2.b shows the exact same population, but the plane has been 'flooded' by raising the expansion threshold, to $C = 4$, such that the BAs containing fewer than 4 units cannot be 'seen'. As such, the y-axis in 2.a shows the raw count of the underlying population, while the axis in A1.b is reduced by 3 (the highest count that does not trigger expansion). As a result, the 'flooded' population is now both clustered and rare.

Figure A1: Examples of a Clustered but non-Rare Population



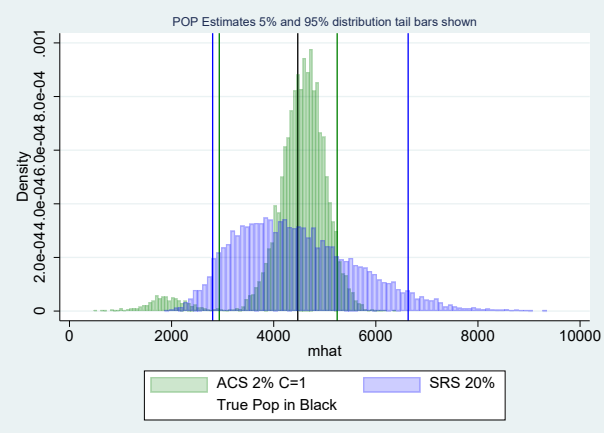
A.4 Summary of Eswatini Economic Census

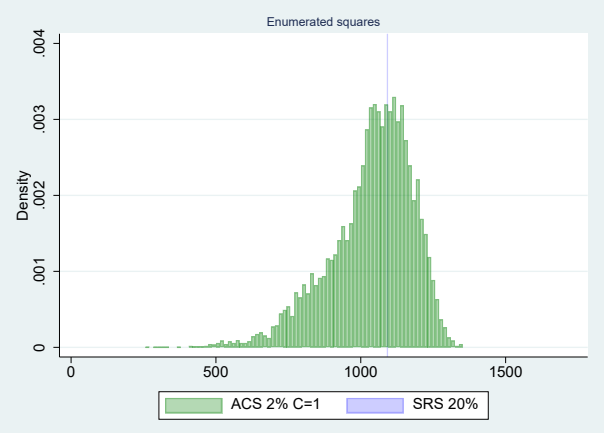




A.5. A Direct Comparison of ACS and SRS, Fixing Specific Parameters

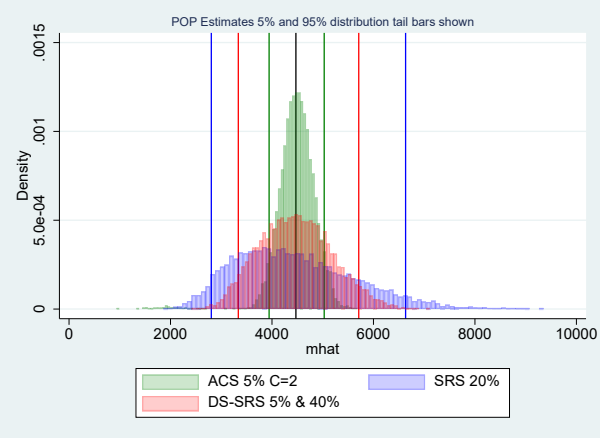
The figures below show comparisons between two sets of simulations, each corresponding to a row in Table 1, with approximately the same (mean) amount of fieldwork effort (as shown in panel b). These are ACS: *C=1,* n=2% (mean fieldwork of 1,035 BAs in table 1), and SRS: n=20% (with associated fieldwork of 1,092 BAs). What is immediately clear, then, in panel a is that ACS provides a mean unbiased population estimate with much lower mean standard error, underlining the intuition that the adaptive discovery process allows for more statistically efficient estimates at the same level of fieldwork effort. Note a small peak of negatively biased estimates (that is, to the left of the distribution) that occurs in cases where a comparatively low initial selection of BAs fails to produce sufficient expansions (or misses networks completely) to estimate the true population mean. Panels c and d show what occurs if the ACS n is increased to 5%, compared to an SRS of n=20%: both mean variance and mean fieldwork effort drop noticeably, with the disappearance in the distribution of a small peak of downwardly biased estimates in panel a. The figures also give an important nuance. While the last columns in Table 1 show the mean fieldwork effort in ACS vs. SRS, panels b and d in Figure 4 show the distribution of these relative efforts. Under ACS with n=2%, *C=1* roughly half of the simulated fieldwork efforts are reductions relative to a (known) effort using an SRS n=20%. When n=5%, *C=2*, virtually the entire mass of simulated fieldwork effort is below the effort in SRS n=20%.

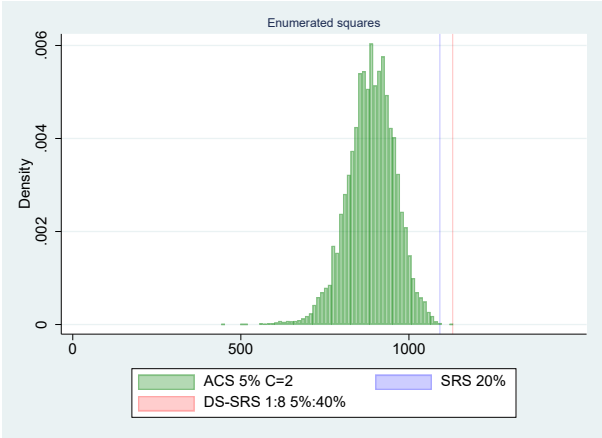




A.6. A Direct Comparison of ACS and SRS, Fixing Specific Parameters with Stratification

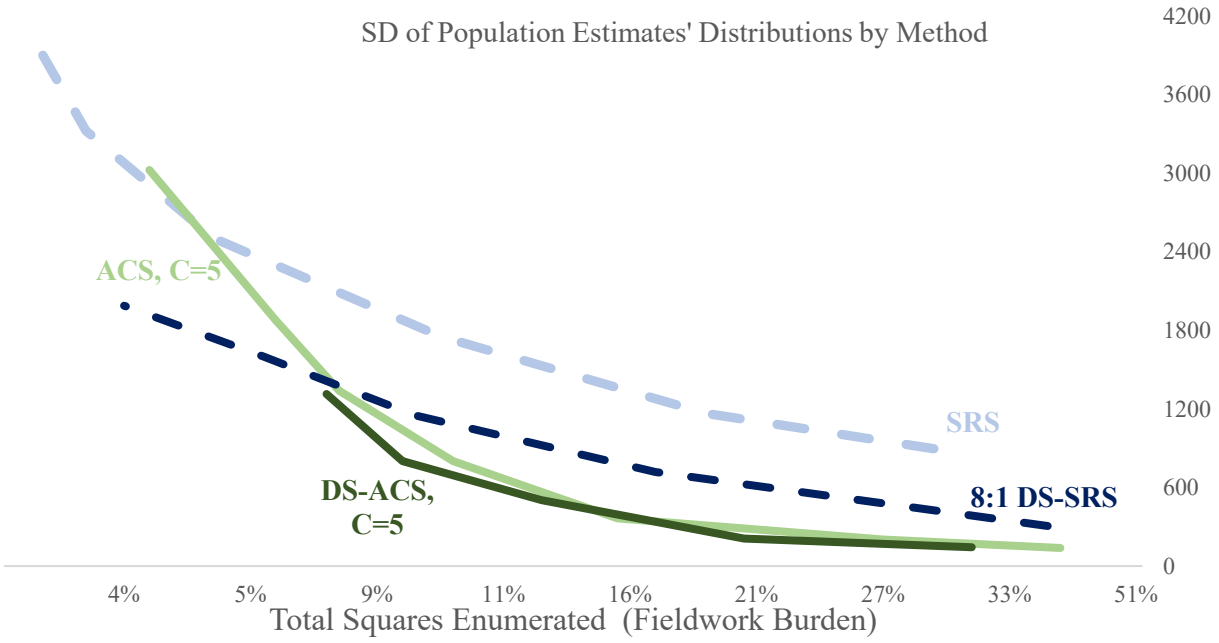
As it was the case of examining proportional SRS, introducing proportional stratification to ACS (S-ACS) results in small but significant improvements in the distribution of population estimates and expected fieldwork requirements. Expected fieldwork requirement distributions appear somewhat sensitive to condition *C* when comparing naive ACS with S-ACS as shown below. In addition to the improvements in fieldwork certainty (and sometimes reductions) shown below, adding stratification to ACS also appears to reduce the left-tail error visible in some of the figures shown above.[34]





[34] Left-tail error (here observations with pop estimates <3,000) is anywhere from 2% to 25% smaller depending on the specific parameters.

A.7. Comparison of ACS and SRS with Fixed Parameters, with and without Stratification



A.8. ACS and SRS with Increased Density

As the population density of the area covered by our data set is relatively low compared to larger cities in other countries, an artificial data set was simulated by duplicating and randomly displacing the population of establishments in Matsapha and Manzini. This synthetic data set is used to evaluate how the density of the underlying population might affect the efficiency of the ACS estimators relative to SRS. The duplication and displacement is conducted through four iterations, so the new simulated population is approximately four times as large. The geographic area remained the same and the proportion of occupied squares rose from 13% to 42%. The resultant data is shown in Figure A.8.1 and visual inspection confirms that clustering is retained through this process.

Figure A.8.1: Synthetic Data

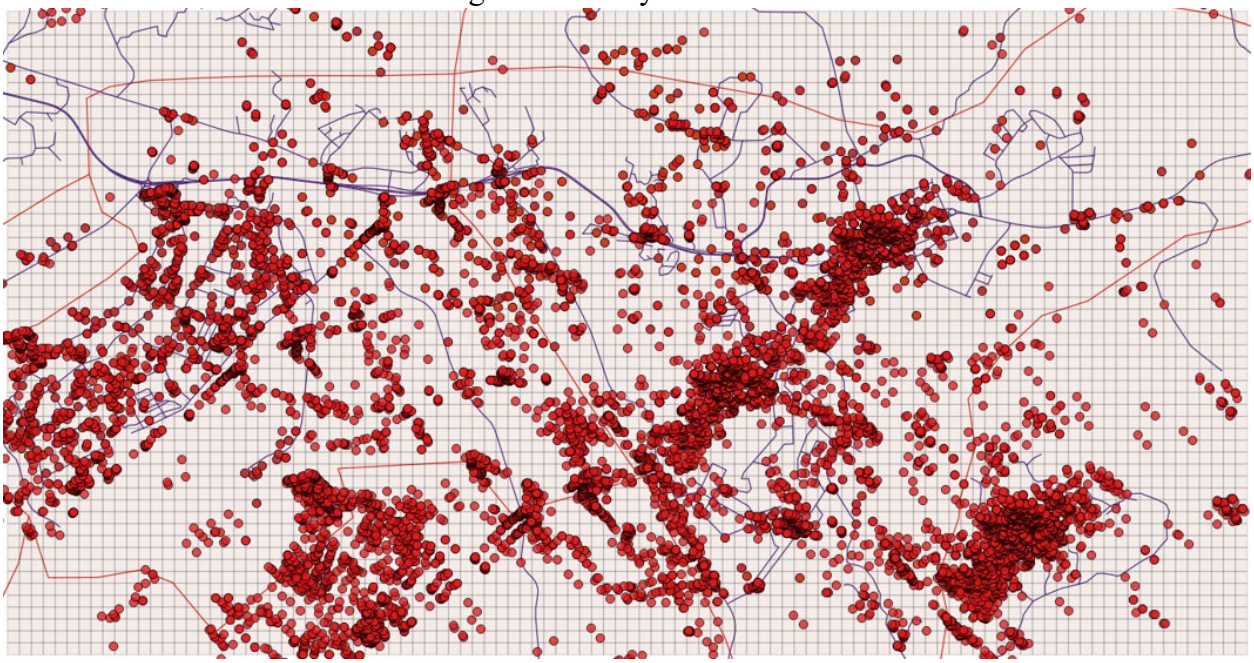
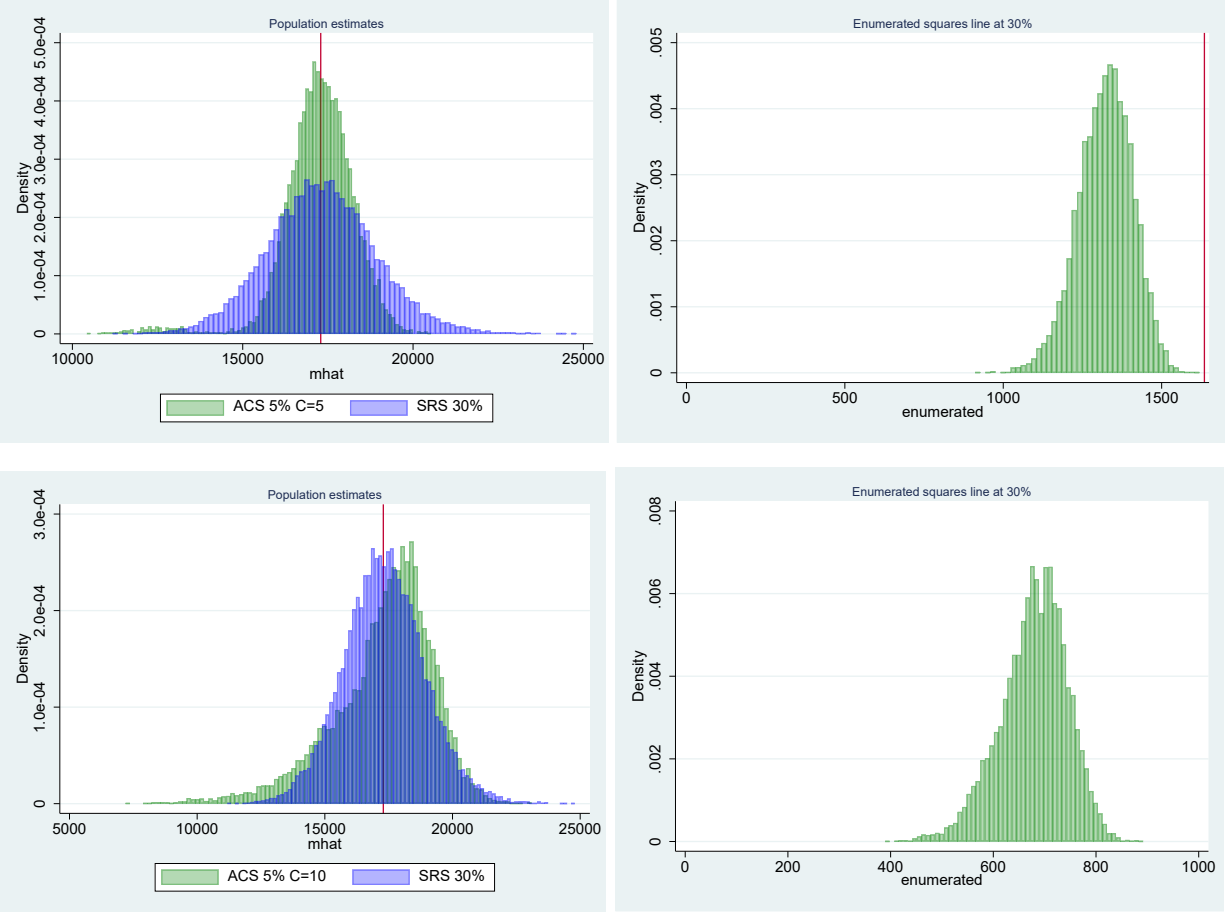


Figure A.8.2: ACS with Greater Density Population



A.9. Useful Categories for Stratification

Strata	Stratification of Blocks	Definition
1	Residential	Any use of land that serves as a place of residence. This includes: multi-family homes (apartment buildings), single family residences (houses, including backyards), or mobile home parks.
2	Commercial/Industrial	Urbanized areas that encompass retail or service activities or land uses for the production, storage, and distribution of manufactured goods. Commercial areas include: business centers, retail centers (shopping malls) or retail stores, service areas (banks, restaurants, repair shops), office space, space of public assembly (auditoriums, convention halls, stadiums), and institutional space (government buildings, hospitals, schools). Industrial areas include: industrial parks, and localities in which industrial production occurs (ex: steel mills, paper mills, chemical plants, oil refineries).
3	Mixed (Residential/Commercial/Industrial)	Includes areas in which there is a combination of the two strata discussed above
4	Market Centers	Any form of land use with a dense concentration of sellers, such as farmer's markets, bazars, etc. This a characterized by near certainty of encountering informal businesses.
5	Open Area	Refers to accessible open space where little commercial or industrial activity is expected. Examples of this are: parks, zoo, gardens, cemetery, camping space, agricultural fields, vacant spaces.
6	Inaccessible	Land in which there are no roads or walkable paths. It includes: open water (any water body such as lakes or rivers), restricted access areas (such as airport runways, military bases), forests, and mountainous areas.