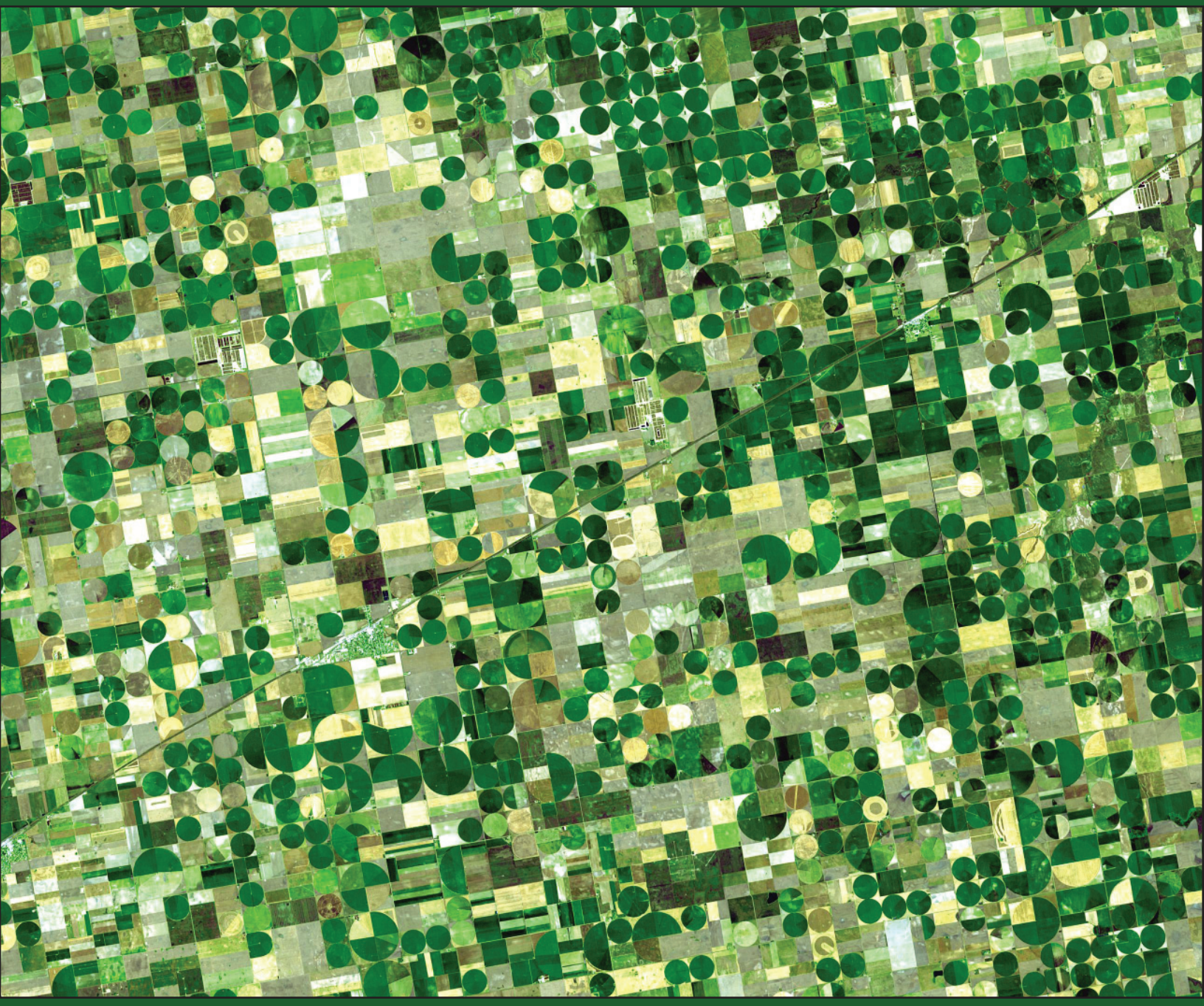
GIS and Sampling



Peter M. Lance • John Spencer • Aiko Hattori

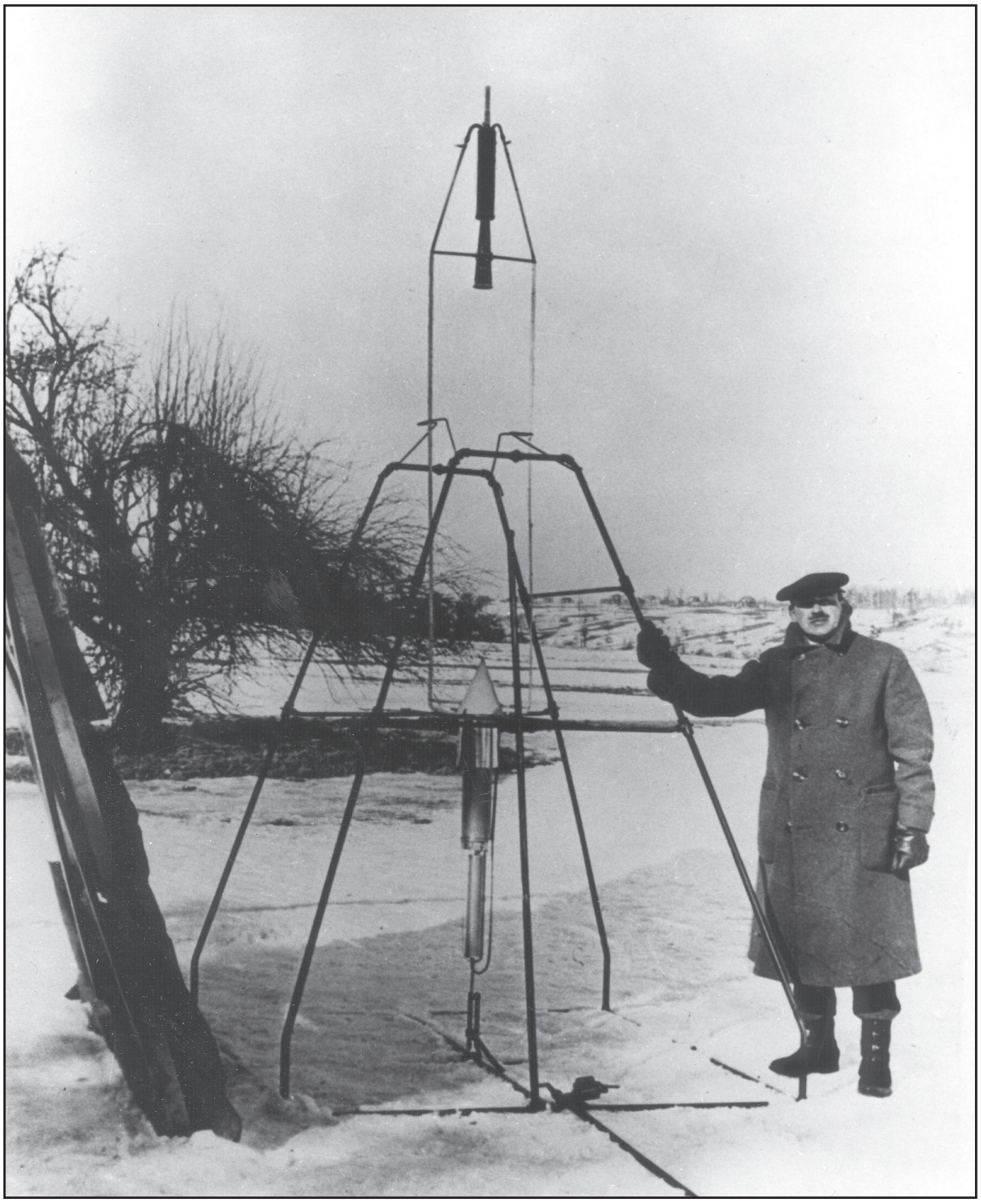
GIS and Sampling

Peter Lance
John Spencer
Aiko Hattori

This research has been supported by the President's Emergency Plan for AIDS Relief (PEPFAR) through the U.S. Agency for International Development (USAID) under the terms of MEASURE Evaluation cooperative agreement GHA-A-00-08-00003-00, which is implemented by the Carolina Population Center at the University of North Carolina at Chapel Hill, with Futures Group, ICF International, John Snow, Inc., Management Sciences for Health, and Tulane University. The views expressed in this publication do not necessarily reflect the views of PEPFAR, USAID or the United States government.

Cover image features a satellite image of crops growing in Kansas, United States. Courtesy of NASA,via Wikimedia Commons through the license PD-USGov.

August 2014 **MS-14-95**



Source: Wikimedia Commons, PD-USGov

It is difficult to say what is impossible, for the dream of yesterday is the hope of today and the reality of tomorrow.

– Robert H. Goddard

Table of Contents

Chapter 1. Introduction 1
Chapter 2. Sampling – Basic Principles and Procedures 5
Chapter 3. Geographic Information Systems 29
Chapter 4. GIS and Sampling 49
Chapter 5. Case Study: A Survey of Urban Health in Bangladesh 79
Chapter 6. Conclusion 97

Figures and Tables

Table 1. President Smith's Approval Rating 7
Figure 1. Efficiency 9
Figure 2. The Population 11
Figure 3. A Random, Representative Sample from the Population 12
Figure 4. A Non-random, Non-representative Sample from the Population 13
Table 2. The Frame of All Women in the City 16
Table 3. The Random Start 17
Table 4. The Second Selection 17
Table 5. The Third Selection Onward 18
Figure 5. Multistage Sampling 21
Figure 6. Census Tract Map of North Carolina 22
Table 6. A Hypothetical Frame for Dhaka 23
Table 7. Cumulative Size 23
Table 8. The First Selection 24
Table 9. The Second Selection 24
Figure 7. Probability Weight and Information Contributions 26
Table 10. A Hypothetical Frame for Dhaka 28
Figure 8. Cote d'Ivoire Continuity of Care for Refugee ART Patients in Ghana, April 2011 30
Figure 9. Two Points at Ohene Djan Stadium 38
Figure 10. Age of Child in Household 39
Figure 11. Zonal Configuration #1 40
Figure 12. Zonal Configuration #2 40
Figure 13. An Image from Google Earth 41
Figure 14. Slum Area in Dhaka, Bangladesh 43
Figure 15. Makoko Section of Lagos, Nigeria 45
Figure 16. Satellite Image of Dhaka, Bangladesh 46
Figure 17. Google Street View Image 48
Figure 18. The Aedes Aegypti Mosquito as an Adult 51
Figure 19. Active Dengue Clusters in Singapore as of April 25, 2014 52
Figure 20. An Aedes Larvae Resting Comfortably (in Clean Water) 53
Figure 21. A Larger PSU (Outlined in Green) Generated from Several Enumeration Areas 55
Figure 22. The Peri-Urban Fringe of Kibera Slum, Nairobi, Kenya 57
Figure 23. A River Runs Through It: Technically These Communities Are Only 225 Meters Apart 59
Figure 24. Three Brick Factories in Bangladesh 60
Figure 25. Google Traffic of Map Ramseur, NC, USA 61
Figure 26. Google Traffic Map of Johannesburg, South Africa 62
Figure 27. The Distribution of West Nile Cases in the Dallas Area, August 2012 64
Figure 28. Urbanization in Wealthy and Lower Income Nations and Bangladesh 80
Figure 29. Slum Sampling from a Census Frame: EAs with Equal Population Sizes 81
Figure 30. Slum Sampling from a Census Frame: EAs with Unequal Population Sizes 82
Figure 31. The Slums of Dhaka in 1996 84
Figure 32. The Korail Slum 85
Figure 33. The Slum Survey Instrument 89
Figure 34. The Slum Map of Dhaka 91
Figure 35. Dhaka, Ward 17 92

Chapter 1. Introduction

This is an era of incredible innovation in programming and policymaking, as increasingly sophisticated intervention design is motivated by the desire to reach particular, often vulnerable, populations with very specifically designed programming tailored to address their distinct needs, frequently in some narrow fashion. A few major areas of programming where this evolution is already well underway include:

- Effective HIV prevention programming revolves around those populations most vulnerable to infection while treatment-focused programming is often concerned with those with the least access to care;
- Programming designed to address the challenges associated with life in densely settled, environmentally vulnerable concentrations of poverty (in short, with life in slums) by definition must focus on the residents of those communities;
- Programs designed to reduce vulnerability to and coping capacity in the face of environmental (e.g. air pollution) or epidemiological (e.g., dengue or malaria) risk must focus on those particularly at risk in these respects;
- Disaster recovery programming must focus first on those hardest hit by them;
- Programming designed to address the costs of climate change must identify those members of a society most vulnerable to the greatest costs from environmental change.

The common theme in all of these cases is that the programming moves beyond addressing the welfare of the average member of society and instead focuses on human welfare gaps that specific vulnerable, marginalized or hard to reach subpopulations of that society particularly face.

This is also an era of increasingly evidence-driven program design at every stage of the programming process. One manifestation of this has been a tremendous increase in demand for survey data that can supply such "evidence" at the population, or in the case of this emerging, more targeted programmatic approach, specific sub-population level.

Surveys, which collect information from a sample of populations of interest, are a powerful tool for informing programming. In short, surveys allow us to learn about the characteristics or circumstances of an entire population (or subpopulation) of interest based on information provided by only a small sample from it. The key to doing so is that the sample be representative, meaning that their experiences, circumstances, and characteristics reflect those of the large population or subpopulation of interest. The goal of survey sampling is to obtain such representative samples.

Surveys can inform the programming process in numerous ways. First, program designers need *ex-ante* information about the profiles of the sub-populations of interest: their characteristics, environmental circumstances, human welfare levels, needs, and constraints. This information is essential for appropriate program design as well as for identifying vital gaps in programming. Next, there is a need for focused information regarding program execution (market share, participant profiles, etc.) that can guide crucial mid-course corrections to program design and execution. Finally, *ex-ante* and *ex-post* information is often required to assess program impact. Average impact of the program across all of society typically is not a particularly important consideration for programs targeted to specific sub-populations. Instead, interest is more likely to be focused on impact within the subpopulation that was the focus of programmatic design and activity.

Efficiently designed surveys that address these demands for program-relevant information require a particularly targeted selection of samples that are representative of the subpopulations of interest to designers and managers of those programs. Unfortunately, such targeted sampling can be a challenging business in many societies with the generally available resources for sampling for population surveys. In particular, the most readily available sampling frames (i.e., lists from which survey participants are selected) are often official frames that tend to be census-based (though what follows typically applies to many if not virtually all official or otherwise "validated" frames).

While these vary considerably from nation to nation and even within some nations over time in terms of their quality and the detail they offer, a general theme is that it is not clear how to use these official frames to sample precisely from the subpopulations of interest to many targeted programs. Official frames more often than not offer little information about detailed population characteristics and, when they do, that information can be quite inaccurate.

This can be an important limitation when one considers that many vulnerable subpopulations of interest tend to be spatially concentrated. The spatial concentration may be due to physical environmental conditions that pose health risks to residents, spatial clustering by types of people that relate to genetic, economic or socio-demographic factors that influence susceptibility to health risks, or both.

Whatever the reason for their potential spatial concentration, straightforward sampling of members of such subpopulations from a national frame can be quite inefficient, as many sampling units selected from them will not contain many or even any members of the vulnerable subpopulation of interest. In short, many selected sampling units are likely to be essentially useless in terms of obtaining information about the subpopulation of interest to the survey, a conclusion that is often reached only after expensive and time consuming field visits.

However, the observation that subpopulations of interest are oftentimes spatially concentrated offers a potential avenue for more efficient sampling: craft a geographic information system (GIS) that captures their spatial distribution so that knowledge of where these subpopulations are concentrated can be used to guide more precise sampling from them. This manual explores how GIS can help inform more targeted sampling for population surveys.

GIS is a powerful tool that allows for the linking of disparate types of information through a common geospatial framework. In other words, GIS allows us to link information from different sources if those various sources tell us the places to which that information applies.

The tools of GIS can enable those seeking to build an evidence base for programming to achieve far more efficient sample selection, particularly for surveys where specific population subgroups are of interest. The basic reason for this is that the information GIS provides can allow for more precise identification of where these subpopulations of interest can be found than is generally possible with traditional frames.

This holds the promise of huge decreases in the costs of obtaining this information, and in several senses. More precise sampling means smaller, more focused samples since there would presumably be less need to screen for potentially comparatively small subpopulations of interest from large samples representative of the general population. This implies a (potentially great) reduction in the financial and time cost of surveys to learn programmatically crucial information at the level of focus subpopulations. This would also likely reduce tremendously the sheer logistical complexity of field work, allowing for more focus on things like data quality. Ultimately, putting GIS in the service of sampling could lead to faster, better and cheaper information to guide programming targeting specific subpopulations.

This greater sampling precision can be achieved either by augmenting the information in official frames traditionally used for sampling or by using GIS as a platform for crafting entirely new sampling frames. To be sure, the scope for merging the information synthesized through GIS with official frames can be limited by the fact that geocode information for the sampling units (often referred to as enumeration areas or census enumeration areas) in official frames is often unavailable or insufficient. (By geocode information we mean the precise, accurate geographic coordinates for sampling units within the frame, including their boundaries.)

Official frames for many poorer societies are simply not geocoded or not geocoded well, though this is starting to change with the widespread availability of increasingly cheap, relatively easy to use tools for geocoding (such as more and more accurate and cheap GPS devices, precise and more timely high resolution satellite photography, etc.).

However, even when a frame is well geocoded access to the geocode information can be problematic given an array of political sensitivities and privacy, ethics, and sometimes, security concerns.

Fortunately, the rapidly evolving tools of GIS are also helping to create the necessary conditions for a kind of "democratization of sampling frame construction": it is becoming easier and easier for teams of investigators to build their own sampling frames. Put simply, GIS provides a natural framework for producing so called "area frames" whereby the area in which a subpopulation of interest is concentrated can be segmented into mutually exclusive and exhaustive sampling units. By allowing us to identify where subpopulations of interest live or are concentrated, GIS can allow for the crafting of precise, focused area frames that capture specific subpopulations of interest with a high degree of efficiency (e.g., far fewer selected sampling units will prove useless by not containing any members of the subpopulation of interest).

A canonical example we discuss involves the construction of innovative sampling frames for the slum and non-slum areas of the major cities of Bangladesh. That exercise was motivated by the desire to conduct a survey of urban health in those cities that would generate indicators representative of slum and non-slum populations. The problem the research team faced was that there were no explicit official frames for slum and non-slum areas or populations in Bangladesh and the oft used official frame might actually generate small samples of slum dwellers. The research team responded by using geo-referenced (i.e., rendered in terms of geographic coordinates) satellite images to identify likely slum concentrations and then, using an innovative synthesis of GIS and traditional field work methodologies, verify them and detect any slums not evident from the photos.

This example will frame much of the discussion of the application of GIS to sampling. The basic principles behind it have potentially far reaching applications. First, GIS allowed for rather precise sampling of subpopulations (slum and non-slum dwellers) not readily detectable *a priori* from existing official (e.g. census-based) sources. Second, the GIS work was integrated into other fundamental methodologies, in this case traditional fieldwork methods. Third, the work was supported by multiple funders and the outputs were designed to be as leverageable as possible, not only as a platform for sampling but also as a tool for more targeted programming. Diverse funding and designing to serve multiple purposes made an undertaking such as this more justifiable from a cost and effort perspective.

Though this slum mapping was designed to parse the study cities into slum and non-slum areas with a fairly high degree of precision, the application of GIS to sampling does not necessarily need to be a perfect process. Indeed, a more realistic, but still quite attractive, goal in many applications is simply to make sampling more precise than it otherwise would be. For instance, there are many potential settings in which GIS can reduce the need for oversampling considerably simply by providing a reasonably more precise avenue for isolating a subpopulation of interest than frames based on, for instance, official sources such as census data.

To be sure, GIS cannot always usefully inform the sampling process and in no sense represents a solution to the challenges of *ex ante* identification of key populations of interest for sampling. However, there are many instances where GIS can render an otherwise infeasible study goal attainable.

There are, of course, intrinsic challenges that may persist regardless of how GIS-related technology or the richness or precision of geocoded information improves. For instance, it might be impossible or prohibitively expensive to identify spatial patterns to many sorts of subpopulations of interest through the existing tools of GIS. However, we are now experiencing an almost mind-boggling pace of technological change in terms of the tools available for GIS work. Indeed, to a certain extent this concern may in fact be evaporating before us.

This manual explores these issues. The next two chapters provide background to the basics of sampling and the tools of GIS. This is necessary to understand the requirements of the sampling process from a statistical perspective and what exactly GIS offers. We then introduce key principles to guide the application of GIS to sampling, and discuss them through the prism of the Bangladesh slum mapping example. We then briefly discuss the exciting new technological possibilities that hold the potential to expand the possibilities for the application of GIS to sampling in a fashion that is nothing short of revolutionary.

Source: Shutterstock

Chapter 2. Sampling – Basic Principles and Procedures

This manual details how GIS can inform the sampling process, opening the door to powerful new possibilities for obtaining samples representative of particular subpopulations of interest, and the considerations surrounding the application of GIS to sampling. In other words, it is concerned with how GIS can serve sampling and what the procedure for applying GIS to the sampling process entails. To understand the discussion that will unfold, it is necessary to have a solid grounding in both the principles of sampling and the basic tools and possibilities of GIS. In this chapter, and the next, both are provided.

The focus of this chapter is sampling. Sampling is considered by many to be a rather arcane, technical area best left to experts. However, sampling more often than not actually revolves around some very simple principles and concepts. Even when rather complex sampling techniques are used, they tend to be motivated to a significant extent by these relatively elementary principles and concepts. In this chapter we walk through some of the basics of sampling. This will help to build our understanding of the process of sampling, and in particular provide a sense of the requirements for successful selection of samples representative of populations (or subpopulations) of interest.

Perhaps it is best to start the discussion by recognizing the most basic goal in conducting most surveys: to learn something about a population. By something, we mean some parameter of interest. The parameter might be the prevalence of a behavior, outcome or characteristic, or some average behavioral outcome, human welfare outcome or characteristic. For instance, we might wish to know the modern contraceptive use rate among all women of reproductive age in Bangladesh, the average wealth of Americans, or what percentage of Egyptian adults finished high school. Whatever the case may be, of particular interest is the true value of that parameter across the population.

Generally, we cannot immediately know the value of such parameters across an entire population. There are two main options for learning something about the population value for a parameter.

One potential avenue for learning the population value of a parameter is a census. A census involves creating a list of every member of a population and, typically, gathering some information about each of them. Their key feature is that they strive to list and capture information for *every* member of the population. A population parameter can be learned from a census because the information determining it (provided it was collected) is known for every individual in the population. For instance, if we know the personal income of every individual in a complete census of a population, then its average across them is population average personal income, its total across them is population total personal income, etc.

Censuses tend to be conducted by national governments per some official mandate that governs the frequency and scope of the information gathered. For instance, the United States census is mentioned in Article I, Section 2 of the United States Constitution:

> "*Representatives and direct Taxes shall be apportioned among the several States ... according to their respective Numbers ... . The actual Enumeration shall be made within three Years after the first Meeting of the Congress of the United States, and within every subsequent Term of ten Years.*"

Similar motivations drive many other national censuses,[1] though in many societies censuses are conducted less frequently, and less information is collected, than in the U.S. (indeed, some nations go decades between national censuses).

Given the typical expense and operational challenge of conducting censuses, it is difficult to gather a wide range of detailed information via a census. Of the examples furnished (the modern contraceptive use rate among all women of reproductive age in Bangladesh, the average wealth of Americans, or what percentage of Egyptian adults finished high school), perhaps only the high school completion rate of adults might plausibly be available through an official census. However, even this variable is sometimes unavailable (or at least not reliably available) in national censuses.

Moreover, censuses are conducted relatively infrequently, limiting their usefulness for learning about parameters that fluctuate more over time. For instance, the United States, Egypt and Bangladesh (the three specific countries mentioned thus far) all conduct decennial (i.e., once every ten years) censuses, with other societies sometimes doing so on a far less frequent basis.

In the absence of an official census, calculating a population parameter by gathering information related to it for every member of the population could be a daunting task. While this is obvious in the case of most national populations, it can be true even of comparatively small populations. Consider, for instance, a randomized control trial (RCT) that, out of a group of a few hundred villages, randomly assigns some to receive a program. The program designers might want to learn about program impact for the populations of the villages selected to participate in the experiment. The combined populations of the treatment and control villages might be small compared with the society as a whole, yet still too large to interview every household or individual within them.

Despite their limited usefulness for learning directly about many population parameters of interest, censuses can still be of great usefulness for operationalizing the alternative method of learning about the population parameter values that lie at the center of this manual. The reason being that the population lists that censuses yield can provide an ideal platform for selecting groups of individuals from the population. We will be discussing this in far greater detail in what follows.

The other major option for learning something about the population parameter is to do so by learning about the characteristics of a few of its members, and from the characteristics of those few members infer by estimation those of the population to which they belong. Surveys are a way of learning about a population by gathering information about a sample from that population that includes only some of its members. Specifically, surveys are essentially a process of observation of or collection of information (e.g., via a questionnaire) from a sample from the population. Because they are motivated by the idea of learning about the population from the information obtained through a sample from it, they are also referred to as population surveys.

Surveys convey the enormous benefit of learning something about a large group of people based on only a few among them. The information from the sample allows one to form statistical estimates of true parameter values across the populations from which the sample was selected.

To cite a classic example, consider political polling. In many societies, there exists a desire to know what percentage of the population approves of the performance of the leaders of that society. Clearly, it would likely not be feasible to interview every member of a society to learn about their political inclinations. In the United States, that would involve, as of this writing, interviewing about 314 million people. Even in a small country such as Belize (ranked by population size to be around 180th out of roughly 244 nations at the time of writing), this would involve interviewing upwards of 350,000 people, an enormous task.

[1] The historical motive for most censuses is to provide an organizing framework for taxation and, sometimes, military service levies. With the spread of democracy, insuring correct (i.e., in true proportion to population share) representation has become another important motivation for conducting censuses.

The typical solution has been to infer political approval at the population level using a small sample from that population. A political poll is a survey using a sample from the population (or some subpopulation of interest) designed to estimate political opinion across that entire population (or, as the case may be, subpopulation) of interest.

Political polls are an excellent example of just how much can be learned about huge populations from comparatively small sample sizes. For instance, Gallup, an American polling company, released poll estimates of President Barack Obama's popularity for the period February 25-27. They based this on a sample of 1,500 U.S. adults.[2] At present, the population of U.S. adults (those aged 18 and older) is probably around 250 million. The Gallup sample is only around 0.000006 percent of that population. Here, we see just how powerful surveys of a population can be: Gallup attempts to infer the net popularity of President Barack Obama among all U.S. adults from just *0.000006 percent* of them.

Nor is this potential limited to political polling: all kinds of things are learned about populations based on surveys of samples from them. For instance, the popular and well-known Demographic and Health Surveys (DHS) attempt to infer population parameters such as the contraceptive prevalence rate of ever-married women aged 15-49 in a society from samples that constitute (typically) a tiny percent of that population. To cite a typical example, there are probably around 9.9 million women aged 15-49 in Kenya,[3] but the most recent (2008-09) DHS there interviewed just under 8,500 women in this age range. In other words, the 2008-09 Kenya DHS attempts to learn things about all Kenyan women aged 15-49 from a sample that constitutes 0.00086 percent of them (and many DHS samples constitute a far smaller percentage of the population of interest).

Table 1. President Smith's Approval Rating

Presidential Approval	Poll Period	Sample Size	Target Population	Polling Firm
57	March 6-8	1,500	Adults	Galluper
48	March 6-8	1,500	Likely Voters	Foxy News
55	March 6-8	1,500	Adults	Un-Associated Press
50	March 2-4	1,000	Registered Voters	USA Tomorrow

We return to the arena of political polling to discuss a few other key qualities of samples. Consider, for instance, a series of national political polls to gauge the popularity of President Smith of the fictional Republic of Samplestan. These polls relied on samples interviewed in early March 2014. The essential features of these polls are given in the preceding table (Table 1). Table 1 presents President Smith's approval rate as well as some key features of the polls themselves.

The polling estimates of President Smith's popularity are quite varied. The polls covering March 6-8 offer a 9 point spread (from 48 to 57 percent) in the estimates of population approval of him as President. However, the two polls designed to provide estimates for all adults in Samplestan offer approval ratings in the mid- to upper-fifties while that targeting likely voters is the comparative outlier with a 48 percent approval rating. An immediately apparent possible explanation for the discrepancy is that the polls target different populations which might have different true sentiments regarding President Smith. This is an important lesson for the craft of sampling: selecting a sample representative of the intended population can be extremely important since the same parameter can have very different values in different populations.

However, this does not explain the different approval ratings (of 57 and 55 percent) provided by the two March 6-8 polls targeting all adults. One possible explanation for the difference in the estimates that they provide might be found in ordinary variation in results typical with surveys like political polls. Political polls produce not the exact population approval rating of the President for their target populations, but instead an estimate of it. This estimate will vary from poll to poll.

[2] As reported on www.realclearpolitics.com on March 1, 2014.

[3] Retrieved March 1, 2014 from the World Population Prospects website (http://esa.un.org/wpp/unpp/panel_indicators.htm).

For instance, had another sample of 1,500 Samplestan adults been interviewed between March 6 and 8, the estimate of President Smith's approval rating might have been 55, 57, or some other number in that vicinity. While surveys can allow us to infer the value of some parameter at the population level, they do so only with a degree of uncertainty. Similarly, a given DHS survey does not tell us the exact value of population parameters such as the contraceptive prevalence rate. Rather, it provides an estimate and multiple DHS surveys at the same time in the same country would not necessarily yield the same estimate value for contraceptive prevalence.

Given that surveys like political polls or DHSs provide only estimates of population parameters, we can never really know with certainty the true value of a population parameter from them or how much a given survey estimate of such a parameter differs from the true population value. How, then, can we assess the quality of a survey?

There are two ways of thinking about this: whether the survey yields "statistically appropriate" estimates of the population parameters of interest, and whether it is efficient. We begin with "statistical appropriateness." Survey estimates are random variables because they depend on the particular composition of a sample which (if sampling is being done correctly) is itself random. For example, if numerous DHS surveys were conducted in a country at exactly the same time and using exactly the same methods it is likely that the estimates of some population parameters (e.g. contraceptive prevalence) might vary between them due to random variation in the distribution of types of individuals between samples.

It seems reasonable to hope some average of such alternative estimates might provide the true population parameter. This is the core idea behind the concept of unbiasedness. To begin with, let us formalize exactly what generates these estimates for the purposes of considering their properties. An estimator is a method for generating an estimate of a population parameter. For present purposes we consider the entire process of sample selection, information collection from the sample (i.e. interview), and parameter estimate computation as the "estimator".[4] So, for instance, we consider the entire DHS survey process, from sampling to indicator estimate computation, to be an estimator.

An estimator is unbiased if the expected value of the estimates generated by it equals the true population value for the parameter it seeks to estimate. In simpler terms, an estimator is unbiased if it is "right on average." For instance, it is unbiased if the average across many, many estimates generated by that estimator equals the actual true value for the population parameter.[5] One would hope that the sampling process would be unbiased in the sense that the sampling phase of the survey yields a sample that can support unbiased estimation of population parameters.[6]

Turning to efficiency, let us first expand a bit on the idea that different samples will generate different estimates of a population parameter. For instance, had Galluper selected several samples over the March 6-8 period in the same fashion, they likely would have yielded slightly different estimates of net presidential approval. This sample to sample variation reflects generally small, random differences in the distribution of types of individuals across samples. The resulting variation in estimate values from sample to sample is referred to as sampling variation.

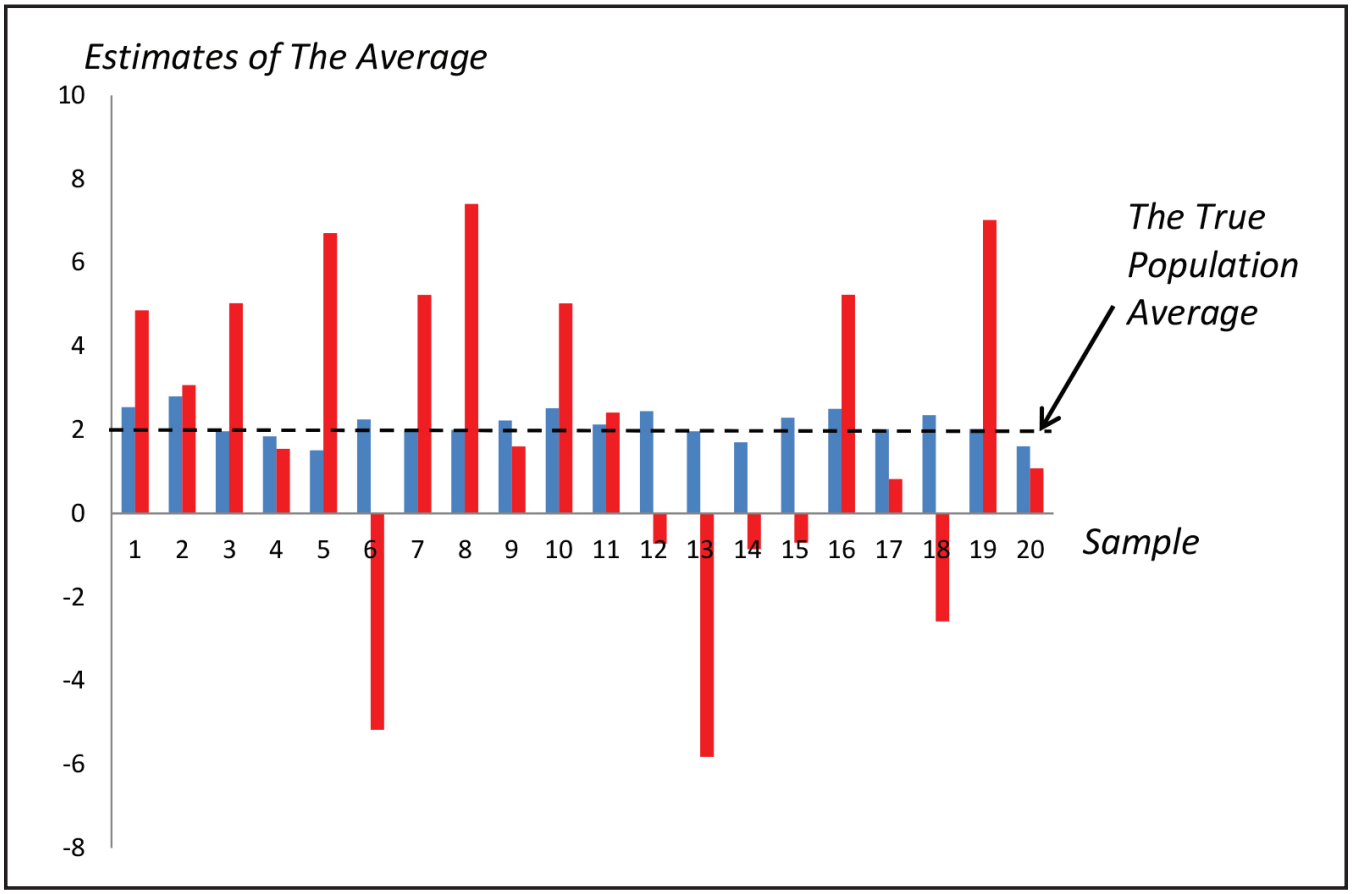
[4] This is a more expansive definition of an estimator than in most econometrics or statistics texts, which tend to assume that sampling and interviewing were done appropriately, and hence focus on methods for computing estimates of population parameters given the data yielded by sampling and interviews of the selected sample.

[5] There are other concepts of statistical appropriateness such as consistency. An estimator is consistent if the distribution of the possible values for the estimates generated by it becomes increasingly concentrated on the true population value as the sample size grows. For present purposes, the relatively simpler concept of unbiasedness is perhaps most useful.

[6] Of course, unbiased sampling is a necessary but not sufficient condition for an unbiased estimate of a population parameter since one can still screw up either the interview or estimation phase of the survey work.

Suppose that we had two different estimators for a population parameter, both of which were unbiased. However, one of them exhibits less variation in the estimates generated from sample to sample. In other words, while both might be "right on average" estimates generated by one tend to differ from the true population parameter value by less than those generated by the other estimator. The estimator that produced estimate values that vary less around the true population parameter value is the more efficient estimator.

Figure 1. Efficiency



The figure above provides a visual illustration of the concept of efficiency. It is a bar graph detailing the estimates of the average of a variable across 20 samples for two different estimators. The samples are numbered on the horizontal axis. The various estimates of the average for the two different estimators are provided by the red and blue bars (i.e., the red bars provide the stream of estimates of the average across 20 samples provided by one estimator, with the blue bars providing estimates across 20 samples for the other estimator). The true population average of the variable is 2, which is indicated with the black dotted line. The two estimators are unbiased, and even across just 20 samples in each case the average of their estimates of the population average for the variable hover around 2 (at 2.13375 in the case of the estimator indicated by the blue lines and 2.058149 for the estimator indicated by the red lines).

However, it is evident that, however similar the average of their estimates across 20 samples, they do not exhibit the same degree of variability in their estimates of the average from sample to sample. The "red" estimates of the average vary far more wildly from sample to sample than the "blue" estimates. This would indicate that the "blue" estimator is likely the more efficient one in that it produces far less sample to sample variation.

There are numerous ways that one estimator can be more efficient than another. Perhaps at the simplest level (and given our rather expansive definition of estimator), it could merely reflect differences in sample sizes selected.

Larger sample sizes generally provide more information and, hence, in principle allow for more precise estimates of population parameters. Thus, for instance, it could be that the "red" estimates are simply based on a smaller sample size than the "blue" ones. However, even for the same given sample size there can be different estimate calculation approaches. Sometimes, one calculation approach involves more variation in estimates from one sample to another.

It is also the case, however, that some sample selection methods can yield more efficient (i.e. precise) estimates than others in the sense of diminished sample to sample variation in estimates. For instance, it could be that the samples behind the "red" estimates were selected in a fashion likely to yield more sample to sample variation in estimates than those samples generated under the "blue" sample selection methodology.

Since our focus is on sampling, when we discuss efficiency our primary interest will lie with the ways that different sampling approaches inform efficiency. We will discuss basic principles of efficiency in sampling as this chapter unfolds. We begin, however, with a simple introduction to basic sampling concepts.

We have learned thus far that population surveys involve the selection of samples from the population of interest in order to learn something about that population. Thus, out of a potentially vast population, a small number of its members are selected in the hope that their behaviors, outcomes and characteristics are indicative of those of the overall population. The process of learning something about the larger population from the (potentially much) smaller sample from it is called statistical inference. In other words, we infer something about the population from the sample through the construction of estimates known as statistics.

Sampling refers at a minimum to the process by which a sample is selected from a population for a population survey. In practice, the term sampling has come to encompass certain other operations (such as probability weight calculations) potentially required to support the calculation of unbiased estimates of a given population parameter. We will discuss sample selection, as well as some of those follow-on tasks, to provide a sense of what is required in terms of sampling to support unbiased estimation of population parameters.

In order for the experiences and characteristics of the sample to provide a useful reflection of the true profile of the experiences and characteristics for the entire population (and thus allow for unbiased estimation of a population parameter), it is necessary for that sample to be representative of the population. For present purposes, this means, in essence, that the sample needs to reflect the population in the sense that the distribution of types of individuals in it should be similar to that in the population from which it is drawn. For instance, if a given subpopulation is largely male and poor, a representative sample from it should be as well. Later, we will expand a bit on our concept of representativeness.

There can be some random discrepancy between the distribution of types in the population and in a particular sample from it. However, sampling that allows for unbiased estimation of population parameters requires that the samples drawn have, on average, distributions of types within them that match that in the population. It is in this sense that a sample is representative: it is selected via a process that insures that, on average, the distribution of types within the samples is the same as that in the population.

Consider separate estimation of average personal income from two representative samples selected at the same time and by the same means. Compared with the distribution of types in the overall population, one of these samples has slightly more high earning and slightly fewer low earning individuals, while the reverse is true for the other sample. All other things being equal, the sample with a greater relative representation of high earners will probably yield an estimate of population average personal income slightly greater than that yielded by the other sample or than the true value at the population level. The process of selecting these random samples will still yield an unbiased (i.e., "right on average") estimate of average income as long as the average representation of different types of earners across samples matches on average that in the population.

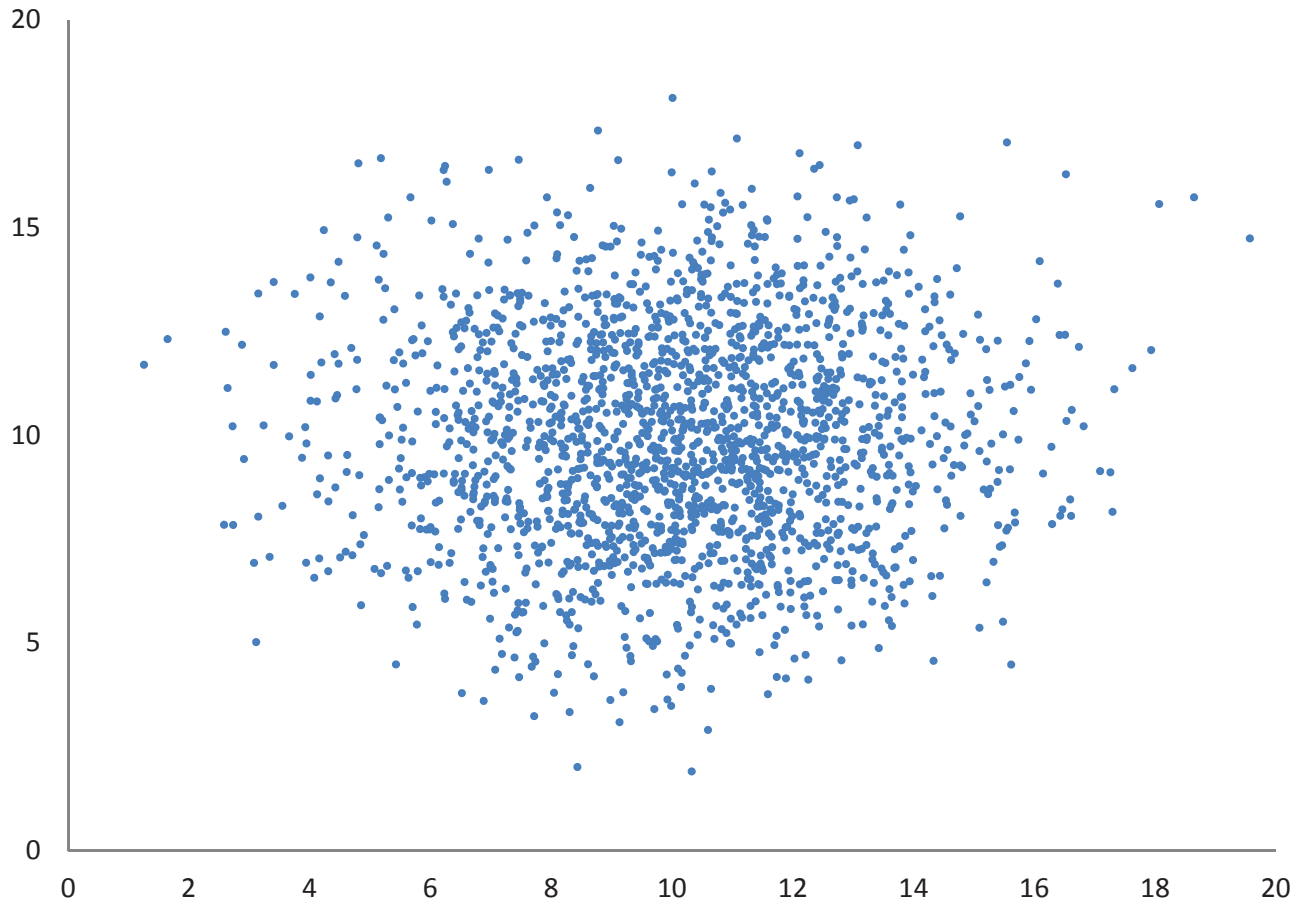
This can be contrasted with a sampling process where, for instance, the samples yielded persistently have a larger proportion of high earners and lower proportion of low earners than their true respective shares in the population. Were one to estimate population average personal income with such samples by averaging observed personal income

within them, the resulting estimates would be persistently higher than the true population value. This estimator (again, using our broad definition of estimator that includes the sampling process) would be biased (i.e., "wrong on average") because of the skewed samples.

The key problem in this case of biasedness is that the samples were not representative. They persistently deviated in the distribution of types of individuals within them from the true distribution across the population, thus providing a distorted representation of the population with the result that samples from them provided misleading inferences about the true value of parameters for that population. Representative sampling is a necessary (though not sufficient[7]) condition for unbiased estimation of population parameters.

To get a better grasp on representative and unrepresentative sampling, let us consider a simple graphical exercise laid out in the three figures that follow. In the first figure we see the scatter plot of a population in terms of two characteristics, one captured on the horizontal axis and the other on the vertical. For instance, suppose that the population is wage earning workers. The horizontal axis could then be hourly wages and the vertical axis might be length of commute in minutes. Each blue dot represents a member of the population, and indicates their wage and commute time.

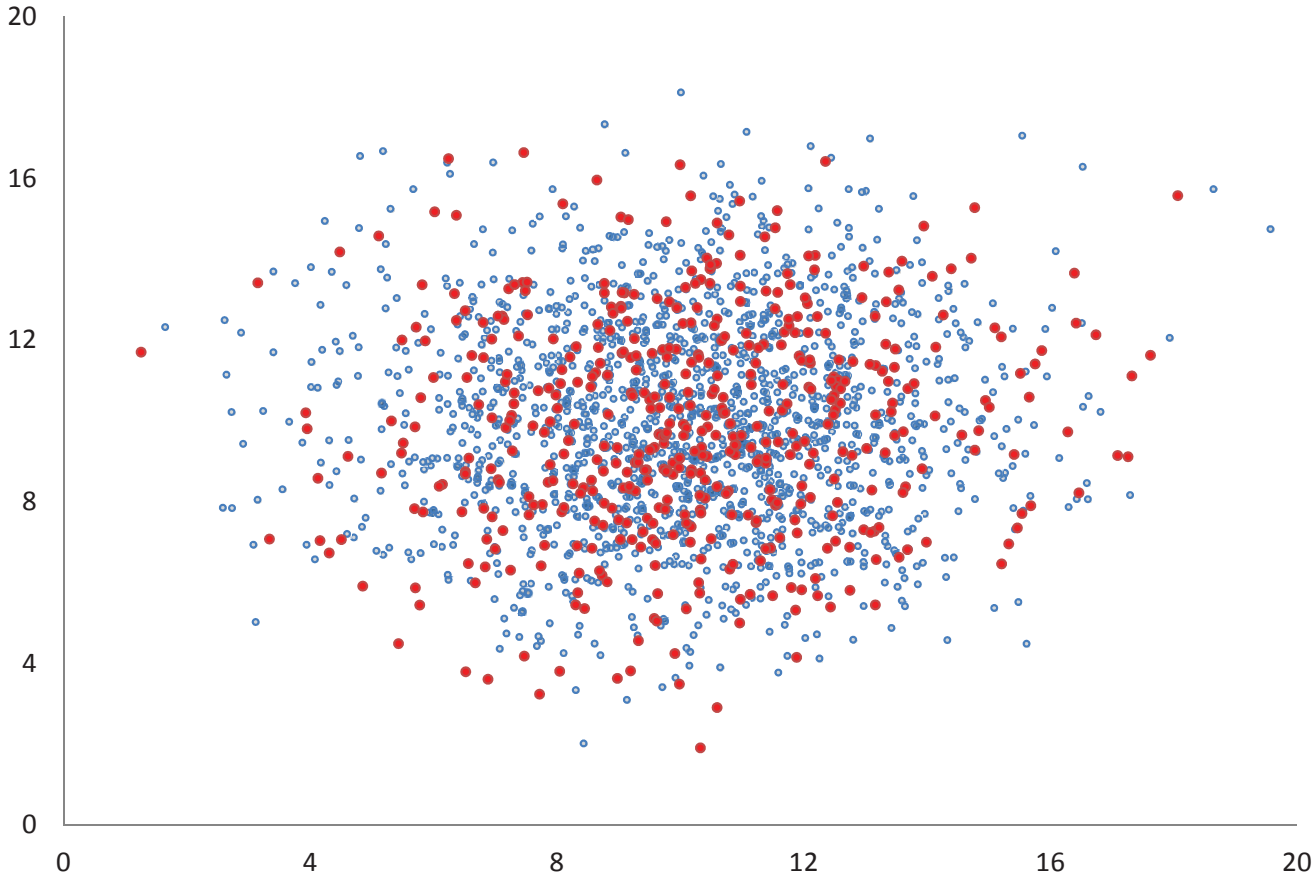
Figure 2. The Population



[7] Again, even if representative sampling has occurred, it still might not be possible to estimate population parameters. Interview and estimate calculation methods must still be appropriate as well.

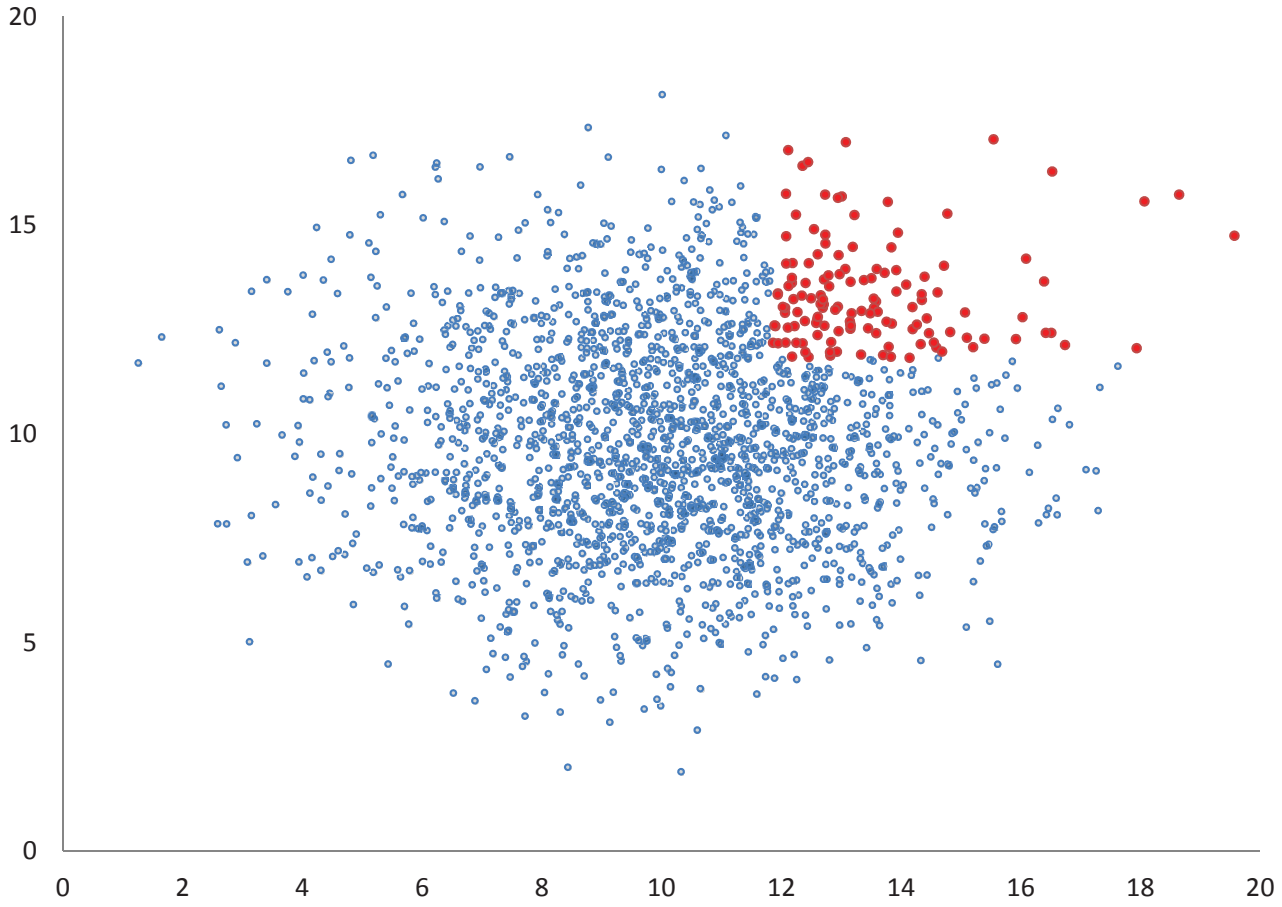
The next figure illustrates a random, representative sample from this population. The selections are indicated with red dots, with the remaining, unselected members of the population presented with faded blue dots. Notice that the spatial distribution of the red dots appears to closely match that of the overall population. Representative selection thus implies that the distribution of types in the sample reflects that in the population. Put differently, the sample is an accurate representation of the population from which it is drawn.

Figure 3. A Random, Representative Sample from the Population



Finally, in the third figure, we provide an example of non-representative sampling. Once again, the selections are indicated with red dots, with the remaining, unselected members of the population presented with faded blue dots. Notice that the distribution of red dots does not really follow the same pattern as the overall distribution of dots for the population. Indeed, compared with the population as a whole, the selected observations have higher wages and commute times than the full population. This would present a serious obstacle to unbiased estimation of average wages or commute times for the population with this sample.

Figure 4. A Non-random, Non-representative Sample from the Population



So how does one insure the representativeness of samples selected from a population? While there can be many facets to a discussion of this question, at the most basic level, it is necessary that the sample be selected randomly. Random selection can be contrasted with non-random selection, under which selection is not indiscriminate. Typically, under non-random selection particular types in the populations are targeted (explicitly or implicitly) for selection, leading to an overrepresentation of such types in the sample relative to their share in the population. For instance, in the figure above the selected workers were those whose hourly wages and commute times both exceeded some threshold.

A major problem with non-random selection is that it generally provides no obvious mechanism for satisfactorily correcting any distortions to the representativeness of the sample as a result of non-random sampling. If certain types have a far greater share in a non-randomly selected sample than their true representation in the population, it is typically not clear how one would either detect this problem or satisfactorily remedy it when the sample was non-randomly selected.

Of course, random selection alone cannot always guarantee representativeness. Suppose, for instance, that one were to conduct a survey of population attitudes for a neighborhood regarding a proposed street widening project in that locale. For it, a sample of households is selected randomly from a list of households in the neighborhood. Each selected household is visited and whichever member is encountered first is asked about their attitude regarding the street widening project. An obvious problem is that those living in larger households have a lower probability of inclusion in the sample since they are less likely to be the ones answering the door: for example, each member of a selected four-person household has in principle a 25 percent chance of selection, while those in selected single person households have a 100 percent chance of selection. The sample of respondents will thus have too many respondents from smaller households relative to the true share of residents of smaller households in the neighborhood population.

It could be, for example, that individuals from large households tend to view the road widening considerably less favorably than those from small households. A sample selection process for which those from smaller households (who tend to look with favor on the project) have a higher probability of selection could tend to provide a skewed sense of the true population perception of the widening project.

As we will see, unless the information required to understand each respondent's true probability of selection is recorded and used to guide the construction of probability weights (which can correct for the distorted representativeness of a sample due to uneven probabilities of selection for individuals in the population of interest), there is no way to correct for this distorted representativeness of the sample. Probability weights correct for uneven probabilities of selection from the population by assigning more weight to the information provided by those with a lower probability of selection. Continuing the road widening project example, with the application of probability weights, the information provided by those from larger households would contribute more to estimating average population approval of the project than that provided by those from smaller households.

Explicit probability weights correct for the possibility that some members of the population have a higher probability of selection into the sample than others. Of course, the potential problems associated with uneven selection probabilities would disappear if all members of the population had the same probability of selection. In this instance, the expected proportional representation of different types in a sample would match their share of the population. In this case, no group has distorted representation (compared with their actual prevalence in the population) and everyone in the sample should contribute equally to computing statistical estimates of population parameters. Such a situation is referred to as a self-weighting sample (as in one that does not require explicit construction of probability weights and instead assigns the same weight to every member of the sample).

We have now developed a broader conceptual approach to what a representative sample means. A representative sample is now simply one for which the estimation of population parameters can be achieved without exaggerating the importance (in terms of the information they contribute) of any individual relative to their actual share of the population. In the case of a self-weighting sample, this is insured by the fact that no individual member of the population had a greater or lesser probability of selection into the sample than any other. In a non-self-weighting sample this is achieved by applying probability weights to the estimation of population parameters.

An emerging point is that random sampling is a necessary but not sufficient condition for samples that can be used to support unbiased estimation of population parameters. The key requirement for random sampling that can support unbiased estimation of population parameters is that it conform to the principles of probability sampling, the requirements for which are met if a sample is selected randomly and:

1. Every member of the population has a positive probability of random selection (i.e., they have at least some chance of being selected); and
2. Their probability of selection is known.

If probability sampling has been satisfied then unbiased estimation of population parameters is generally possible, at least from the sampling standpoint. Even if the sample per se presents a distorted picture of the true population (in the sense that different individuals from the population had different probabilities of selection), probability sampling makes possible the construction of probability weights which can be used to correct that distortion.

There are, to be sure, non-probability sampling methods. A few examples include:

1. Convenience sampling — Members of the sample are selected based on relative ease of selection. The classic example is opinion polling of passers-by in shopping malls;
2. Snowball sampling — This is basically referral sampling, where the first selected survey respondent refers or otherwise somehow leads surveyors to the next respondent;
3. Quota sampling — The goal in this case is to sample a certain number or, more usually, proportion from a population, with the actual selection achieved by some sort of non-random means.

Notice that each of these in some fashion or another fail the two key standards for probability sampling.

In non-probability sampling, the chance of selection of some or all individuals can be unknown or zero. Because of this, non-probability sampling, including these examples, cannot ensure truly representative samples that can support unbiased estimates. First, if some individuals from the population have zero probability of selection, it is not possible for them to appear in any resulting sample. The sample thus cannot be representative of the population since some types of individuals from the population cannot appear in it. Second, even if this is not the case, if one does not know the probability of selection for each individual in the sample, then there is no mechanism to assess whether all members of the population had the same probability of selection or, if they didn't, to correct for the possibility that the sample is a distorted representation of the population since there is no information (regarding the true probability of selection) with which to correct for uneven probabilities of selection across the population.

With the introduction of probability weights and probability sampling, we in some sense complete our journey to understand the basic properties of a sample selection process that can support unbiased estimation of population parameters. In short, this requires that the sampling process generate samples that are representative, in which case the weight assigned to the information provided by each individual in the sample to calculate estimates of population statistics reflects that individual's share in the population. Probability sampling is crucial since it provides a means to assess whether the sampling process involved equal probability of selection into the sample across the population (in which case the resulting sample is self-weighting) or whether there were unequal probabilities of selection across the population (in which case explicit probability weights will be required to support unbiased estimation of population parameters). In the latter case, probability sampling insures (as we will see shortly) that we have sufficient information to compute probability weights.

The next obvious question then is how to select a sample by means of probability sampling (and, by extension, how to utilize the information from probability sampling to construct appropriate probability weights in the event that the sample is not self-weighting). Typically, selection of samples by probability sampling is done by selecting the sample from a list that somehow contains all members of the population. Let us start with a simple case.

Suppose that our goal is to estimate the marriage rate and total personal income for the population of adult women living in a city and we have determined that a sample of 5,000 women is required to estimate these with the desired degree of precision (i.e., desired degree of sampling variation).[8] Suppose as well that there was an accurate list of all 15,000 adult women in that city. In other words, there were no adult women in the city left off the list and no women on the list who do not reside in the city (or are not alive). Finally, we assume that there is only one entry on the list for each woman in this city.[9]

This list, from which we can select a sample of women, is called a sampling frame. A sampling frame is a list of sampling units from a population. If the population in question is children under age 5, then the children under age 5 are the sampling units. If the population is households in a community, then the sampling unit is the household. In our case, the population is of women across the city, hence the sampling units in the frame are women.

[8] Sample size determination is a separate topic beyond the scope of this manual.

[9] Alternatively, we could assume that any duplicate entries that do exist for any particular woman can be easily identified and corrected.

The characteristics of a sampling frame determine whether probability sampling is possible. We have already mentioned the first property necessary for this list to support probability sampling (i.e., that it is an exhaustive list of all of the women in the city, which insures that no woman is left off the list and thus has zero probability of selection). The fact that there is one entry per woman naturally insures that the second condition (a known probability of selection) is met. Thus, to be absolutely clear, a frame that can support probability sampling must provide:

1. An exhaustive list of all sampling units from the population of sampling units, insuring that no sampling unit has zero probability of selection; and
2. The opportunity to select a sample from that list with each unit having a known probability of selection.

Thus, the most crucial qualities in a good sampling frame are that it must exhaustively and uniquely list all sampling units from a population. Uniqueness insures that it is possible to know each unit's probability of selection from the list.

Recall our assumption that there were no adult women in the city left off the list and no women on the list who do not reside in the city (or are not alive). Per the conditions for probability sampling, it is the first of these assumptions (that no adult women were omitted from the list) that is critical. If some women in the city were omitted from the list, those women would not be subject to selection from the list and hence have zero probability of selection. The second assumption, that every woman on the list is indeed a living female resident of the city, is less important and simply insures that all 5,000 women selected will in fact be eligible for interview.

For practical reasons (one must be able to find selected sampling units to interview them), a good frame must also contain the information necessary for us to find and identify selected units (women) to interview. Table 2 illustrates our hypothetical frame. It not only lists the women, but also provides their addresses, making it possible to find and identify them for interview.

Table 2. The Frame of All Women in the City

Woman Number	Woman's Name	Woman's Address
1	Jennifer French	204 Dewar Street
2	Emily Addison	337 Avenue A
3	Elizabeth Blue	214c Park Street
4	Carolyn Smith	25 Herdon Avenue
5	Anne Hurley	1644 Highway 251
6	Rebeccah Jeffers	88 Lacebark Lane
7	Barbara Gingrich	107 Ipswich Street
8	Patricia Clarkson	202 Morgan Avenue
…	…	…
15000	Beatrice Mandel	1414 Highway 251

Since every woman in the city has one entry on the list, the most straightforward way of selecting the sample of women would be via the equal probability of selection method (or epsem, for short).[10] Epsem sampling is fairly straightforward and can represent a fallback position when other options (particularly size sampling, to be discussed below) are not possible. To begin with, we need to calculate what is known as the sampling interval, which is just the

[10] There are multiple approaches to epsem sampling, in addition to the use of a sampling interval as introduced here. One such example is Bernoulli sampling, which randomly assigns a value from a uniform distribution (0,1) to each unit and selects a unit if its assigned value is smaller than the sampling probability. While the procedures of sampling are different, the common principle of these approaches is that units are selected randomly with equal probability. Interested readers may be referred to Särndal, C.E., Swensson, B., & Wretman, J. (2003). *Model assisted survey sampling*. Springer.

size of the frame (in the epsem case, the number of sampling units in the frame) divided by the number of sampling units to be selected from the frame. In this case the sampling interval is

$$15{,}000/5{,}000 = 3$$

The sampling interval is the number of entries on the list that will separate selections. In this case it is 3, so we will select every third woman. What remains is to determine the first woman to be selected. To do so, we draw a random number between 1 and the sampling interval (3). This is called the random start (RS). Suppose the number selected is 2. This is the first woman selected, as shown in Table 3.

Table 3. The Random Start

Woman Number	Woman's Name	Woman's Address		Selected
1	Jennifer French	204 Dewar Street		No
2	Emily Addison	337 Avenue A	RS (2)	Yes
3	Elizabeth Blue	214c Park Street		
4	Carolyn Smith	25 Herdon Avenue		
5	Anne Hurley	1644 Highway 251		
6	Rebeccah Jeffers	88 Lacebark Lane		
7	Barbara Gingrich	107 Ipswich Street		
8	Patricia Clarkson	202 Morgan Avenue		
.	.	.		
.	.	.		
.	.	.		
15000	Beatrice Mandel	1414 Highway 251		

To select the next woman, we simply add the sampling interval to the random start. Since, the random start is 2 and the sampling interval is 3, this means that the fifth woman on the list will be selected. This is illustrated in Table 4.

Table 4. The Second Selection

Woman Number	Woman's Name	Woman's Address		Selected
1	Jennifer French	204 Dewar Street		No
2	Emily Addison	337 Avenue A	RS (2)	Yes
3	Elizabeth Blue	214c Park Street		No
4	Carolyn Smith	25 Herdon Avenue		No
5	Anne Hurley	1644 Highway 251	RS+SI=2+3=5	Yes
6	Rebeccah Jeffers	88 Lacebark Lane		
7	Barbara Gingrich	107 Ipswich Street		
8	Patricia Clarkson	202 Morgan Avenue		
.	.	.		
.	.	.		
.	.	.		
15000	Beatrice Mandel	1414 Highway 251		

The third woman selected is simply the woman one sampling interval ahead of the second woman selected. Since the second woman selected was the fifth woman on the list, this means that the third woman selected is the $5+3=8^{th}$ woman on the list. This is illustrated in Table 5.

Table 5. The Third Selection Onward

Woman Number	Woman's Name	Woman's Address		Selected
1	Jennifer French	204 Dewar Street		No
2	Emily Addison	337 Avenue A	RS =2	Yes
3	Elizabeth Blue	214c Park Street		No
4	Carolyn Smith	25 Herdon Avenue		No
5	Anne Hurley	1644 Highway 251	RS+SI=2+3=5	Yes
6	Rebeccah Jeffers	88 Lacebark Lane		No
7	Barbara Gingrich	107 Ipswich Street		No
8	Patricia Clarkson	202 Morgan Avenue	RS+2*SI=2+6=8	Yes
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
15000	Beatrice Mandel	1414 Highway 251		No

Let us look at this in general terms. The second woman selected is in the

$$RS+SI=2+3=5^{th}$$

place on the list while the 3^{rd} woman selected is at the

$$RS+2*SI=2+2*3=8^{th}$$

place on the list. Then, the 4^{th}, 5^{th} and 6^{th} women selected as selected follows:

The 4^{th}: RS=2+3*3=11 (the 11^{th} woman on the list is the 4^{th} selection);

The 5^{th}: RS=2+4*3=14 (the 14^{th} woman on the list is the 5^{th} selection);

The 6^{th}: RS=2+5*3=17 (the 17^{th} woman on the list is the 6^{th} selection);

and so on. Then, in general, the K^{th} selection is the RS+(K-1)*3 entry on the frame.

The basic take-away is that probability sampling is usually done from a list that exhaustively and mutually exclusively enumerates a population. The selection method is then fairly straightforward. First, one must determine the interval that must separate selections given the size of the list and number of selections to be made. Then, within the first such interval on the list, randomly choose a start point. Each successive selection is then made by advancing down the list by the sampling interval. One final point is that in applying this method, it is important that the list is randomly ordered so that there is no "periodicity" in variables of interest that can coincide with the sampling interval. For instance, in this example, suppose that women are enlisted by their personal income and the listing interval coincides with the sampling interval. Depending on the random start, the sample can consist of women of a certain level of personal income that may not represent the population.

The sample thus selected will be what is called self-weighting, which essentially means that everyone shares the same probability weight for the purposes of calculating estimates of population parameters of interest. A probability weight is a weight based on the probability of selection from the population.[11] The probability weight can be thought of as how many women in the population each woman represents statistically in the event that they are selected. To say that each woman has the same probability weight is equivalent to saying that each woman has the same probability of selection into the sample.

Probability weights can be very complex, but at their core, they are simply the inverse of the probability of selection. In other words, if a given sampling unit had a probability P (where $0<P<=1$) of selection into the sample, its probability weight w is

$$w = \frac{1}{P}$$

In our example, all women in the frame have the same probability of selection P:

$$P = \frac{5{,}000}{15{,}000} = \frac{1}{3}$$

Since each woman in the frame and each woman selected to be in the sample has the same probability of selection, P, each woman in the sample will therefore have the same probability weight w:

$$w = \frac{1}{P} = \frac{1}{\frac{5{,}000}{15{,}000}} = \frac{15{,}000}{5{,}000} = 3$$

Each woman thus has the same probability weight of 3.

This constant probability weight implies that each woman in the sample represents three women in the population. For estimation of a rate such as the marriage rate, the weight from a self-weighting sample is not consequential. For instance, for the purpose of calculating a weighted estimate of an average, rate or proportion, it is common for statistical software to "normalize" the probability weight of each woman in the sample by dividing her own probability weight by the total of the probability weights in the sample. In this self-weighting case, this would yield for each woman in the sample a normalized weight of

$$\frac{3}{5{,}000 \cdot 3} = \frac{1}{5{,}000}$$

Note, however, that this normalized weight is simply the weight attached to each woman's contribution to a straightforward average.

[11] There are other sorts of weights. A frequency weight represents the number of cases in a sample that an included case represents. For instance, suppose that five women in the sample have the same marital status and income. Then perhaps the data manager might retain only one of them, with a frequency weight of 5, because she represents five cases just like her. Importance weights are subjective weights that reflect how important the analyst thinks a case should be. Finally, analytical weights are a rarely applied tool in instances where observations are actually averages based on different numbers of cases. These weights are often based on the inverse of the variance, so that observations based on less precise estimates of the average (i.e. higher variance) tend to have smaller weights. Compared with probability weights, none of these are commonly employed.

Suppose, for instance, that the marriage rate is coded as a binary variable M_i that is equal to 1 if the ith woman in the sample is married and 0 otherwise. The marriage rate is then

$$\sum_{i=1}^{5,000}\left(\frac{1}{5,000}\right)\cdot M_i = \sum_{i=1}^{5,000}\frac{M_i}{5,000} = \frac{\sum_{i=1}^{5,000} M_i}{5,000}$$

where

$$\sum_{i=1}^{5,000} M_i$$

is the sum of M_i for the 1st (i=1) to 5,000th (i=5,000) woman in the sample. This is just the simple average of M_i across the 5,000 women in the sample, under which each woman has the same weight:

$$\frac{1}{5,000}$$

In other words, for a self-weighting sample the probability weight is not consequential for estimation of statistics such as averages, rates, proportions, etc., since it happens to be the same as the weight under a simple average.

For estimation of a total such as total income, it is more meaningful. For instance, suppose we were to estimate total personal income across women in the city by adding up the reported personal incomes of all 5,000 women selected for our sample. Those women still constitute only one-third of all of the women in the population. Therefore, the sum of their personal incomes cannot be total personal income across women in the city since it omits the personal incomes of the 10,000 women in the city not selected for our sample. Multiplying the personal income observed for each woman in the sample by her probability weight of 3 and then totaling (or, since this is a self-weighting sample, simply multiplying the total of the personal incomes observed across the women in the sample by 3) corrects for the fact that the women in the sample constitute only one-third of the population.

Even in this simple case, if some adult women in the city were not included the frame it could create serious obstacles to unbiased estimation of either the marriage rate or total income. For either parameter, there is the problem that the excluded women might not be similar to those included on the frame (for instance, they might have a higher marriage rate or lower income than those included on the frame). For the estimation of total income, the exclusion of some women would additionally preclude us from knowing the true size of the population, and hence what proportion of total population income the sample's total income represents. This reinforces the importance of frame completeness in the sense that the frame must include every member of the population of interest.

While this hypothetical example has been a useful way to ease ourselves into the mechanics of sampling frames and sample selection (and to understand the properties essential for a frame to support unbiased estimation), it is highly unrealistic in at least once sense: rarely does one actually have a list of every individual person in a population. The usual solution to this is to select a sample in stages across which various sampling units above the individual are selected until the final such unit selected is small enough that it is feasible to enumerate all of the individuals in it. This is referred to as multi-stage sampling.

For instance, in the typical Demographic and Health Survey (DHS), a sample of census enumeration areas is typically selected from the national list of them.[12] Census enumeration areas are small plots typically drawn to contain roughly

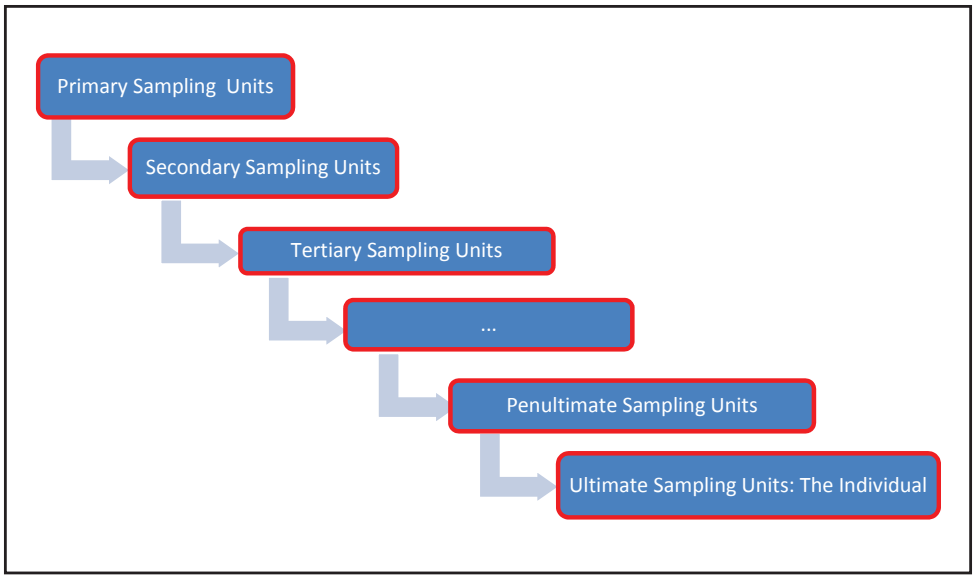
[12] Sometimes selection is directly from the complete list of enumeration areas in a nation, and sometimes it is done from a list that is a randomly drawn (usually by the national statistical agency or census) sample of enumeration areas designed to provide a manageable sized frame for surveys. The latter case (selection of enumeration areas from a list of them randomly preselected from the national list) does not necessarily violate probability sampling because every enumeration area in the nation still had a positive probability of selection into the DHS sample.

50-300 households (depending on the country). Within selected enumeration areas households are then listed in order to construct a sampling frame of households. Then, a sample of households is selected and the women within them are interviewed. The point is that there typically is no national list of women in the population of interest (generally women aged 15-49 in most DHS surveys). Therefore, larger units (first enumeration areas and then households) are selected until one gets down to a unit small enough to enumerate all of the eligible women within it.

In this instance, the first stage of selection is for census enumeration areas while the second stage is for households. Units selected at the first stage of a multistage sampling process are primary sampling units (PSUs). Those selected at the second stage are secondary sampling units (SSUs), and so on.

In principle, there can be many stages of sampling in a multi-stage sampling scheme, as illustrated by the figure below. In some surveys all ultimate sampling units are selected. For instance, in the typical DHS all women in some age range (women aged 15-49) in selected households are interviewed. In this instance, ultimate sampling units have a probability of selection of 1 within selected penultimate sampling units.[13]

Figure 5. Multistage Sampling



To insure that the final sample is capable of supporting unbiased estimation of population parameters, it is necessary that the requirements of probability sampling be satisfied for the population of sampling units at each stage of multistage selection. For instance, in the DHS example above, all census enumeration areas must be included in the first stage frame of census enumeration areas, while all households within selected enumeration areas should be listed for second stage selection.[14]

One other important detail of this example is that the frame of census enumeration areas constitutes what is commonly referred to as an area frame. An area frame is one that divides a physical space (such as a country, region, city, etc.) into a set of mutually exclusive and collectively exhaustive geographic units from which a sample of such units can be

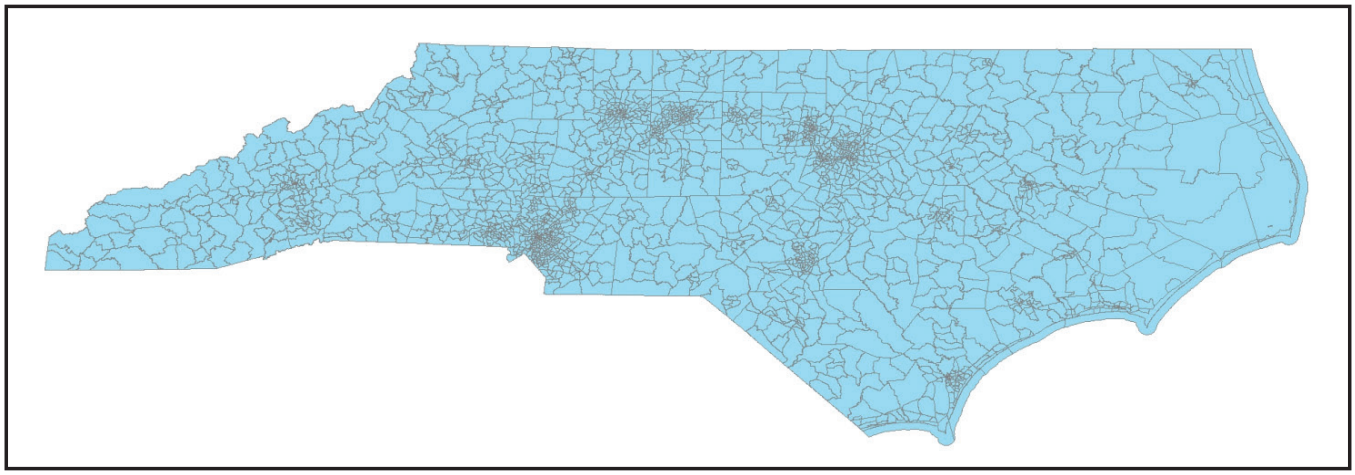
[13] We use terms such as penultimate and ultimate sampling units simply to denote order of selection (second to last and last, respectively) in a multi-stage process of arbitrary length. Suppose, for example, that a survey had three stages of selection: census enumeration areas, the households within selected enumeration areas and, finally, individuals within selected households. Then, census enumeration areas are the primary sampling unit. In terms of the figure illustrating multistage sampling, households would be both the secondary and penultimate sampling units, while individuals would be both the tertiary and ultimate sampling units.

[14] In the case where the national statistical office or census provides a frame comprised of a subset of enumeration areas, this still constitutes probability sampling as long as all enumeration areas still have a positive probability of selection into that subset.

selected. In other words, each portion of a physical space is in one and exactly one sampling unit of any area frame covering it. This is particularly important in this manual, since many of the most interesting ways that GIS can inform sampling are in the context of area frames.

The figure below is a map showing the location of sampling units of an area frame for the authors' home state of North Carolina. Specifically, this is a map of U.S. Census tracts for the state of North Carolina. No two tracts overlap and, collectively, the tracts account for the entire physical area of North Carolina. The Census tracts therefore constitute an area frame for North Carolina.

Figure 6. Census Tract Map of North Carolina



Source: http://www.census.gov/geo/maps-data/data/tiger-line.html

In principle, each stage of a multi-stage sampling process can rely on epsem sample selection. However, there is sometimes another option for selection which can lead to more efficient sampling. By more efficient sampling we mean sampling by means that will lead to less sample by sample variation in the ultimate estimates of the population parameters of interest. This method essentially assigns a higher probability of selection to sampling units that contain more population. It is commonly referred to as probability proportional to size (pps) sampling.

We will illustrate this with a fictional multi-stage sampling process. The situation is motivated by the task of selecting a sample of adults that is representative of the adult population of urban Dhaka, Bangladesh. There is (to our knowledge) no list of adults residing in urban Dhaka in existence. Instead, a sample could be selected by the following multi-stage sampling process:

1. Select a sample of census enumeration areas (EA) from the complete list of EAs for Dhaka;
2. In each selected EA, list households and then select a sample of a fixed number of households from that list by epsem selection;
3. Interview the adults in the selected households.

In this fictional example we assume that we have access to and can select from the full census list of EAs for Dhaka. We present a fictional frame for Dhaka EAs in Table 6.

In this frame, we suppose that there are 5,000 EAs in Dhaka and that, across these EAs, there are 1,450,000 households (both are just arbitrary numbers for this example). For each EA, we indicate the Thana/Upazila (an administrative unit in Bangladesh) and ward (an administrative unit below the Thana/Upazila) in which the EA is located, and the number of households in it per the last national census.

Table 6. A Hypothetical Frame for Dhaka

EA	Thana/Upazila	Ward	Households
1	Dhanmondi	47	225
2	Mohammadpur	45	307
3	Hazarigbagh	58	144
4	Lalbagh	59	397
5	Hazaribagh	48	255
6	Chawkbazar	56	411
7	Shabagh	57	144
8	Chawkbazar	56	505
9	Paltan	36	300
…	…	…	…
4,999	Chawkbazar	56	199
5,000	Motijheel	31	257

The size measure is the number of households in each EA. The first step in sampling with probability proportional to size sampling is to calculate the cumulative size of the frame as of each entry in the frame. This is done in Table 7. Thus, the cumulative size as of EA number 2 is 532 (=225+307), while for EA number 3 it is 676 (=532+144), etc. The cumulative size of the entire frame is simply the cumulative size as of the last (i.e. 5,000th) entry, which is 1,450,000 (note as well that the 4,999th entry has a cumulative size of 1,449,743, which is simply 1,450,000 minus the 257 households in the 5,000th EA).

Table 7. Cumulative Size

EA	Thana/Upazila	Ward	Households	Cumulative
1	Dhanmondi	47	225	225
2	Mohammadpur	45	307	532
3	Hazarigbagh	58	144	676
4	Lalbagh	59	397	1,073
5	Hazaribagh	48	255	1,328
6	Chawkbazar	56	411	1,739
7	Shabagh	57	144	1,883
8	Chawkbazar	56	505	2,388
9	Paltan	36	300	2,688
…	…	…	…	
4,999	Chawkbazar	56	199	1,449,743
5,000	Motijheel	31	257	1,450,000

The next step is to determine the sampling interval. To do this, we need to know the size of the frame (1,450,000 households) and the number of sampling units (for this frame, EAs) to be selected. Suppose that the multi-stage sampling plan calls for the selection of 2,000 EAs at this stage of the selection process.[15] The sampling interval (SI) is 1,450,000/2,000=725. Finally, we must determine a random start within that sampling interval. This means that we must randomly select a number between 1 and the sampling interval (of 725). Suppose that we select 502 (RS=502).

[15] The determination of the number of sampling units to be selected at each stage is, again, beyond the scope of this manual.

This means that the first EA selected in the frame is that EA containing the 502nd household. When we say the "502nd" household, we mean the 502nd household in the cumulative count of households. The 502nd household, so defined, occurs in the second EA on the list. The reason we say this is that the cumulative figure for the first EA on the list is 225 (so the 502nd cumulative household does not fall within that one), hence the first selected EA must be farther down the list than the first EA on the list. The cumulative household number of household for the second EA on the list is 532, meaning that the 502nd household is within that EA given that it was not in the EA that preceded it on the list. This is illustrated in Table 8.

Table 8. The First Selection

EA	Thana/Upazila	Ward	Households	Cumulative		Selected
1	Dhanmondi	47	225	225		No
2	Mohammadpur	45	307	532	RS=502	Yes
3	Hazarigbagh	58	144	676		
4	Lalbagh	59	397	1,073		
5	Hazaribagh	48	255	1,328		
6	Chawkbazar	56	411	1,739		
7	Shabagh	57	144	1,883		
8	Chawkbazar	56	505	2,388		
9	Paltan	36	300	2,688		
…	…	…	…			
4,999	Chawkbazar	56	199	1,449,743		
5,000	Motijheel	31	257	1,450,000		

To make the next selection, we add the sampling interval (SI) to the random start (RS): 725+502=1227. This indicates that the second EA selected should be the EA containing the 1,227th household. This is the fifth EA on the list, as illustrated on Table 9.

Table 9. The Second Selection

EA	Thana/Upazila	Ward	Households	Cumulative		Selected
1	Dhanmondi	47	225	225		No
2	Mohammadpur	45	307	532	RS=502	Yes
3	Hazarigbagh	58	144	676		No
4	Lalbagh	59	397	1,073		No
5	Hazaribagh	48	255	1,328	RS+SI (502+725=1,227)	Yes
6	Chawkbazar	56	411	1,739		
7	Shabagh	57	144	1,883		
8	Chawkbazar	56	505	2,388		
9	Paltan	36	300	2,688		
…	…	…	…			
4,999	Chawkbazar	56	199	1,449,743		
5,000	Motijheel	31	257	1,450,000		

To determine the third EA selected, simply add twice the sampling interval to the random start (RS): 2*725+502=1952. The EA containing the 1,952nd household (on the list, that is EA 8) is selected. The fourth EA selected is then the EA that contains the 3*SI+RS household. In general, the Kth EA selected is that containing the (K-1)*SI+RS household.

As with epsem selection, the probability of selection is at the center of the weight calculation process. For a given EA (let's say EA number *j*) the probability of selection is

$$P_{1j} = \frac{N_1 \cdot S_j}{S}$$

where N_1 is the number of primary sampling units to be selected in the first stage (2,000 in our example), S_j is the size of (i.e., number of households in) EA number *j* and *S* is the overall size of the frame (1,450,000 in this example). P_{1j} simply stands for the first-stage probability of selection for EA *j*.

Once EAs are selected, the households within each selected EA are listed and from each selected EA's household list a sample of a fixed number of households is selected by epsem sampling. Suppose, to fix ideas, that *H* households are selected per selected EA and that there are H_j households listed in EA number *j*. The second stage probability of selection is then

$$P_{2j} = \frac{H}{H_j}$$

where P_{2j} stands for the second-stage probability of selection for EA *j*.

The overall probability of selection of households is then

$$P_j = P_{1j} \cdot P_{2j}$$

In other words, in multi-stage sampling, the overall probability of selection is the product of the probabilities of selection from each stage. Assuming that all adults in each selected household are targeted for interview (that is, assuming that there is no further meaningful "selection" with probability less than 1), the weight for adults interviewed in EA *j* is

$$w_j = \frac{1}{P_j}$$

This would at first glance seem to indicate a different weight for the adults in each EA.

Let us delve a bit more into P_j:

$$P_j = P_{1j} \cdot P_{2j} = \frac{N_1 \cdot S_j}{S} \cdot \frac{H}{H_j}$$

Suppose that the frame was accurate, in that the number of households in each EA actually equaled the number indicated on the frame. Then, $S_j=H_j$ and

$$P_j = P_{1j} \cdot P_{2j} = \frac{N_1 \cdot \cancel{S_j}}{S} \cdot \frac{H}{\cancel{H_j}} = \frac{N_1 \cdot H}{S}$$

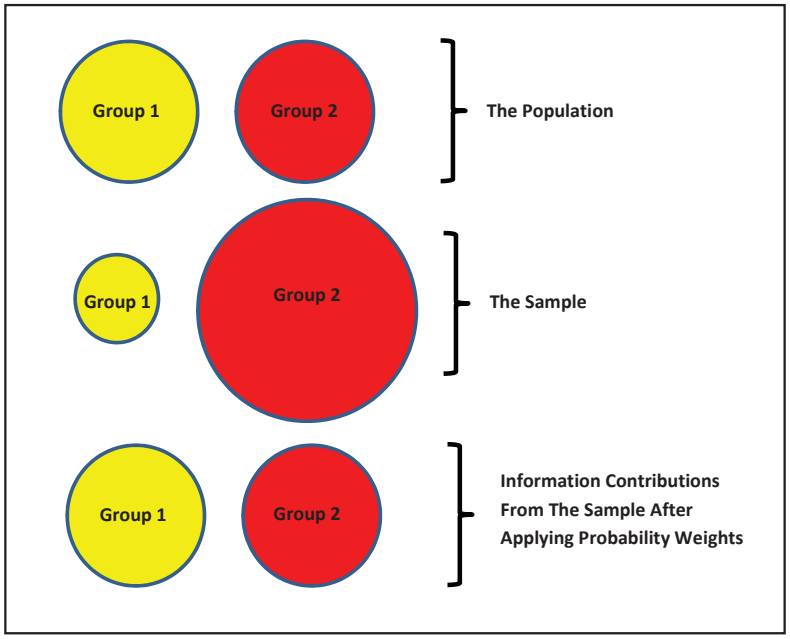
(Recall that N_1 * H is the total number of HHs in the sample and S is the total number of HHs in the sampling frame.) In other words, the probability of selection of households and adults (and hence their weight) would be the same across EAs.

This is the self-weighting case in the framework of frames allowing selection with probability proportional to size (with size measured by the number of households) and it carries essentially the same interpretation as in the epsem case: a self-weighting sample is one in which all interviewed units of observation have the same probability of selection and hence the same weight. Of course, if the size measure in the frame was not accurate then the terms S_j and H_j would not cancel out and the weight for interviewed adults would depend on their EA.

This once again brings up a key point about what we mean by a representative sample and how representative sampling relates to probability sampling. In the case where S_j and H_j do not equal the sample is not by itself necessarily a perfect representation of the population because, ultimately, the individuals (in our case adults in Dhaka) do not all have the same probability of selection. This means that some types of individuals are more likely to be selected than others.

However, the probability weights allow us to correct for any sense in which the sample provides a distorted representation of the population due to the fact that different individuals have different probabilities of selection. For instance, those with a higher probability of selection are overrepresented in the sample, but because their probability of selection is comparatively high they have a comparatively low probability weight. They then contribute less to the estimation of the parameter of interest than their share in the sample would suggest. However, their contribution to the estimation *is*, via probability weights, in line with their share in the population. The weights thus bend the information contribution from the units in the sample back toward true population representativeness.

Figure 7. Probability Weight and Information Contributions



The Figure above provides conceptual illustration of this. Groups 1 and 2 are equally common in the population. However, size sampling (with less than fully accurate size measures) has resulted in a sample in which Group 2 is overrepresented (and Group 1 underrepresented) relative to their shares of the population. However, the effect of applying probability weights in estimation with the sample is to pull the information contributions of the two groups back into line with their shares in the population, even if the sample by itself (i.e., with no probability weights) would provide a distorted representation of the population.

Probability weights are thus critical. Correct probability weights depend on correctly calculated (with reference to the entire population) overall probabilities of selection, which depend in turn on each member of the population actually being in the frame at each stage of selection. This is another stark reminder of the importance of complete frames that include all members of a population.

We conclude with a brief discussion of two other topics of potential relevance to sampling and GIS. First, we consider the topic of stratification. Stratification essentially involves grouping a population into separate (mutually exclusive and collectively exhaustive) frames for different subpopulations in order to sample separately from each of those subpopulations. Stratification effectively implies sampling independently from each strata and hence fixing the sample size from each strata. There are a number of reasons to do this. Sometimes sampling goals (i.e., what one wishes to learn from a survey) are different for different subpopulations (for instance, in a survey of the health of American adults, prostate and breast cancer rates are of greater interest among males and females, respectively). Another popular reason is simply to insure some minimum sample size for smaller subpopulations in order to preserve flexibility for future analyses with the data. Additionally, stratification can reduce sampling variation to the extent that the average value of the outcome differs across the strata.[16]

Whatever the motivation, the process of stratification is fairly straightforward. It usually proceeds in two conceptual steps:

1. From the overall frame, create frames for each subpopulation; and
2. Select samples separately from the frames associated with each subpopulation.

Since the second step (i.e., how to select from a sampling frame) is fairly well established, we focus on the first step.

To fix ideas, suppose that we wanted to be able to estimate some indicator for the adult populations of each Thana/Upazila of Dhaka. In other words, sample selection (and eventual estimation) is to be done independently, or stratified, by Thana/Upazila. To select independently one must develop a separate frame for each Thana/Upazila. This process is illustrated in Table 10 for two Thanas/Upazilas, Chawkbazar and Motijheel.

The process is quite simple. The separate entries for Chawkbazar and Motijheel are collected into distinct and separate lists which then become the new, independent sampling frames for Chawkbazar and Motijheel. Sample selection can then proceed by whatever means the researcher feels is appropriate. The same process is then repeated for every other Thana/Upazila in Dhaka.

[16] The reason for this is actually intuitively rather straightforward. When the average value of an outcome varies across strata, in the absence of stratified sampling one source of variation in the estimate of the average outcome is variation in representation of the strata across samples. Suppose, for instance, that men are heavier than women. In repeated sampling across the population of adults, one reason for variation in the estimate of the average weight of adults is random variation in the shares of men and women across the samples. By fixing the sample sizes from each strata (in the example of the last sentence, by fixing the sample sizes of men and women) one eliminates this source of variation in the estimate from sample to sample.

Table 10. A Hypothetical Frame for Dhaka

EA	Thana/Upazila	Ward	Households
1	Dhanmondi	47	225
2	Mohammadpur	45	307
3	Hazarigbagh	58	144
4	Lalbagh	59	397
5	Hazaribagh	48	255
6	Chawkbazar	56	411
7	Shabagh	57	144
8	Chawkbazar	56	505
9	Paltan	36	300
10	Motijheel	31	224
11	Mohammadpur	45	266
12	Shabagh	57	156
13	Motijheel	31	198
...	...	...	...
4,999	Chawkbazar	56	199
5,000	Motijheel	31	257

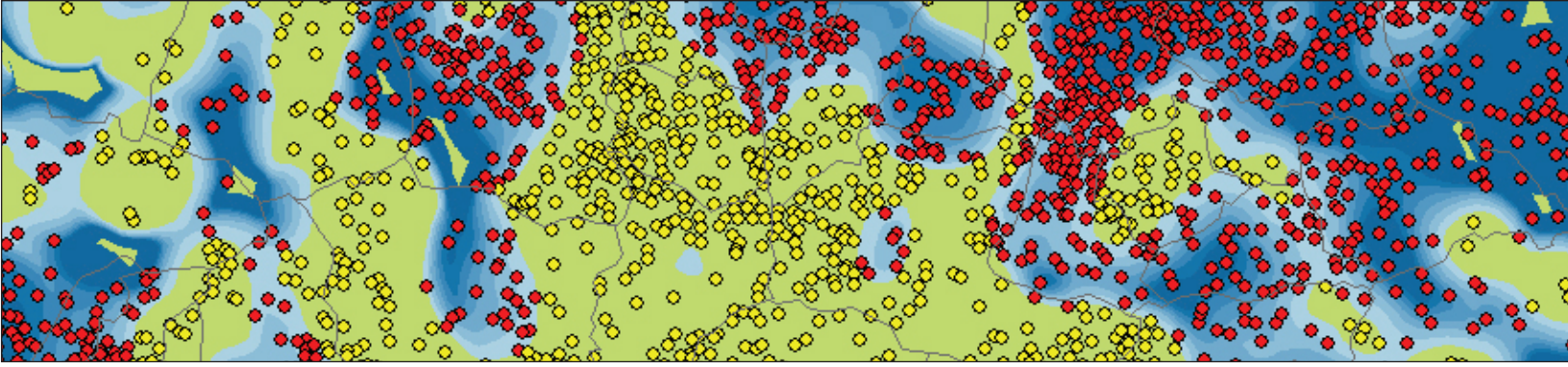
A Hypothetical Frame for Chawkbazar			
EA	Thana/Upazila	Ward	Households
6	Chawkbazar	56	411
8	Chawkbazar	56	505
...	...	...	...
4,999	Chawkbazar	56	199

A Hypothetical Frame for Motijheel			
EA	Thana/Upazila	Ward	Households
10	Motijheel	31	224
13	Motijheel	31	198
...	...	...	...
5,000	Motijheel	31	257

Thus far we have focused on sampling situations where sampling units appear just once in one frame. In practice, selection of a sample sometimes must be done with several frames. A given sampling unit might appear in just one or several of those frames. In the latter case, there are thus several ways that they can be selected.

A common instance of this arises in the context of our original example, political polling. Sample selection for political polling often involves the selection of phone numbers and subsequent interview of whoever answers the selected numbers. The trouble is that there is no one list of all of the phone numbers in most societies. Rather, there are often different lists for alternative telecommunications firms and types of phones. Many polling companies possess (more or less) two lists: one for landlines, and one for mobile phones. Selection typically proceeds from these two lists. However, some individuals possess only a landline or a mobile phone (and hence have a positive probability of selection from just one of the two lists) while others have both. The latter group has a positive probability of selection from either (i.e., land line or mobile) frame.

This is a situation known as overlapping frames. The terminology is motivated by the idea that there are sampling units in both frames. The fact that these units could be selected from either frame with positive probability must be accounted for in the course of weight calculation, sometimes leading to rather complex probability weights. However, to be able to calculate weights correctly, one must know which sampling units appear in which frames (in other words, if a unit appears in more than one frame).



Chapter 3. Geographic Information Systems

In this chapter we introduce and discuss the possibilities presented by a geographic information system (or GIS). In the last chapter we learned about sampling frames, and in particular about the importance of area frames, which parse a study area into mutually exclusive and collectively exhaustive geographic units. Such frames are typically used in stratified sampling and multistage sampling (particularly in the earlier stages of multistage sampling) for selecting samples of individuals or households. The samples of individuals or households can then be used to support estimates of population parameters that can inform programs geared toward influencing human welfare.

Because area frames are based on geographic sampling units, there is a clear natural opportunity for a geospatially oriented framework for information (in other words, a GIS) potentially to enrich understanding of the characteristics and circumstances of the populations in those sampling units, allowing for the identification of vulnerable populations which would not otherwise be straightforward to identify prior to sampling. This, in turn, can allow for far more efficient, effective and focused sampling of many vulnerable populations that are of such increasing interest for programming. However, this is not the only way that GIS can support the process of estimating population parameters of programmatic interest. By providing some sense of the features and possibilities of GIS, this chapter lays the other key piece of groundwork for subsequent discussion of how GIS can inform sampling.

A GIS is a system that integrates software, hardware and data for gathering, managing, analyzing and displaying data using a geographic context. When incorporated into a GIS, the geography associated with data makes it possible to see how phenomena, such as the prevalence of a behavior, outcome or characteristic, represented by the data, relate to their location. When multiple datasets associated with an area are brought together in a GIS, it is possible to explore the relationships between the data through maps or the analysis capabilities of the software. In this way a GIS can provide a multidimensional sense of the spatial patterns of the indicators contained in these datasets.

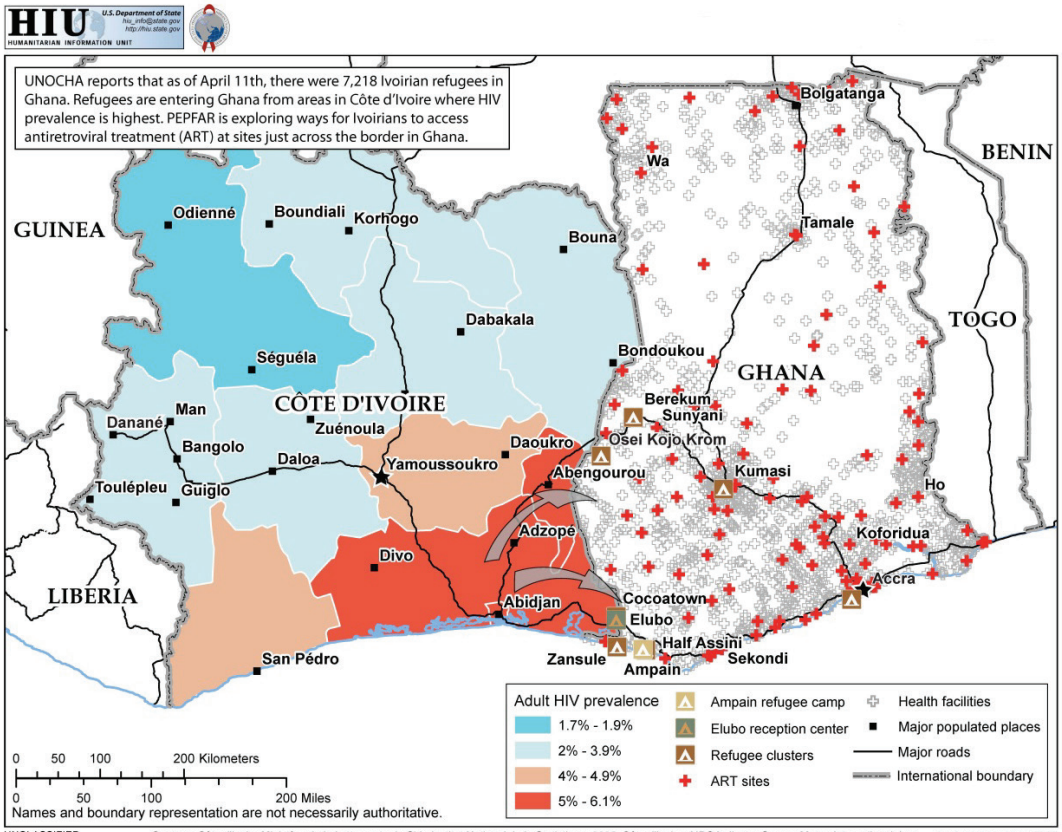
To illustrate, when confronting a complex problem such as ensuring continuation of antiretroviral treatment (ART) for HIV positive refugees there are many facets to consider, including the demand for and supply of necessary health and social services. To understand the potential demand, it is critical to know the characteristics and the scope of the HIV problem among the refugee population, including the HIV prevalence, number of patients, socio-demographic factors, and so on. While it would be ideal to find all necessary information on the refugee population in one data source, such data rarely exists due to their displacement and loss/lack of records, among other factors. When the data doesn't exist, we must rely on other and often multiple alternative datasets. For instance, knowing whether the refugees are coming from high or low HIV prevalence areas may allow for estimation of the HIV prevalence among the refugee population, and it might be possible with the available data. Additionally, it would be important to know something about the destination of the refugees for understanding the supply side of necessary health and social services for HIV positive refugees. For example, are the destination areas equipped to provide ART?

A GIS makes it possible to bring all of those data layers together:

- Location of the refugees' origin and the HIV characteristics of those locations;
- Location of likely destinations and their characteristics;
- Health facilities in the destination area and ART capacity.

The following figure shows a map based on the incorporation of all of those data elements. Produced by the Humanitarian Information Unit at the Department of State, the map illustrates the potential for refugees from Cote d'Ivoire to continue to receive ART in their destination in Ghana. From the programming perspective, this illustrative exercise may suggest that the HIV prevalence among the refugee population is high given that their origin area has an HIV prevalence higher than the other parts of the origin county. Additionally, mapping of the refugee clusters and ART facilities facilitates understanding of the demand and supply for health and social services in particular areas to focus programming efforts and resources.

Figure 8. Cote d'Ivoire Continuity of Care for Refugee ART Patients in Ghana, April 2011



However, GIS is more than just a program that makes maps. To truly be considered a GIS, a software program must also have the ability to build, query and modify data. While maps are the most common output from a GIS, the system can be used effectively without a single map being produced. The strength of GIS is the ability to link disparate data sets using a common frame of geographical or spatial reference. There is a geography to all human activity and that fact can be leveraged to link virtually any data associated with a specific geographic location. This means that a GIS represents a potentially powerful platform for conducting all sorts of analyses. In the subsequent paragraphs, we walk through the basics of GIS features and capabilities.

Data in a GIS is separated into two domains, geographic and attributes. The geographic domain contains the data's reference to the location on Earth. Depending on whether the data represents a point, line or polygon, the geographic domain could represent a single point or a series of points connected to form a line or a polygon. For each element in the geographic domain there is an associated attribute also stored by the GIS. The attribute domain contains the data of interest (i.e., the information about that location). GIS permits queries and analysis of either domain. In other words, it is possible to search for data as it relates to a geographic location or identify locations that correspond to specific aspects of the attribute data.

For instance, if a GIS contained a dataset with school locations in an area, the geographic domain would contain the geographic coordinate of where the school was located and the attribute domain would contain information about the school such as number of students, teachers and so on. It would be possible to query the geographic domain to identify the school that is the farthest from any other school. It would also be possible to separately query the attribute domain to identify the school with the fewest number of students.

The data in a GIS is most frequently conceptualized as layers, with each layer referenced to the same geographic context and representing a separate entity. For instance, a spatial database might contain a layer representing population, a layer representing roads, a layer representing administrative boundaries and so on. GIS permits querying and analysis within and across layers based on either attributes or geographic characteristics. To continue the preceding example, if in a GIS there is a layer for schools and a layer for roads it would be possible to identify which schools were within 1km of a road.

In the preceding discussion of sampling, we learned that it is critical that a sampling frame for a population of interest satisfy certain properties in order to unequivocally support unbiased sampling from that population. For instance, it is important that a sampling frame contains all of the members of a population of interest from which one wished to sample. Similarly, there are factors that contribute to the quality of a GIS, in the process determining whether it can be an asset or liability to applications such as the sampling process.

Perhaps most essentially (and obviously), it is important that data intended for use in a GIS be properly geo-referenced. Geo-referencing data is the act of defining the location associated with the data. Sometimes the geo-referencing can be as basic as including the name of an administrative unit in the data and then linking it via that name to a GIS file whose geographic domain contains validated shapes for the boundaries of the administrative units.

Other times, geo-referencing requires highly precise coordinates collected using GPS or surveying equipment. Regardless of the method used, it is important that the accuracy associated with the data is consistent with the needs of the user. For sampling, higher levels of accuracy are preferable.

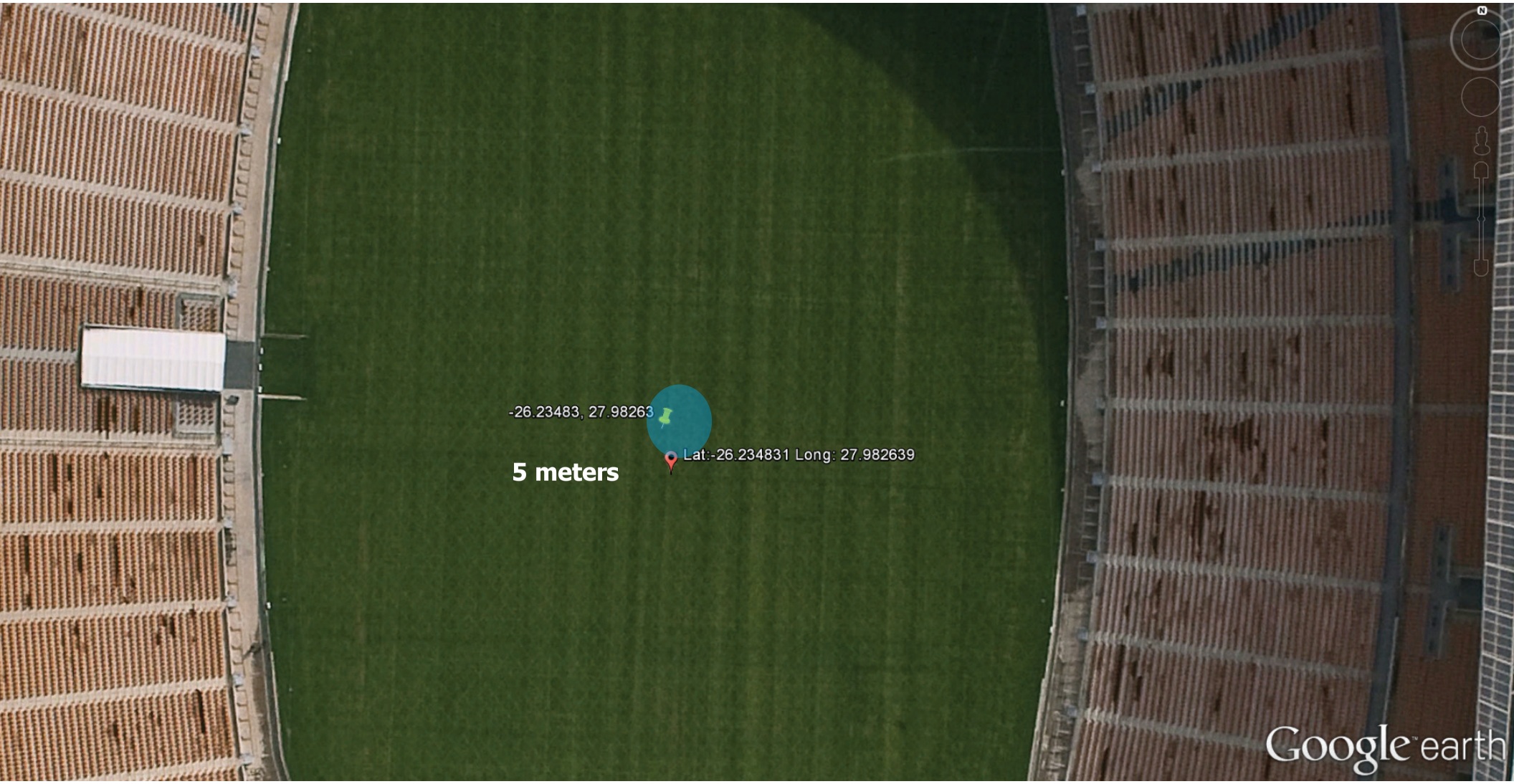
It should be pointed out that accuracy is a function of the numerical precision of the coordinate as well as how far away the coordinate is from the true location of the object the coordinate represents. These issues of precision and accuracy are crucially important. Precision in coordinates is crucial and not including enough significant digits can introduce many kilometers of error.[1]

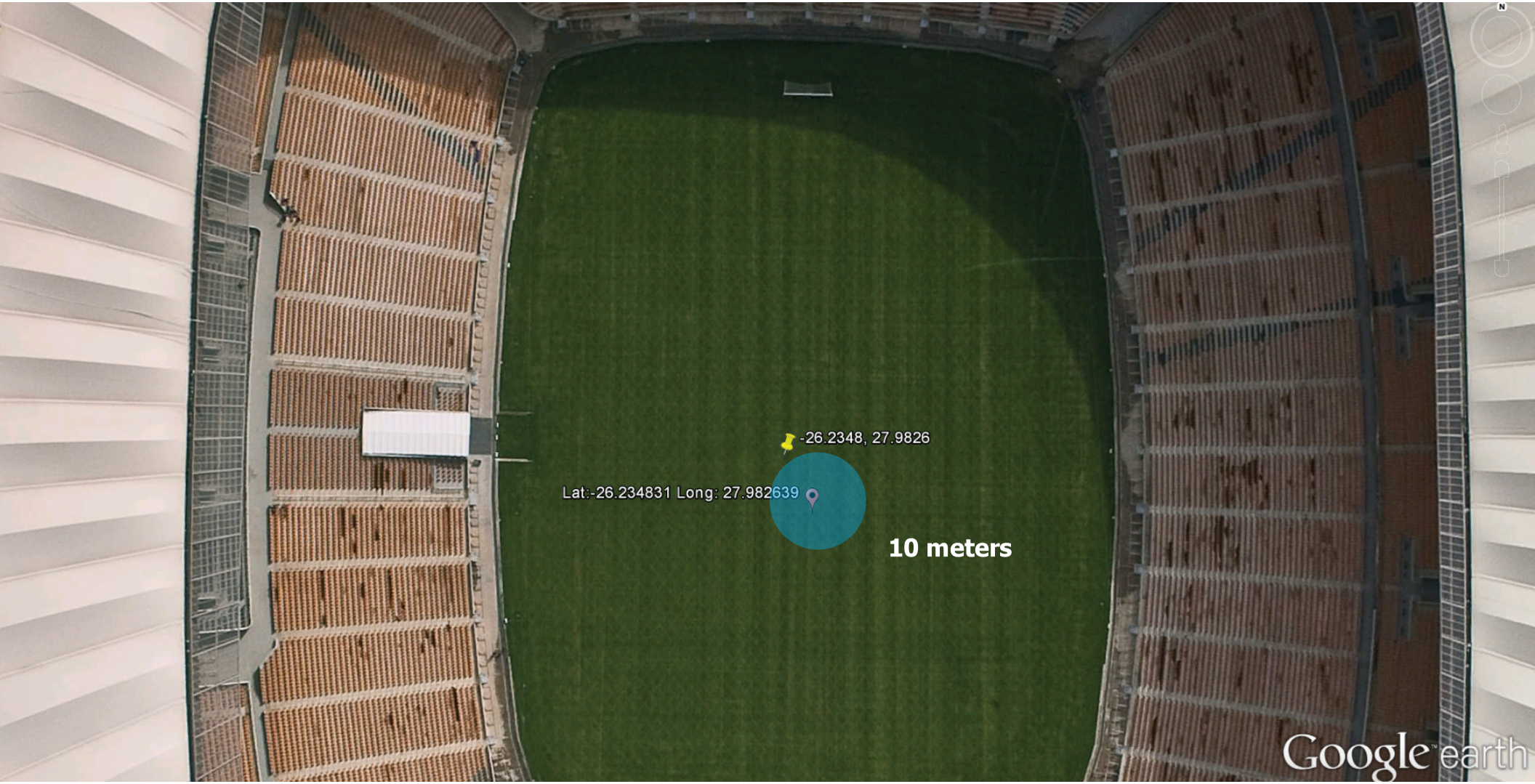
The following graphics illustrate the importance of precision. The first image shows the Latitude/Longitude coordinate of a point mid-pitch of Soccer City, in Johannesburg, site of the 2010 FIFA World Cup as Lat:-26.234831, Long: 27.982639. Dropping the trailing digit so that the coordinate is -26.23483, 27.98263 introduces an error of five meters. Continuing to drop trailing digits results in errors of 10 meters, 115 meters, 600 meters, all the way to 9.1 kilometers for a coordinate of -26.2, 27.9.[2] It is not difficult to imagine that such inaccuracies with coordinates when referring to a health facility or household location could mean that households could be incorrectly assigned to a facility.

In addition to precision, spatial accuracy is also important. Spatial accuracy refers to whether the coordinate accurately reflects the true location of the point on Earth. Errors in accuracy can be introduced because a coordinate was recorded incorrectly on a data collection form or because of errors inherent in the devices used to obtain the coordinate. For instance GPS devices are subject to both random and systematic errors. The effect of these errors can be as small as a few centimeters or potentially many kilometers.

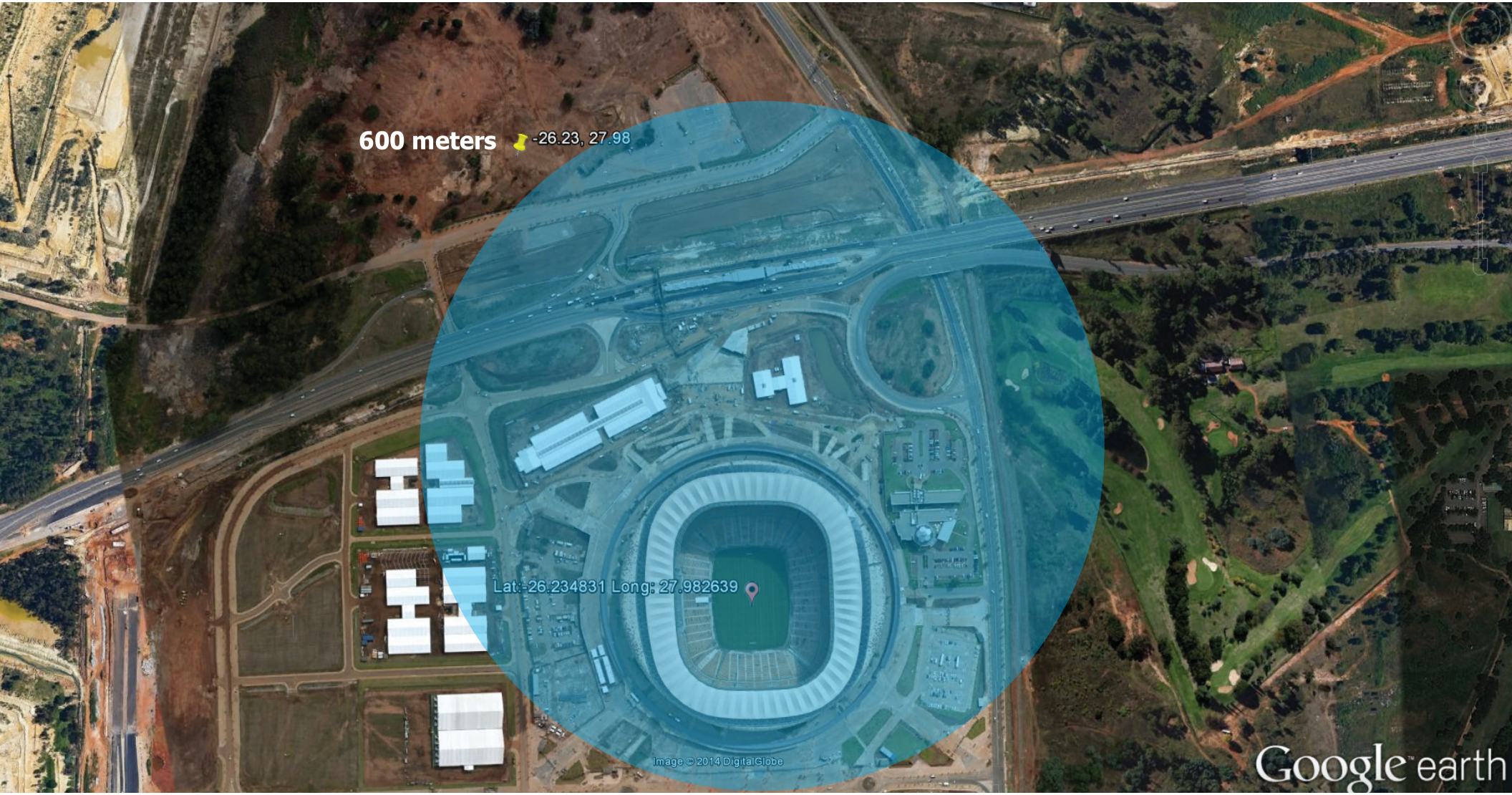
[1] Coordinates can be presented in many different coordinate systems: latitude and longitude are the most common, but there are other coordinate systems that can be found in use in many locations. Latitude and longitude coordinates can be presented in different formats, degrees, minutes, seconds, decimal degrees, etc. In general, GIS software typically requires latitude/longitude expressed in decimal degrees. These are important issues but are beyond the scope of this manuscript. The citations section contains references to documents that provide an overview of specifics about using geographic coordinates.

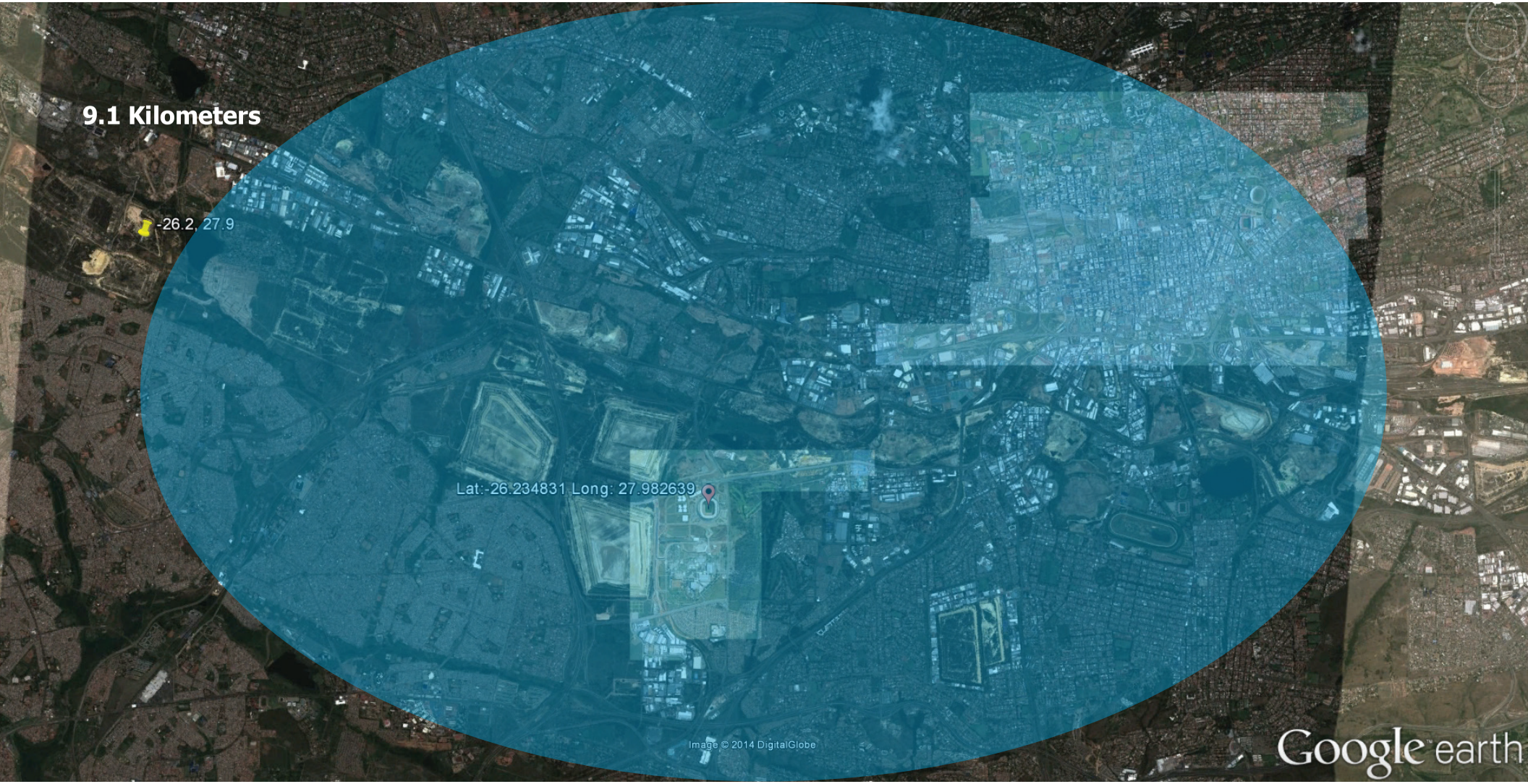
[2] The amount of error will vary depending on the location. Dropping significant digits near the equator will introduce a greater error than dropping significant digits near the poles.


-26.23483, 27.98263
Lat:-26.234831 Long: 27.982639
5 meters
Google earth


-26.2348, 27.9826
Lat:-26.234831 Long: 27.982639
10 meters
Google earth


115 meters
-26.234, 27.982
Lat: -26.234831 Long: 27.982639
Google earth


600 meters
-26.23, 27.98
Lat: -26.234831 Long: 27.982639
Image © 2014 DigitalGlobe
Google earth


9.1 Kilometers
-26.2, 27.9
Lat: -26.234831 Long: 27.982639
Image © 2014 DigitalGlobe
Google earth

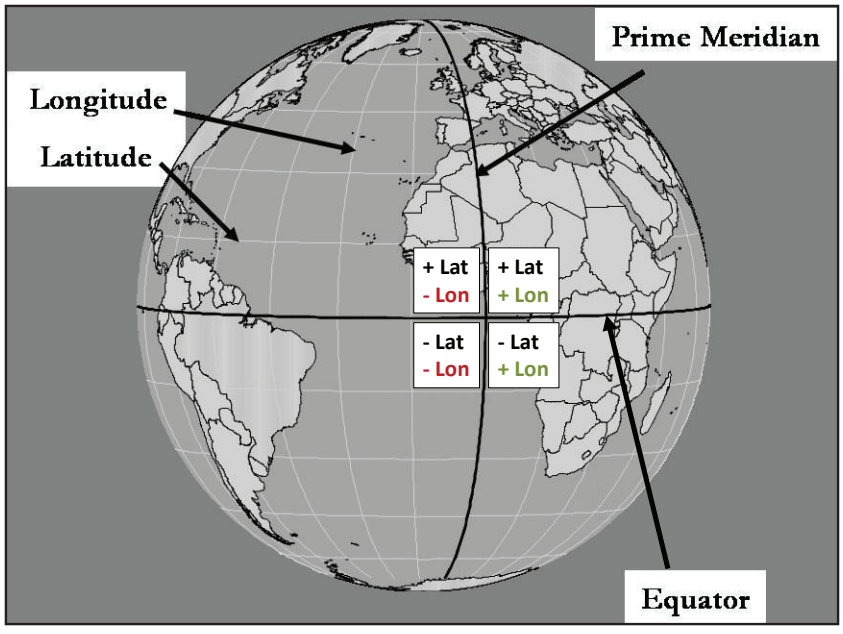
GPS errors can be addressed in several different ways and the best approach will depend on the data collection protocol and the equipment used. Proper data collection protocols can improve spatial accuracy greatly. For instance, making sure the GPS device is receiving a strong signal from at least three satellites will result in a much more accurate coordinate. GPS devices vary but some will provide an indication of the number of satellites from which they have been able to receive signals.

One approach to minimize GPS errors is to collect multiple readings at a location and calculate the average of those readings. This technique, known as point averaging, has been shown to result in coordinates with an accuracy of five meters or greater. Some GPS devices can automatically average points, while in other cases it may be necessary to manually average the points using a GIS. The greater the number of points collected, the greater the accuracy of the averaged points. For devices that do point averaging, collecting one point a second for three minutes (for a total of 180 points) should result in an accuracy of one meter. If the device does not do point averaging, 180 points is impractical but collecting at least three points will minimize error.

Another important consideration involves issues of coordinate systems and datums. Coordinate systems are the reference system that makes it possible for locations to be represented by numbers. There are many different coordinate systems that are in use, some of them global and others specific to a country. One of the more common coordinate systems is latitude and longitude.

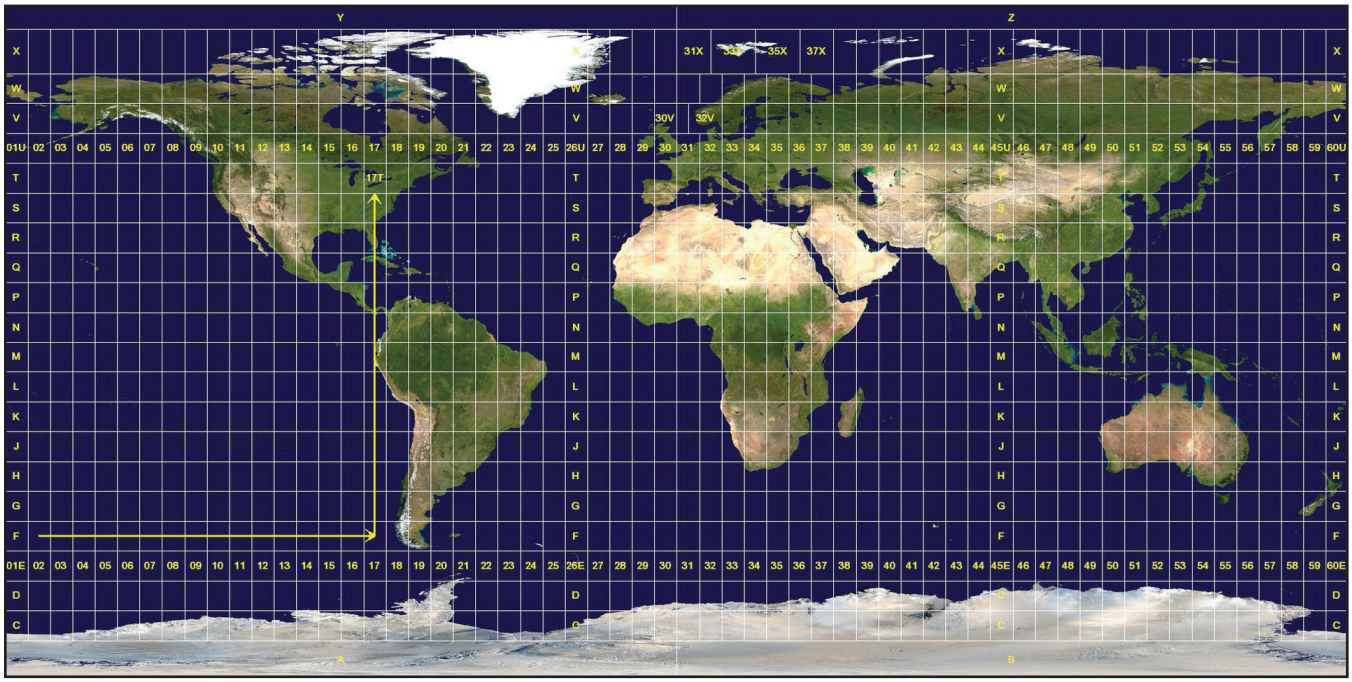
Latitude measures how far north or south a location is from the equator, while longitude measures how far east or west a location is from the Prime Meridian. As illustrated in the graphic below, coordinates north of the equator are recorded as positive numbers while coordinates south of the equator are recorded as negative numbers. Locations east of the Prime Meridian are recorded as positive numbers while negative numbers represent locations west of the Prime Meridian. While latitude and longitude can be represented in several different formats, the preferred format for GIS is decimal degrees.

There are other coordinate systems in use such as Universal Transverse Mercator (UTM). With the UTM system, a series of zones have been applied to the earth's surface and coordinates refer to a location's distance from the origin of one of the zones. UTM coordinates along the X axis of the zone are known as *eastings* and coordinates along the Y axis are known as *northings*.



While some GIS software can accommodate data using different coordinate systems, not every software package can. It is considered best-practice to use a consistent coordinate system across all data layers in a GIS.

In order to be plotted on the earth, every coordinate system needs to be referenced to a datum. Datum is the reference point or surface that corresponds to the specific mathematical model of the earth used when the data was collected. It is becoming increasingly common for coordinates to be stored using the WGS84 datum. This datum is a global datum and as such, is suitable for use anywhere in the world. However, local datums are still in use in many countries.



Source: http://en.wikipedia.org/wiki/Universal_Transverse_Mercator_coordinate_system#mediaviewer/File:Utm-zones.jpg

A local datum uses a local reference point and is best suited for use in the area near the reference point. The graphic below illustrates the effect of displaying a coordinate with a local datum in an area far from the reference point by showing two points at Ohene Djan stadium in Accra, Ghana. One point was collected using WGS84 datum (a global datum), the other point shows the error introduced if that point is added to a GIS with the incorrect datum specified. In this case we illustrate OSGB36, a datum created for use by the Ordinance Survey in Great Britain.

Figure 9. Two Points at Ohene Djan Stadium

Source: Google Earth

The shape of the earth is an ellipsoid—slightly flat at the poles and bulging along the equator. In order to display the spherical Earth on a flat surface such as a piece of paper or computer screen, it is necessary to transform the coordinates to accommodate the different shapes. This process is known as projection.

While a detailed discussion of map projections is beyond the scope of this manual, it is sufficient to point out that issues around projections can affect the way that locations are displayed in a GIS. Locations that have been displayed in one projection may be placed incorrectly if they are overlaid on a layer that is in another projection.

The process of geo-referencing data, collecting and formatting geographic coordinates properly, managing projection and datum issues all require a more detailed discussion than is possible in this manual. More information on using geospatial data can be found in the MEASURE Evaluation Publications *An Overview of Spatial Data Protocols for HIV/AIDS Activities* and *An Overview of Spatial Data Protocols for Family Planning Activities.*

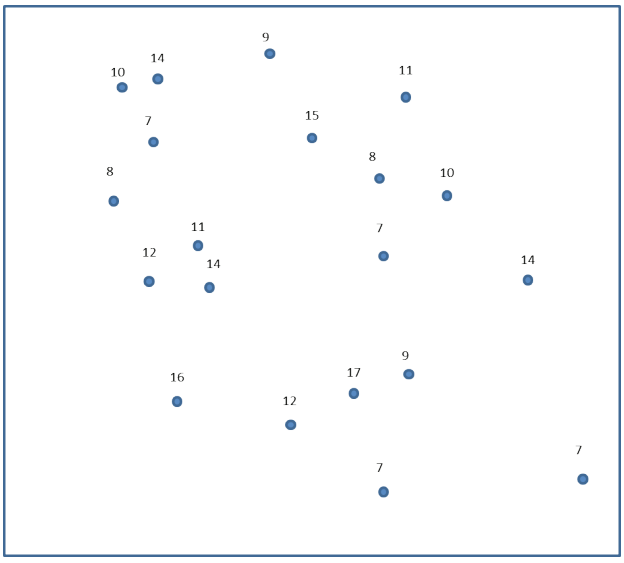
The scale of the data is also an important consideration when linking and analyzing data. Data can be collected at a local scale, such as school or health facility locations in a community, at a district scale, or higher. GIS permits analysis of data collected at multiple scales, meaning that community level data can be analyzed in relation to district level data. However, it is important that the scale of the data be considered during analysis.

Otherwise, data could be misinterpreted due to biases from the Modifiable Area Unit Problem (MAUP) or Ecological Fallacy. Aggregating data collected at a finer scale to larger, coarser scales can introduce bias based on the aggregation approach used. The MAUP refers to this bias.[3] Ecological Fallacy occurs when assumptions about individuals are made based upon aggregated data—in other words, assuming that aggregated data applies to lower levels of aggregation. Both have important implications for sampling, since data from multiple scales can be used to develop sampling frames.

MAUP can become an issue when an area is subdivided into zones and the data at the individual level is aggregated into those zones. To illustrate, consider the following simplified example. Imagine a hypothetical community where every household has one child. The community wants to create school districts in an area. Each district should have the youngest average age possible.

The community conducts a census in the area and collects the age of the household's child. The dots in the figure below correspond to household locations and the child's age.

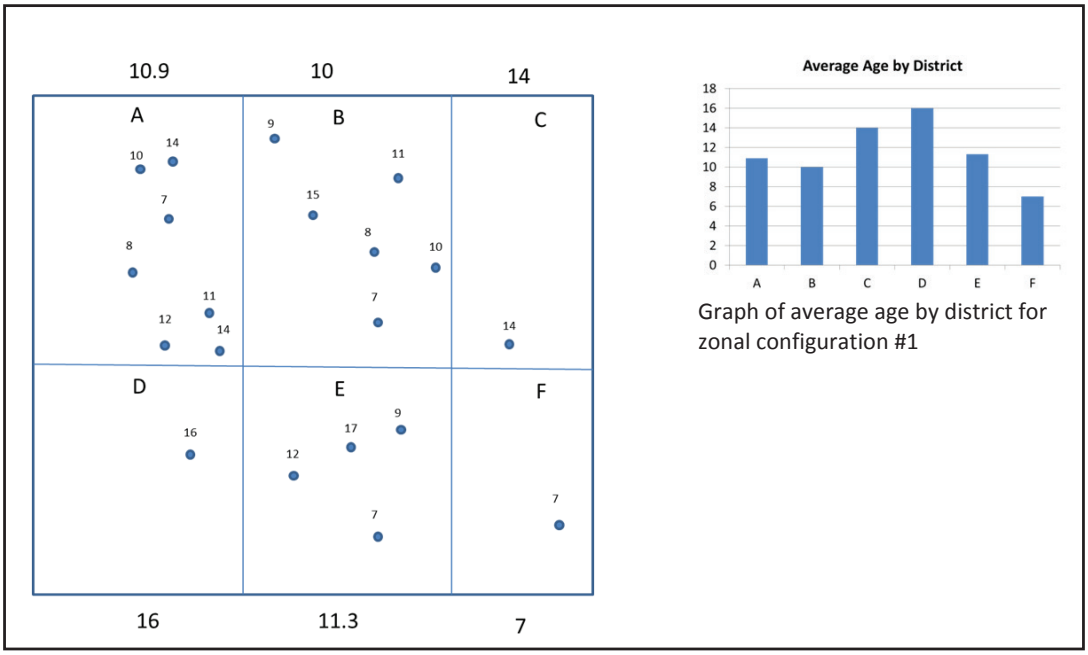
Figure 10. Age of Child in Household



[3] Fotheringham, Wong 1991, *Environment and Planning* Vol. 23, http://www.envplan.com/abstract.cgi?id=a231025

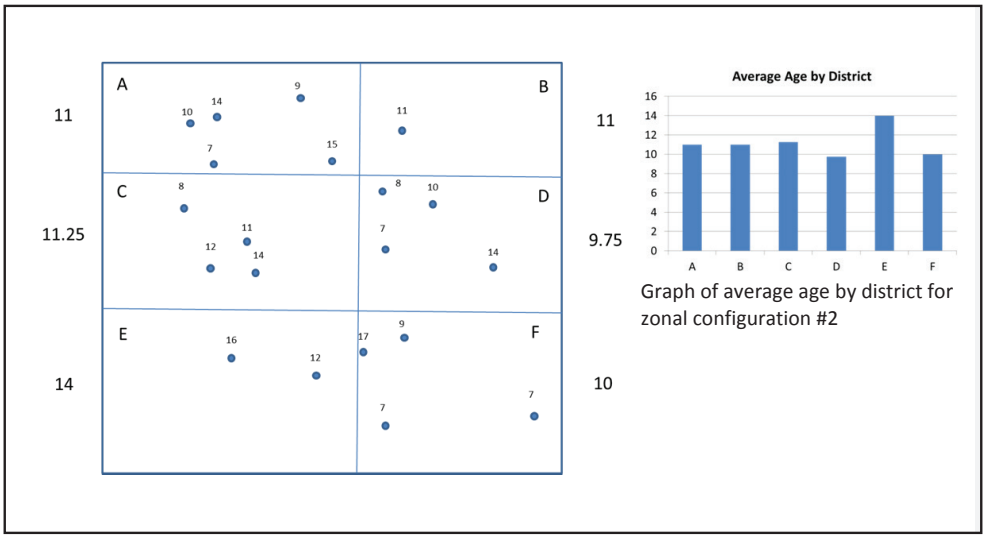
One member of the school board draws the following boundaries over the area and calculates the resulting average age:

Figure 11. Zonal Configuration #1



Another member of the school board creates the following zones and calculates the resulting average age:

Figure 12. Zonal Configuration #2



MAUP is the explanation for why each approach results in different average ages for the zones. There is not a "correct" or "incorrect" way to draw the zones, each is legitimate. Whichever approach is used should minimize unacceptable biases.

GIS offers many different techniques for aggregation and creation of zones. It is up to the user to select the one that best fits the purpose at hand. When subdividing an area for development of sampling frames, it is important to consider any biases that may be introduced based on how the area is aggregated.

There are two strategies recommended for avoiding MAUP (Waller, Gotway, *Applied Spatial Statistics for Public Health* 2004): avoid, if possible, using aggregate data to make inferences on individuals and/or use data collected at the scale at which inferences are needed. Additionally, scale-independent statistics should be used when possible. Tobler[4] recommends avoiding use of Pearson correlation coefficient and employing a cross-variogram technique. Gottway and Young[5] highlight many of the issues to be addressed when dealing with MAUP issues.

In addition to issues around scale, temporal issues are important when using GIS. Data collected at different time periods can introduce bias or error. For instance, if there is significant seasonal migration, population-based data could be skewed depending on when it was collected and the importance of capturing or adjusting for migration.

These considerations (accuracy, scale and temporality) are the foundation of robust geographic datasets suitable for the development of a sampling frame. With a solid data foundation in place, the analysis capabilities of GIS can be utilized. GIS permits analysis of data within and across layers based on attributes or geographic characteristics. Basic analysis could include identifying features within a certain proximity (measured in distance or time) of other features, and identifying features that intersect other features or have specific attributes. There are more advanced analysis features available that can support sampling.

Use of remotely sensed data is an example of the advanced capabilities available when using GIS. Before describing the analysis techniques available, it is important to clarify what is considered remotely sensed data. Remotely sensed data is any data collected using a platform that is not confined to the ground. The sensors are typically on satellites orbiting the earth or air craft flying above the area of interest. The platforms can have active or passive sensors.

Passive sensors collect data corresponding to a spectral signature. Examples include visible light, infra-red or heat. Active sensors emit energy and collect the reflected radiation. Examples include radar, whose reflected signals can be used to identify topography. The remotely sensed product with perhaps the most current utility for sampling is a high resolution aerial image, though the opportunity to use other products exists.

Many people have had exposure to high resolution aerial images through the use of Google Earth. Google Earth uses a mix of satellite images and images from airplanes to provide a birds-eye view of the earth. The image below is a high resolution aerial image of Rome from Google Earth. It provides striking detail about an area, where buildings, streets and even cars are identifiable.

Figure 13. An Image from Google Earth

Source: Google Earth

[4] *Frame Independent Spatial Analysis, "Accuracy of Spatial Databases"* eds M. Goodchild and S. Gopal, London: Taylor and Francis pp. 115-122.

[5] Gottway, Young, Combining Incompatible Spatial Data, *Journal of the American Statistical Association*, June 2002, Vol 97 #458.

Google Earth's imagery, while often helpful at identifying populated areas through visual inspection, does not have adequate spatial accuracy for use in a GIS. The primary purpose of the Google Earth software is to produce a visually consistent and appealing product, and it sometimes distorts images to accomplish this.

In order to use high resolution aerial images and maintain spatial accuracy it is necessary to use a GIS or a specialized image processing software package and obtain images which have been geo-referenced. In other words, the images have been referenced with a high degree of accuracy to the earth and a coordinate for each pixel of the image can be identified. The vendor providing the image will include information in the metadata necessary for the user to geo-reference the image. These images are like the images in Google Earth except they have maintained spatial integrity.

There are several commercial satellites in orbit providing high resolution images. A partial list includes: Quickbird, IKONOS, SPOT, WorldView, and GeoEye. Each platform offers different resolutions and capabilities. The satellites can be tasked by potential customers to capture images of a specific area at a specific time or images can be purchased from the company's inventory. Imagery from the inventory is cheaper than instances where the satellite has been tasked to collect specific images.

If not purposefully tasking a satellite, obtaining high resolution imagery from a satellite such as Quickbird is fairly easy. First, it is necessary to identify a vendor who sells the images and is authorized to conduct business in your area. Most vendors have a website where potential customers can identify the area of interest, set other parameters such as date range or acceptable cloud cover and then see whether images are available. It is then often possible to download thumbnails to ensure that the area of interest is visible. Once images have been identified as being suitable they can be purchased from the vendor.

There are two elements to consider when contemplating the acquisition of high resolution remotely sensed images:

1. Spatial Resolution: The size of the area represented by one pixel in a remotely sensed image. For instance, some data from older, earlier generation satellites may be collected at a thirty meter resolution, which means that any data variation that may happen within a 30 meter by 30 meter patch on the ground won't be captured. High resolution satellite images from a source such as Quickbird can be as high as 25 centimeter resolution.[6] A spatial resolution should be selected that is appropriate for the analysis being conducted. For sampling purposes, 30 meter resolution may be too coarse to identify areas of population and it may be preferable to obtain higher resolution.
2. Cloud cover: Clouds can obscure features on the ground for the visible-light sensors used to create high-resolution images. In some parts of the world, especially in the equatorial region, it can be difficult to find a completely cloud free image. It may therefore be necessary to purchase imagery from multiple time-periods and mosaic them together to obtain a complete cloud-free image for an area.

High resolution aerial images can be very helpful with one of the important tasks in sampling: differentiating between populated and unpopulated areas. This can be especially important in areas prone to rapid changes and where census data no longer reflects the reality on the ground. For instance, slum areas may change dramatically based on population changes or areas being cleared for development. The illustration on the following page shows the changes in a slum area in Dhaka, Bangladesh over the course of seven years.

[6] In June of 2014, restrictions on the resolution of satellite images that are available commercially was modified to permit release of images with 25cm resolution.

Figure 14. Slum Area in Dhaka, Bangladesh

Compared with an image taken 2/26/2010, you can see the large number of settlements that have appeared along the river.

Approximately three years later the settlements have been cleared as indicated in this image taken 2/9/2013.

Source: Google Earth

Census data, which is typically collected decennially or on a less frequent basis, cannot keep up with such rapid change and the use of remotely sensed data such as imagery may often be the only way to adequately assess the population patterns in an area. High resolution images can help to identify clearly areas of population and basic characteristics such as relative density of the population. These areas can be delineated through manually or automated analysis of images. However, there are limitations to the approach that should be considered.

First, it can a daunting or challenging task delineating areas through the use of satellite images. There are several ways to analyze such satellite images. The most basic type of analysis would be a simple visual inspection of the image. An analyst could delineate areas with evidence of population from those areas without population. High resolution images from a vendor such as Quickbird provide a bird's-eye view of an area which can be used to identify individual structures or even vehicles. Polygons can then be created in the GIS that correspond to the populated locations.

For example, in the figure on the following page, an image taken from Google Earth shows a portion of the Makoko section of Lagos, Nigeria. Makoko is a community on the water. Because it is water based, some maps of the area may not adequately reflect where the population is. Thus, a polygon representing a "block" in Makoko has been drawn.

Figure 15. Makoko Section of Lagos, Nigeria

Source: Google Earth

It may not always be practical or efficient to manually delineate areas of population. It is possible to automate the process by using the GIS to analyze the image and identify areas of interest through a technique known as signature analysis. Signature analysis relies on the reflected energy of objects on the ground and the fact that different objects reflect different amounts of the energy emitted towards them. This energy could be in the form of reflected sunlight or in the form of energy emitted from the platform such as radar signals. For instance, metal roofs of structures reflect sunlight differently than vacant lots.

It is possible to have the GIS automatically process images to identify areas with spectral signatures of interest. These spectral signature can then be analyzed by the program to identify areas corresponding to roofs or bare ground, thereby identifying potentially populated buildings.

However, the second limitation to delineating areas by using satellite images pertains to signature analysis. In some cases it may not be able to capture fully the reality on the ground from spectral signatures as many structures are more complex than their appearance from above, even in high resolution images, would suggest.

For example, the figure that follows shows an area in Dhaka, Bangladesh. While some structures are visible there is an area (circled in red) where it is less clear. Therefore, while automated delineation may be efficient, images should, wherever possible, be manually verified to ensure accuracy once structures are identified.

Figure 16. Satellite Image of Dhaka, Bangladesh

Source: Google Earth

Image taken 2/03/2009: Location (23.786311N, 90.414933E)

To fully understand the potential population of the area the imagery needs to be combined with on the ground verification. On the ground, the reality is:

People in tents

Roofs covered with branches

As evidenced by these pictures, the area is populated by people, however the imagery may not adequately identify their presence. Augmenting high resolution imagery with on the ground verification, also known as "ground truthing," is the only way to be certain of populations such as these or to validate the assumption about the area used to estimate with confidence the population.

Aside from the satellite images discussed previously, GIS makes it possible to link together multiple datasets in pursuit of an accurate understanding of the population of interest, which in turn facilitates development of an effective sampling frame. Below are potentially useful GIS datasets that can help provide additional context about an area.

- Global night-time lights dataset from the U.S. National Oceanic and Atmospheric Association (NOAA) (http://ngdc.noaa.gov/eog/dmsp/downloadV4composites.html). This global dataset is a composite of images showing lights visible from space at night. The data is 1km resolution, so it is not suitable for sub-city or small area analysis but for larger areas it can be used to quantify populated areas, it has even been used as a proxy for socio-economic data (Bleakley and Lin (2012),[7] Henderson et al. (2012),[8] Michalopoulos and Papaioannou (2013),[9] Lowe (2014),[10] Storeygard (2012),[11] Pinkovskiy (2013)).[12]
- WorldPop (http://www.worldpop.org.uk/) is a modeled global population database. The data is at 100m spatial resolution and is created through the aggregation of multiple datasets to produce a population surface.
- Census or other official datasets are also useful dependent on their availability. When available, they offer the opportunity to include not only population characteristics but potentially other information such as housing value, tenure, etc. that may be important when creating a population database of an area. However, in some locations the data may be out of date or it may be difficult to find or acquire.
- Weather data can be useful, especially when combined with other data sources such as a digital surface model to identify areas prone to flooding or other events that can affect population. However, weather data at a local scale can be difficult to obtain in many parts of the world. The World Data Center for Meteorology, part of NOAA (http://www.ncdc.noaa.gov/oa/wmo/wdcamet.html), maintains some data online (http://www.ncdc.noaa.gov/cdo-web/) though its coverage and scale varies.

There are potentially other datasets that would be of value beyond these. Regardless of the specific dataset, utilizing a mix of well-chosen spatial databases in a GIS can provide a more robust foundation from which to conduct analysis.

In the effort to accurately locate populations, emerging technologies are offering the opportunity to go beyond static data and take advantage of new techniques such as dynamic data capture. Many of these technologies have proven themselves in other applications such as humanitarian relief, commercial or military contexts, but have not found full utilization in sampling activities.

Geo-referenced video through GPS enabled cameras offers the opportunity to canvass an area in a vehicle and then classify and identify population characteristics at a later date through a GIS. Such an approach can cut down on time and number of staff needed on the ground. In Haiti, researchers from Kent State University in the U.S. used off-the-shelf $250 video cameras to map health risks (A ubiquitous method for street scale spatial data collection and analysis in challenging urban environments: mapping health risks using spatial video in Haiti; Andrew Curtis, Jason K Blackburn, Jocelyn M Widmer and J Glenn Morris Jr; *International Journal of Health Geographics* 2013, 12:21).

---

[7] Portage and path dependence, Bleakley & Lin, Quarterly Journal of Economics, Vol 127 #2, pp 587-644 (http://qje.oxfordjournals.org/content/127/2/587.full).

[8] Measuring Economic Growth from Outer Space, Henderson, Storeygard, Weil, American Economic Review 2012, 102(2):994-1028 (http://www.econ.brown.edu/faculty/David_Weil/Henderson%20Storeygard%20Weil%20AER%20April%202012.pdf).

[9] Divide and Rule or the Rule of the Divided? Evidence from Africa, Michalopoulos, Papaioannou, NBER Working Paper #17184 (http://www.nber.org/papers/w17184).

[10] Night lights and ArcGIS: A brief guide, Matt Lowe, January 2014, White paper, (http://economics.mit.edu/files/8945).

[11] Farther on down the road, Transport Costs, Trade and Urban Growth in Sub-Saharan Africa, Storeygard, World Bank Development Research Group, May 2013 (https://openknowledge.worldbank.org/bitstream/handle/10986/15586/wps6444.pdf?sequence=1).

[12] Economic Discontinuities at Borders: Evidence from Satellite Data on Lights at Night, Pinkovskiy, Whitepaper (http://economics.mit.edu/files/7271).

In some parts of the world, Google Street View can provide a virtual ground-level view of an area. While Street View images aren't able to be imported into a GIS, they can be used in conjunction with high resolution aerial images to differentiate between commercial and residential structures or validate building height.

Figure 17. Google Street View Image

Source: Google Street View

To summarize, the growing spatial data infrastructure, as well as emerging technologies, offers many opportunities for GIS and geo-spatial methods to support sampling. However, regardless of the technology or method, a certain amount of rigorousness of methods is necessary. Because GIS makes it easy to combine data together using geography, it is possible to combine data that have conflicting resolutions or biases. Just because two datasets can be combined together does not mean they support robust analysis. GIS is a tool for science, and with any scientific tool the inputs and their limitations should be well understood and the outputs should be in line with those limitations and be reproducible.

As stated previously, the use of data at multiple scales increases the risk of compounding MAUP and ecological fallacy associated errors. It can also introduce a false precision that the data does not support. For instance, if a facility location was identified based on the name of the community in which it was located, the GIS operator may locate it at the center of the community. Using that location to calculate distance from that facility to a household whose location has been obtained via GPS is not appropriate. Each does have a location in the GIS and it is thus possible to have the software calculate the distance, but because there is a false precision resulting from the facility location it is not appropriate to do so.

Sampling benefits from tools that enhance the ability to derive efficiently and effectively a sampling frame, and GIS can do so either through its analytical capabilities or through creation of products to help direct or manage the field effort. The contribution of GIS is likely to only increase as new technologies and richer datasets have the potential to reduce the time needed in the field and to reduce costs associated with conducting sampling dependent activities. However, effective use of GIS is a balance between the data, techniques and questions being addressed. The geography that underlies data potentially opens a door to a rich mosaic that can increase understanding of an area, but careful considerations of the data and methods are necessary to ensure that the results are robust enough to support sampling-based activities.

Chapter 4. GIS and Sampling

In our exploration of the principles of sampling, we learned the importance of probability sampling for selecting samples that could support unbiased estimation of population parameters of interest. Probability sampling is most readily accomplished via selection of a sample from a list. For the key requirements of probability sampling (namely, that every element of the population from which one is selecting the sample have a known, positive probability of selection) that list must contain every member of the population of sampling units from which the sample is to be selected.

For the purpose of selecting samples of households or individuals, for whom exhaustive lists at the population level typically do not exist, multi-level sampling is standard practice. In that case, successively smaller sampling units typically based on administrative units or some other spatial units are selected until we have obtained a sample of such units with sufficiently small populations that it is feasible somehow to list units of observation (e.g. households, individuals that will actually be the units for estimation) within them. The analogy to probability sampling in this case is that the frame for the first stage units, the primary sampling units, must contain the entire population of primary sampling units. For instance, if the ultimate goal is to select a sample representative of all American households and the primary sampling unit is the state, the first stage frame should contain all 50 U.S. states. The frames for the selected sampling units at each successive stage of selection must then contain all of the sampling units for that stage of sampling for the population within the selected sampling unit from the preceding stage of selection. Continuing the example, if the county is the second stage (i.e. secondary) sampling unit then the frame for selecting counties for a state selected in the first stage should contain all of the counties in that selected state.

The emphasis on probability sampling, along with a traditional general focus on obtaining statistics representative at national or regional levels, has led to an emphasis on first stage frames that are as expansive as possible to insure that no members of the population are left out of the sample selection process. For this reason, there has been a tendency to rely on national sampling frames comprised of census listing units (i.e. "enumeration areas") or larger administrative units for the first stage of selection. The classic example of this is the typical Demographic and Health Survey (DHS), for which this is an entirely appropriate starting point for sampling given the goal of obtaining estimates of indicators such as fertility or modern contraceptive use representative at national or regional levels. This is altogether appropriate in such cases, and will remain so for the foreseeable future, as so much programming is essentially oriented toward the national level (e.g., the multi-year national health, nutrition, and population sector programs in Bangladesh).

We are, however, now witnessing the gradual emergence of a tendency toward programming to be more narrowly targeted toward specific, often vulnerable, sub-populations of interest. This includes programs oriented toward refugees, the HIV positive, those environmentally vulnerable to particular infectious diseases, those maintaining their existence under particular agricultural circumstances, those living in concentrated pockets of poverty and environmental vulnerability in urban areas (i.e. slums), those working in certain industries (e.g. garment workers), etc. Beginning with the broad brush approach to the first stage of sample selection typical of DHS surveys, we would likely end up with a sample of enumeration areas that often contained few or no members of these subpopulations of interest.

One reason for this is that we would generally not expect such vulnerable populations to be distributed evenly across society in a spatial sense (another reason might, of course, simply be that these subpopulations may be small in terms of their share of national population in the society under study). Instead, we might expect that some geographic areas of a given society represent particularly concentrated pockets of individuals belonging to such subpopulations. An obvious case of this is slum populations. Slums in many societies (e.g., in South Asia) tend to contain a large

proportion of the population of cities. However, they are densely settled and hence, despite the large populations within them, tend to be sited on land that represents a tiny fraction of the entire land area of those cities. Reliance on a general urban, or even specific city-wide, first stage sampling frame would thus likely yield a first stage sample with many selected units of no relevance to the estimation goals of a survey of slums because in many cases the selected first stage sampling units (e.g., census enumeration areas) might contain no slums. Equally importantly, it might yield very small samples of slum dwellers (the practical reasons for this will be explored in greater depth in the next chapter).

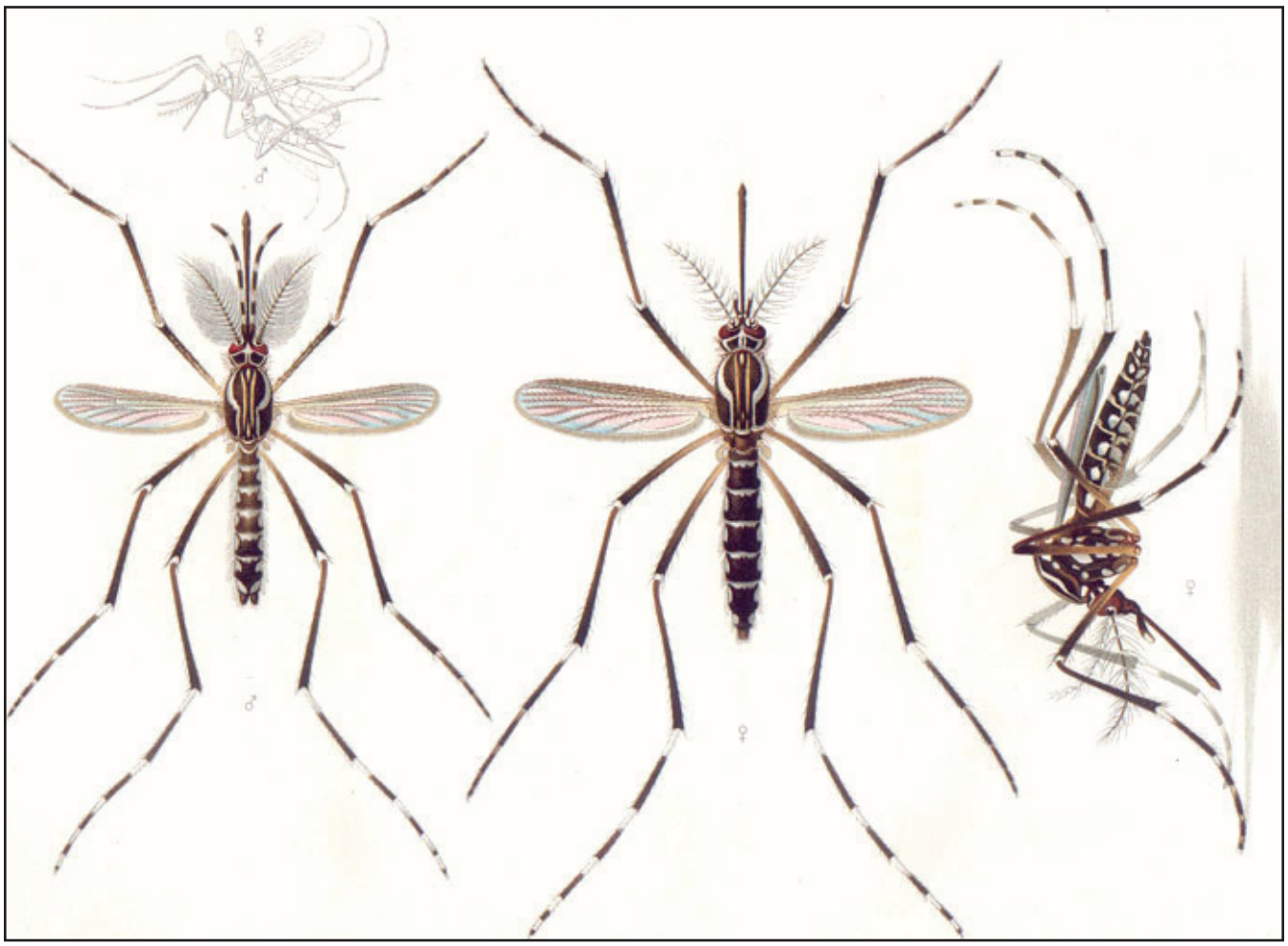
To the extent that such vulnerable populations are spatially concentrated, we would thus expect that multi-stage sampling that began with national or general population frames might be quite inefficient. This will translate into a survey that is likely unnecessarily costly and slow (as field work becomes the setting for weeding out the units not useful for studying the vulnerable population in question) and, to the extent that overall survey resources are constrained, will yield smaller samples from the vulnerable populations than might otherwise have been the case since scarce financial resources need to be dedicated to the weeding process. Thus, when the subpopulation of interest is spatially concentrated, the basic problem with the "broad brush" sampling approach of surveys such as the DHS, with their orientation toward indicator estimation at the broad national or regional overall population levels, is that many selected sampling units will prove useless or nearly so for possessing few or no members of the subpopulation of interest, while the final sample of members of the subpopulation of interest might be quite small.

If there were some means of developing sampling frames for the primary and other, higher level (e.g., above the household) sampling units that somehow focused on, or more effectively isolated, higher level sampling units containing concentrations of the members of these vulnerable subpopulations, then it might be possible to sample more specifically from these subpopulations than is the case under the more "broad brush" approach. This could cut survey time and cost considerably per unit of observation (e.g., the household or individual) actually successfully sampled from the subpopulation, in the process delivering vital programmatic information in a far more efficient and effective fashion. Put differently, for a given expenditure of time and money on a survey, far larger samples from the subpopulation could be obtained than under the "broad brush" approach that ignores the possibility of such spatial concentration to the subpopulations of interest.

It should be evident that this is true in the extreme case where a subpopulation of interest lives in only a few areas of a country—by isolating just the sampling units from those areas, we can avoid selected sampling units from other areas that will contain no members of the subpopulation. However, considerable efficiency gains could be had even if this process of developing more focused sampling frames simply parsed a nation into areas where the members of a subpopulation were concentrated and those where they were sparse (as opposed to identifying the subset of areas where all members of the subpopulation live, which might not be possible since there can be pronounced concentrations of that population even as at least some small numbers from it live in all areas of the country).

Consider a fictional but highly realistic example involving a survey in a particular city to study the social epidemiology of dengue fever with the goal of understanding the social factors determining the severity of outbreaks. Dengue is spread primarily by the *Aedes (A.) aegypti* mosquito[1] (other *Aedes* mosquitos that can carry it include *A. albopictus*, *A. polynesiensis* and *A. scutellaris*). Humans are the primary hosts for dengue in urban settings.[2] The transmission cycle involves an *A. aegypti* mosquito biting an infected human, in the process becoming a carrier that can then spread dengue to any other person that that mosquito bites.

Figure 18. The Aedes Aegypti Mosquito as an Adult

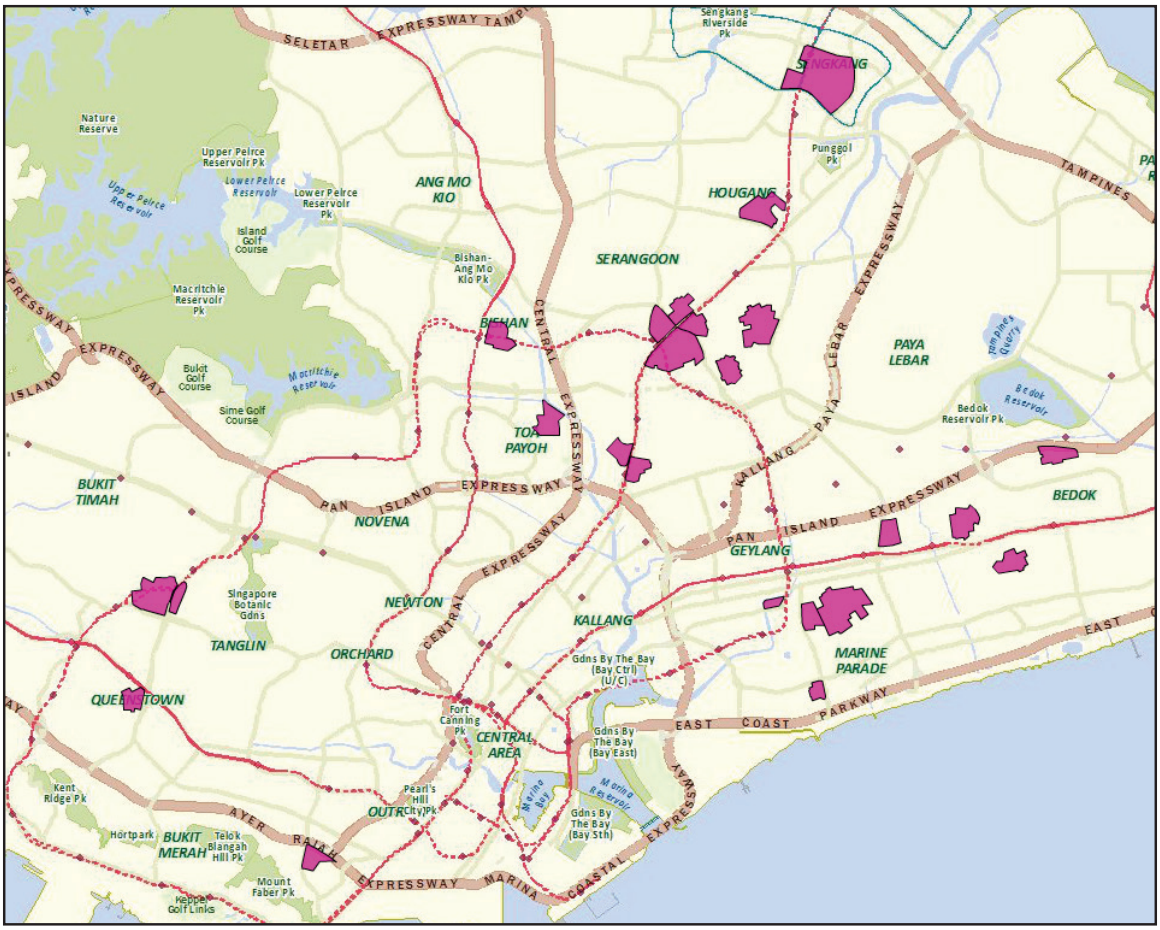


Source: Wikimedia Commons, CC-PD-Mark

[1] *A. aegypti*, the gift that just keeps on giving, also plays a key role in Yellow Fever transmission.
[2] Other primates also play a significant host role in less urban settings.

There are three particularly interesting features of the *A. aegypti* mosquito for the purposes of studying the social epidemiology of dengue. First, its range is rather limited. An *A. aegypti* mosquito might in its lifetime (consisting of a couple weeks to maybe a month) fly *at most* only a few hundred meters from its birth place. Likely because of this, dengue outbreaks tend to occur in pronounced clusters, even in dense urban environments (the primary transmitter *between* clusters is then probably infected humans). In the figure below, we see dengue clusters evident in Singapore as of April 25, 2014. These clusters were often large in terms of cases numbers, with tens to up to 160+ cumulative cases in each cluster as of that date.

Figure 19. Active Dengue Clusters in Singapore as of April 25, 2014



Source: National Environment Agency, http://www.dengue.gov.sg/subject.asp?id=74

Second, the mosquito tends to favor stagnant but clean (i.e. not contaminated) water for laying eggs. A number of settings have been associated with their egg laying, including:

- Middle and upper class areas where receptacles such as flower pots that catch rain water are common;
- The same in indoor settings where water might pool, such as showers or wash basins;
- Lower income areas with limited piped water systems, where water is often stored in containers that can be imperfectly sealed off from *A. aegypti*;
- Construction sites often become a locus for pooled water and hence *A. aegypti*;
- Areas where items are either stored or discarded that can readily catch and hold pools of rainwater (e.g., tire storage and disposal sites often have large numbers of *A. aegypti*).

By contrast, well drained areas served by municipal water pipes but with few rainwater retention opportunities tend to have less substantial *A. aegypti* populations.

Figure 20. An Aedes Larvae Resting Comfortably (in Clean Water)

Source: Amir Ridhwan/Shutterstock

Third, pronounced dengue outbreaks (as opposed to some typical low level background prevalence in non-epidemic times) usually require some critical mass of *A. aegypti* to be in circulation. Essentially, considering that at any given time (even during epidemic periods) probably only a small percentage of *A. aegypti* are carrying dengue, a larger overall number of these mosquitoes are required to support transmission rates sufficient to generate an outbreak. This, coupled with *A. aegypti's* need for clean water to lay their eggs, helps to explain why dengue outbreaks tend to occur during periods of heavy rain (such as monsoons)—those heavy rains generate expanded egg laying habitats, which cause the *A. Aegypti* population to surge until it reaches a critical mass large enough to sustain a human outbreak given human settlement density. It should seem clear then that those areas offering the most abundant catchment for clean rain water might be most at risk for seasonal dengue outbreaks.

There is thus clearly a *spatial pattern* to dengue outbreaks, even within cities—they occur in circumstances under which the mosquito is most likely to thrive, but then only in clusters. The investigators for this fictional survey plan to select a sample of enumeration areas (EAs) from the census frame for the city and then monitor the selected EAs (perhaps through local health care providers) during the rainy season to determine whether they are experiencing a dengue outbreak (defined as some threshold prevalence of likely cases). If and when an outbreak is identified in one of the selected EAs, extensive field interviews will be conducted to collect a community-level instrument, as well as a household level instrument collected from a sample of the households in that dengue-afflicted EA. The population of interest is the households in EAs with concentrations of dengue cases.

Straightforward selection of a sample of EAs from the census frame for the entire city is not likely to be particularly efficient since dengue outbreaks would likely occur in only a small percentage of them. Suppose for instance that only around 4.9 percent of the EAs in the census frame for a given city would likely experience a dengue outbreak

in the next rainy season. Thus, for every 100 EAs selected from a general population frame we might expect dengue outbreaks in only around five of them. Given the limited resources available for many surveys, this represents a significant obstacle to obtaining reasonable final sample sizes.

To fix ideas, suppose as well that the survey budget for monitoring and fieldwork is $300,000, and that monitoring for outbreaks costs $200 per EA, while field interviews cost $2,000 per EA with an outbreak. With this budget and the expected percentage of EAs experiencing an outbreak, the survey team has the resources to select up to 1,006 EAs: the monitoring costs for these EAs would be 1,006*$200=$201,200, while the fieldwork costs in the 49 (roughly 4.9 percent) where an outbreak is expected to occur would be 49*$2,000=$98,000. In the end, given the cost structure and for $300,000 dollars, straightforward selection from the census frame for the city would be expected to yield a sample of 49 EAs with dengue outbreaks.[3]

Suppose, however, that, behind the overall 4.9 percent risk number, was a circumstance where five percent of the city's EAs had a 60 percent risk of an outbreak (due to their environmental circumstances), while the remaining 95 percent had a two percent risk (thus yielding an overall average risk of 4.9 percent). If the team was able to identify which EAs were in high risk areas prior to sample selection, they could stratify the city frame into frames of high and low risk EAs in the city.

This would afford the team tremendous flexibility in sampling that could yield a tremendous increase in the dengue afflicted EAs in their sample. For instance, one possibility financially feasible under this stratification scheme would be to select 200 EAs in low-risk areas (with an expected yield of four EAs with dengue outbreaks) and 180 in high-risk areas, with an expected yield of 108 or so high-risk EAs with actual outbreaks. In other words, the total yield of EAs with dengue outbreaks would now be about 112 for the same survey budget.

Thus, simply by such a stratification scheme one would, for the same budget, more than double the expected sample of EAs experiencing outbreaks! Further, this would still be a probability sample of populations and EAs exposed to dengue outbreaks since every EA in the city, and hence everyone living in those EAs, still had a known, positive probability of selection. The overall benefit from stratification into high and low risk EAs would be substantial. Even if the parsing into high- and low-risk areas was less precise, there would still likely be a substantial increase in the yield of dengue outbreak sites over what would be obtained with simple selection from the overall frame for the entire city.

The question then becomes how one could parse the city into low- and high-risk areas. It is extremely unlikely that the census frame itself will contain much useful information for doing this. However, there might be other sources of information that could be useful. For instance, it is likely that information is available about the sites of past outbreaks within the city from its public health department or local epidemiologists. Presumably there are maps of the water mains in the city, thus revealing which areas likely are and are not well served by piped water and thus the areas where residents are most likely to collect and store water. Topographical and sewage system maps of the city would likely reveal which areas are least well drained and hence most prone to the formation of the stagnant rainwater pools that can serve as *A. aegypti* breeding platforms. Finally, analysis of satellite images of the city could help to pinpoint all sorts of areas where risk is likely heightened. For instance, piles of tires would likely be evident, as well as middle and upper class areas where gardens, and hence urns, flower pots, other rain collection containers, are common. Indeed, consultation with local epidemiologists would likely lead to all sorts of circumstances that promote *A. aegypti* populations and can, with reasonable accuracy, be assessed from satellite imagery.

These sorts of information sources usually come with some sort of geographical identifying information, and hence can be merged according to that information. For instance, satellite imagery (e.g. Quickbird) is typically georeferenced "out of the box." For the cities of lower income societies where they are available, maps of things such as waterlines are often hardcopy but generally accurately to scale. In such cases they can be scanned, loaded into GIS software, and georeferenced using the coordinates of identifiable reference points (perhaps taken with GPS devices, or simply with careful overlaying on the georeferenced satellite imagery). The same process can be applied to topographical maps,

[3] The total cost would actually be $299,200, leaving $800. However, this would not allow for selection of any more EAs since there would not be enough resources left to conduct fieldwork even if dengue was found.

if they are not already digitized and georeferenced. Finally, information on events like past outbreak clusters may or may not be available in scale map form. However, the administrative areas where such outbreaks occurred and other descriptors can usually be used to add them to a GIS. For instance, the location and borders of the low level administrative areas of a city where these outbreaks occurred can usually be added to a GIS using the scale maps of administrative units that typically are available.

Once entered into a GIS, this data can be synthesized to form measures of spatial dengue risk. These can often be based on algorithms, the execution of which through GIS software can be quite fast. If the census frame can be brought into the GIS (which is sometimes, though not always, possible) higher and lower risk EAs can be readily identified. In cases where the census frame cannot be easily brought into GIS (typically because the maps of the EAs are not digitized or to scale, though digitized scale maps are slowly becoming more common) then primary sampling units can be defined by GIS analysts (presumably working with samplers to define them in a fashion most consistent with efficient, unbiased sampling).

Figure 21. A Larger PSU (Outlined in Green) Generated from Several Enumeration Areas (EA Boundaries in Red)

This need not even necessarily be a laborious process of replicating thousands of EAs. For instance, even working with non-scale EA maps as a reference, it should be possible to create primary sampling units that encompass a few dozen EAs each but within which dengue risk is homogeneous. Then, those new primary sampling units can form the basis of high and low risk strata and, for instance, should one be selected several of the EAs within it could then be selected as a secondary sampling unit. In the figure above, several EAs with similar roof typology and density are combined to form a single new sampling unit.

The point is that there are options for building spatially-oriented risk ratings through a GIS, and this can then be used as a basis for stratified sampling from areas with different risks. GIS then opens the door to the sort of far more efficient sampling that we have shown to be a potential fruit of appropriate stratification by risk.

The basic themes in this example have been the power of being able to stratify by the likely concentration of a subpopulation of interest (in this case those most vulnerable to dengue) in sampling units and the possibilities offered by GIS for making that happen. This example was in some sense all the more powerful because there was a certain risk of dengue everywhere in the city. However, there were great efficiency gains to be had simply by identifying the places at *particular* risk, even if the goal was to develop a profile of the social demography of dengue for the entire population in places where outbreaks occurred. And GIS is the tool that can open the door to such stratification.

In other instances, the subpopulation might reside only in a subset of the areas covered by, for instance, a census frame. The appeal of identifying these areas is more obvious (why select sampling units *guaranteed* not to hold the subpopulation of interest?). However, even in these cases there will likely still be many instances where GIS holds the key to more efficient sampling by providing a framework for identifying many (even if not all) of the sampling units that should be discarded before selection because they do not contain any members of the subpopulation of interest.

This is not in any sense an isolated or singular example—there are many other sorts of circumstances where one might improve the efficiency of the sampling process by exploiting the information available through a well-constructed GIS to identify areas of concentration for a subpopulation of interest. We offer several more examples.

One might wish to survey recent migrants to a city to develop a profile of some component of their human welfare. If migrants tend to settle across the city in a rather even fashion there is probably little scope for efficiency gain by stratification through the information provided by a GIS (since the GIS will simply suggest that the migrants are scattered all over, with no particular concentration). However, there often are uneven patterns to migrant settlement. In high income societies such as the United States, migrants are typically from other urban areas and are often more prone, other things being equal, to settle in newer neighborhoods or residential developments closer to the center of economic activity in the city. In lower income societies the migrants often come from outlying rural areas. They tend to settle in the rapidly expanding informal or slum settlements, often at the periphery of the city, but are probably also more likely to settle in recently developed areas more generally. Another likely settlement area would be around urban enterprises that tend to employ recent migrants. Whatever the case, consultation with local officials, program personnel, and researchers can usually identify the kinds of environments within the city where recent migrants might settle.

This information about migrant settlement patterns can often be utilized for building a GIS that informs sampling. Specifically, data sources that contain spatial information regarding the prevalence of indicators of the kinds of areas where migrants settle can be synthesized into a GIS to identify areas of the city where they are most likely to settle. Suppose, for instance, that much of the settlement of recent migrants occurs in areas where there has been a great deal of construction (often there are specific stretches of the urban periphery that are sites for particularly intense urban structural expansion, while in other cases it is more a situation of "filling in" peri-urban pockets within the city). These sorts of areas can typically be readily identified through analysis of satellite photographs of the city taken over time.

Figure 22. The Peri-Urban Fringe of Kibera Slum, Nairobi, Kenya

Source: Google Earth

Shifting gears, let us move toward a more rural setting and consider the challenge of surveying nomadic populations in order to estimate a health profile for their population in some region. There are various definitions of nomad populations but most involve elements along these lines:

1. Their livelihood depends on livestock;
2. They are frequently on the move seeking water sources and pasture;
3. They have no permanent place of residence.

Thus, the population of interest for this survey is one that is on the move.

Ideally, the sampling scheme would be one under which all members of the nomadic population have some probability of selection. Given their mobility, they present a tricky intrinsic challenge. For instance, suppose that selection of survey clusters and fieldwork are conducted in the tradition of surveys of fixed households, with their reliance of multi-stage sampling of successively finer areas, until one gets down to a spatial designation small enough to conduct household listing. Nomadic families might keep moving during fieldwork, generating the risk that some nomadic families might be selected more than once (as they appear in more than one selected area while field work is under way), while other nomadic families might never appear in an area while fieldwork is being conducted in it.[4] Aside

[4] For an excellent discussion of the challenges of sampling nomads, see Kalsbeek, W. and Cross, A. (1982) "Problems in Sampling Nomadic Populations," *Proceedings of the American Statistical Association, Survey Research Methods Section* p. 398-402.

from these substantive scientific considerations, there is also the challenge that a lack of specific information about where nomads are likely to be (even micro-regions where nomads might be congregating can still be vast, with the nomads in them concentrated in a few small sub-areas at any given time) can lead to survey efforts being spread over a large area with few or no nomads located in most EAs/clusters selected and a great waste of survey resources.

Some other approaches to sampling nomads have been suggested. One approach is to use the tribal structure as the sampling framework (with, perhaps, the tribe serving as the primary sampling unit and successively lower level social units within selected tribes serving as secondary, tertiary, etc., sampling units until one gets down to a unit at which well-defined households can be selected). This approach is often quite time consuming to execute in practice. Another approach is to seek the assistance of government officials and local leaders in persuading the nomads to assemble at some fixed point for interview. Noncompliance is a challenge under this approach and it also raises ethical issues (there are potentially risks for participants under this strategy).

Another method exploits the fact that nomads do not just wander aimlessly—there is usually a pattern to their migrations dictated by season and ecological circumstances (in particular, rain fall patterns). In many societies, at certain times of the year the pastoral lifestyle can only be sustained in specific places per rainfall and snowmelt patterns (nomadism tends to be particularly common in places where overall annual rainfall levels are fairly low, making exploitation of seasonally shifting rainfall patterns a key to successful pursuit of the nomadic lifestyle). Even then, however, nomads are likely to make relatively small local adjustments to their location given small intraregional variations in rainfall patterns.

A GIS can be built around factors that predict nomadic presence. A key component of this GIS is likely to be weather radar information, which is typically digitized and georeferenced and from which fairly accurate and timely maps of running seasonal totals for rainfall can usually be constructed. Topographical maps can improve on predictions of where nomads are likely to be. Indeed, if one wishes to get really sophisticated about this, recent historical information on factors such as rainfall and satellite imagery that perhaps reveals nomadic migration tracks can be used to predict where nomads are coming from, making predictions about where they will move to in response to local variation in rainfall patterns even more specific.

Using this GIS, high, low and zero risk areas for nomadic activity in the can be not identified. Indeed, areas of nomadic activity can likely be *predicted*—remember, nomads are no better at predicting the weather than we are. Hence, their migratory patterns within local areas likely follow the rains with some lag. At any rate, this stratification by risk can make fieldwork far more efficient, shortening the geographic scope and length time-wise required to obtain a given sample of nomads. This will imply far lower survey costs for a given final sample size, but also conveys substantive scientific benefits. Faster fieldwork lowers the probability of "double counting" given nomad families and more focus and speed to fieldwork will likely mitigate frame undercoverage (i.e., the possibility for nomad families that effectively have no chance of selection). In short, it holds the promise of making a traditional enumeration area/cluster based approach to sampling far more likely to succeed by quickening and focusing fieldwork.

GIS can also help to identify "activity spaces." These are places where human activities of various kinds occur. Identifying where activity occurs can once again help to narrow the geographic scope for observing a given phenomenon of human endeavor, potentially increasingly tremendously the efficiency of sampling to study it. We briefly consider an example.

First, consider the task of studying the behavior of catchment areas for health care facilities. Typically, the catchment areas of facilities are for analytical purposes defined as some sort of radius around the facility. This simple principle can be hard enough to execute on the ground (for instance, determining precisely which residential structures lie within the radius can be a tricky and error prone business). More subtly, the simple radius rule approach has its limitations. Suppose, for instance, that a river lies within the specified radius around a given facility. It seems unlikely that those on the far side of that river from the facility would represent meaningful residents of its catchment area. Major roads and highways, large tracts of state property (e.g., government administrative buildings, military facilities, etc.), and other factors can also serve as impediments to utilization that should preclude inclusion of households otherwise within the catchment radius of the facility.

A GIS of roadmaps, topographical and river maps, satellite images ,and health facilities can be formed to capture this complexity to the likely catchment space of a facility. Once all of these items are gathered in a well-designed, digitized GIS, more nuanced definitions of facility catchment areas can be operationalized, often through carefully crafted algorithms.

Figure 23. A River Runs Through It: Technically These Communities Are Only 225 Meters Apart

Source: Google Earth

A catchment area for a facility is an activity space in the sense that it defines the spatial area within which an activity (trips to visit a given facility) occur. GIS can help to define all kinds of activity spaces for the purpose of studying the activities, and people who perform those activities, within them. An absurdly partial list includes:

- Locating particular types of industries or commercial enterprises. For instance, the ubiquitous cyclo drivers of much of urban South and Southeast Asia roam their respective cities, but tend to operate out of fixed bases, which can often be readily seen from satellite imagery (they tend to be points where various numbers of cyclos are parked, there are evident cyclo repair and manufacturing sites, etc.). In rural and peri-urban areas many types of light industry are readily identifiable from satellite photos due to distinctive tell-tale visual signatures. For example, in the figure that follows we have a satellite image showing brick factories in South Asia, the tell-tale signs being the distinctive ground coloration around them and the tall smokestacks for their kilns (proximity to a river/canal/stream is another sign since bricks are often most economically transported via water in lower income countries);

- Identification of patterns of activity at night, including markets and enterprises active at night. Thus far, virtually every satellite image shown has been during the daytime, but they are also available at night. Patterns of activity can often be discerned by comparing patterns of light at night with known features of the spatial environment in question;
- GIS can be used to identify likely commuter routes, allowing for a better understanding of the origins and destinations of those making routine trips in the pursuit of various human activities.

These are just a few examples of how a well-crafted GIS can identify human activity spaces, allowing for targeted sample selection for studies concerned with the nature of those spaces and those who carry out activity within them.

Figure 24. Three Brick Factories in Bangladesh

Source: Google Earth

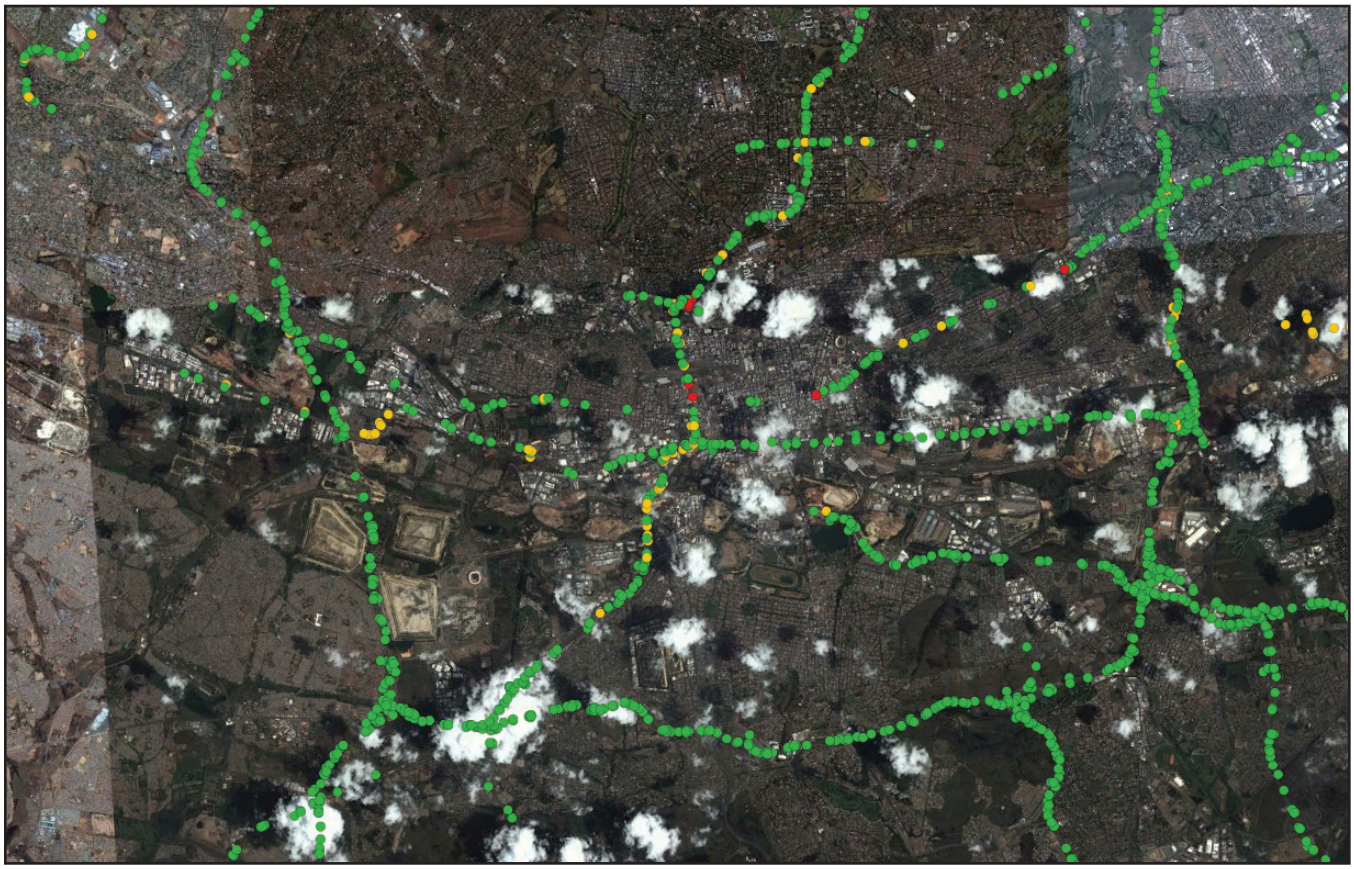
Before moving on from these examples, we also note the somewhat recurring quality to the GIS information sources we have described: satellite images, georeferenced (or georeferencable) maps of various dimensions of infrastructure (roads, water lines, etc.), official statistics, etc. These indeed have been common traditional building blocks of GISs in wealthy and lower income countries alike. However, we live in a time of tremendously expanding technological possibilities, and so the information sources that could inform a GIS are continuously expanding, often in surprising and innovative ways. Many of these technologies are either currently in use primarily in a few wealthy nations, or most ubiquitously used there, but will likely become increasingly used in lower income nations, increasing tremendously the possible information sources for a GIS.

This is a topic to which we will turn in more detail later in this manuscript, but for now we give the reader one example that provides a sense of the range of development. As mobile phone technology becomes increasingly prevalent, all sorts of possibilities for capturing human activity remotely are emerging simply by recording statistics regarding interactions between phones and specific cell sites (also known as "cellular towers"). Consider traffic flows, which traditionally could once be recorded only with expensive (to place and to maintain) monitoring equipment. As cell sites and mobile phone ownership rates have soared, it has become increasingly feasible to capture traffic simply by monitoring contact between phones and towers. A famous and easily accessible example of this is Google's nifty traffic monitoring feature. It has reached the point where Google's coverage now extends to extremely small, largely rural communities in places like the United States. In the following figure, we show a screen capture of a Google Earth image of Ramseur, North Carolina (with a population of approximately 1,700) with the traffic monitoring feature enabled. The dots indicate discrete phones in contact with cell sites (the colors indicate the speed with which each phone is moving). However, the coverage of this tool is rapidly expanding and is now available in, for instance, parts of South and Southeast Asia and Sub-Saharan Africa. For instance, the figure on the following page shows a screen capture of the Google Earth traffic tool for Johannesburg, South Africa.

Figure 25. Google Traffic Map of Ramseur, NC, USA (March 27, 2014 Around 1pm EST)

Source: Google Earth

Figure 26. Google Traffic Map of Johannesburg, South Africa (March 27, 2014, Around Noon EST)



Source: Google Earth

What we have learned thus far is that, by improving our understanding of where subpopulations of interest are concentrated, a GIS can potentially increase tremendously the efficiency of sampling in terms of the final sample size from that subpopulation of interest yielded by the sample selection process. We have seen this through one (in some sense canonical) example involving a survey of conditions in local areas experiencing dengue fever outbreaks. However, subsequent examples provided just the briefest taste of the possibilities for using GIS to inform sampling.

It is impossible to predict the ways in which a GIS might be applied to sampling because the information sources that can inform a GIS are rich and growing with incredible speed, and sampling is done in all sorts of settings, to obtain samples from an infinite number of potential subpopulations, and to support surveys to study myriad topics. It is therefore not possible to offer a "cookbook approach" to the application of GIS to sampling—it would, at best, be useful for perhaps a small portion of the sampling challenges for which GIS might serve as a useful tool.

The course we take is to discuss general principles of best practices for applying a GIS to a sampling challenge and, in the next chapter, offer a detailed case study of an actual instance where GIS was applied to a sampling challenge, to illustrate the application of those concepts of best practice. These principles form a framework that should have clear implications for specific guidance in particular applications where one wishes to use a GIS to inform a particular sampling challenge.

Knowing When GIS Can Inform Sampling

Perhaps the most important principle of practice is to determine when GIS can in the first place usefully inform sampling. A well-constructed GIS can improve the efficiency of the sampling process in many, many applications, yielding accurate samples of subpopulations faster, better, and more cheaply than would be the case without the additional information that it provides. However, not every sampling process can be informed by a GIS. It is important, therefore, to differentiate circumstances where a GIS might or might not prove an asset to sampling.

First and foremost, a GIS is a tool for capturing and analyzing spatial patterns to various phenomena. Therefore, a GIS is most useful in circumstances where there is some sort of predictable spatial pattern to the concentrations of the subpopulation of interest. In the absence of that it is hard to see how a GIS can improve the sampling process. For instance, in the motivating example of this chapter (identifying localized pockets of dengue outbreaks) all of the gains to GIS came from the notion that some areas of the city were at higher risk than others for dengue outbreaks. Had that not been the case (e.g., had the risk of dengue outbreak been uniform throughout the physical area of the city), the entire mechanism by which a GIS could have informed sampling would have broken down. There would have been no way that a GIS could have divided the city into distinct areas according to dengue risk in order to allow stratification of primary sampling units (census EAs for the city) according to that risk.

In practice, the question is not simply whether there is a spatial pattern to some phenomenon or population, but whether a GIS can capture the pattern sufficiently well for predictive purposes in sampling. There are lots of examples of phenomenon or subpopulations where it is clear that in each time period they exhibit some kind of apparent spatial pattern. This does not automatically mean, however, that we would be able to use information associated with these historical spatial patterns to predict what the pattern would likely be at the time of sample selection for some future survey. It is important for investigators to ask themselves how well a GIS could ever predict spatial patterns to inform sampling in their particular application.

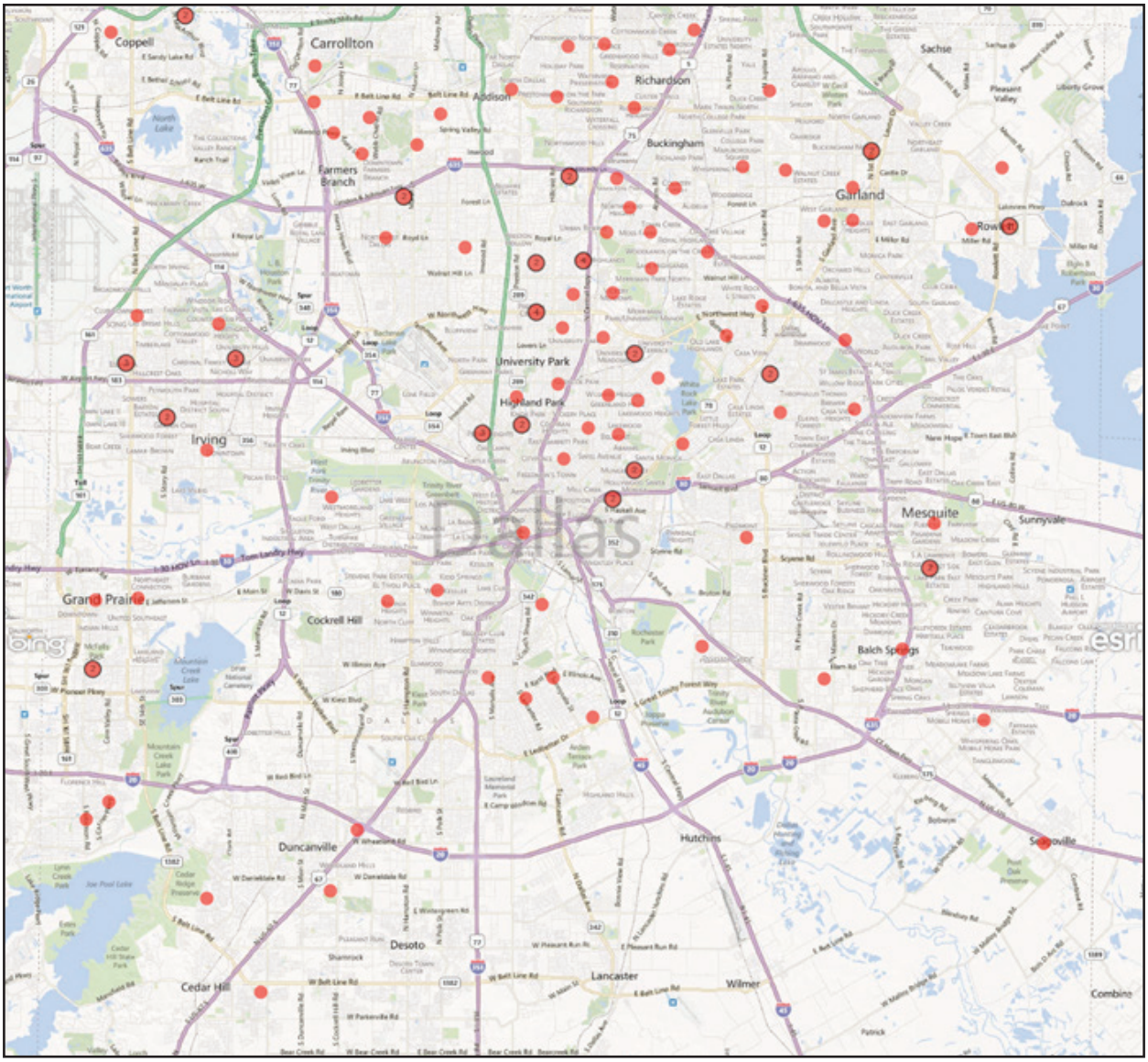
For instance, in the motivating example of this chapter it was important that we ask ourselves whether we can realistically predict what areas are and are not at high risk for dengue, even if we knew that in principal dengue outbreaks appear to have a spatial pattern. In the case of dengue, our confidence relied largely on some particular characteristics of the main vector, *A. aegypti*, that allowed us to have a sense of where its populations would likely be most dense and hence dengue outbreaks most likely.

An easy way to see how fragile this kind of predictive confidence can be is to consider another mosquito borne illness: West Nile virus (WNV), the most common cause of arbovirus disease in the United States. WNV is spread in the U.S. primarily by various species of *Culex* mosquitos, which transmit the disease between birds in much the same fashion that *A. Aegypti* spreads dengue among humans. A mosquito infected during this circular exchange between birds and mosquitos can then infect a human. Like dengue, there is typically a seasonal pattern to human cases, with a peak during the summer months when mosquito populations are largest. There is also a degree of clustering to WNV cases. For instance, in the summer of 2012 there was a serious national outbreak, with Texas being particularly hard hit and, within Texas, Dallas County experiencing a severe epidemic.

Capturing the determinants of spatial patterns to WNV so that areas of severe outbreaks can be predicted is, however, a very tricky business. WNV outbreaks have occurred in places where *Culex* mosquitos were dense on the ground, and where their populations were not particularly large. Interestingly, the severe Dallas outbreak of summer 2012 came at a time of severe drought in the area (which one would surmise would limit *Culex* breeding opportunities). Certain species of bird are particularly associated with WNV, but various measures of the density of their population have not necessarily been very good predictors of outbreaks. Indeed, the role that birds play in transmission, and their often extreme mobility, may in some sense confound prediction of outbreaks.

The figure below illustrates the distribution of West Nile cases in Dallas, Texas. "Clusters" are indicated by a single red dot. The majority of the dots represent just one case. The largest clusters exhibited three or four cases, while the majority of the clusters with multiple cases had just two cases. This can be contrasted with the far more pronounced dengue clusters evident in Singapore as of late April, 2014. It seems evident from these two examples (which were chosen simply for the beauty of their maps, and not because they exhibited particularly different spatial patterns to dengue and WNV cases) that there is far less of a spatial pattern, at least in terms of concentrated cases, to WNV than dengue.

Figure 27. The Distribution of West Nile Cases in the Dallas Area, August 2012



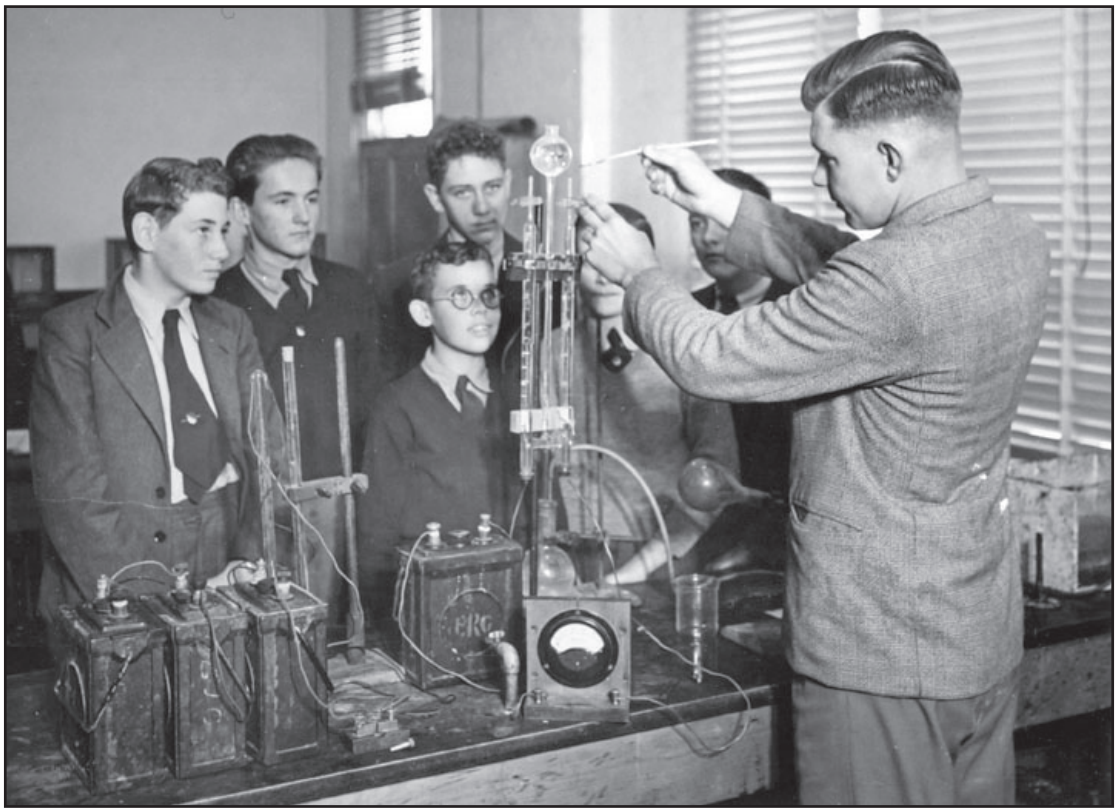
Source: Dallas County Health and Human Services, www.dallascounty.org/department/hhs/WestNileVirusAug7.pptx

The point is that it is hard to see is how spatial patterns to prior risk of WNV could be predicted with anywhere near the confidence that is possible (at least in principle) with dengue—when the epidemiological fundamentals for dengue are in place in a dengue endemic area, the risk of an outbreak is significantly enhanced. The same cannot really be said of WNV. Indeed, it is easy to see how particular circumstances can undercut the value of GIS as a predictive tool in nearly every example provided. For instance, to the extent that the main driving factor in local nomadic movements are instead the highly unstable and usually unpredictable dynamics of inter-clan violence, it is difficult to see how GIS could provide much finer local guidance for sampling their populations.

Even if the spatial patterns to a phenomenon can be predicted through a GIS, one still must balance the costs and benefits of crafting a GIS to inform sampling. Building a GIS can, depending on the circumstances, be expensive and possibly time consuming. Given how pronounced the relevant spatial patterns are, or the precision with which they can be predicted with a GIS, survey investigators should ask themselves whether the cost savings from the enhanced sampling efficiency that the GIS will yield justifies the cost of building that GIS. Sometimes it is simply more cost effective to put up with highly inefficient sampling.

However, it is not hard to envision circumstances where the application of GIS to support stratified sampling will lead to a lower sampling cost per unit of the subpopulation of interest actually successfully sampled. The point of this first principle of practice is simply to appraise the circumstances of one's survey carefully since there are circumstances where either GIS cannot inform sampling (because there is no predictable spatial pattern to the concentrations of the subpopulation of interest) or should not inform sampling. The latter case would presumably involve situations where either it would be prohibitively expensive (financially, time wise or both) to craft the GIS or sufficiently costly, in light of the gains in sampling efficiency, that the application of the GIS would not in fact lower the cost per unit to sample successfully from the subpopulation of interest. Survey investigators need to assess soberly the usefulness of a GIS for sampling in their particular study.

A Science Class in Brisbane, Australia, 1951



Source: Wikimedia Commons, Queensland State Archives, Digital Image ID 1638

Apply a Scientific Standard

Perhaps the other most important principle of practice is to apply an appropriate standard of science to crafting a GIS. At the most basic level, this evokes a fundamental question: what does science mean in the construction of a GIS? We would argue that the fundamental standard of scientific legitimacy for building a GIS should be what it is in many other arenas, such as the canonical setting of the laboratory experiment—the ability to replicate results. In laboratory science, this means essentially that others can repeat an experiment and obtain the same (within some understood reasonable bounds of tolerance) results.

The question is how to apply that concept to the task of building a GIS, at least as crafted to inform sampling. An answer can be found by considering what exactly motivates us to build a GIS to inform sampling in the first place. The advantages of GIS as an instrument to assist sampling are numerous, but at the core of most applications of GIS to sampling the main mechanism by which it can inform sampling is by revealing additional information about the spatial distribution of a subpopulation of interest. This then allows for the possibility of more targeted sampling from that subpopulation.

In other words, we craft a GIS to support sampling because we think that that a GIS could help to reveal the spatial patterns to the subpopulation of interest for the survey at hand. This suggests a simple notion of replicability. A GIS, as so applied, has been constructed in a replicable fashion if another party possessed of the same information sources (maps, databases, satellite photos, etc.), GIS resources (e.g. software, etc.), and following the same protocol would, with a high degree of concurrence, identify the same spatial patterns to the distribution of that subpopulation of interest across some reference area (a neighborhood, city, village, region, state, entire society, etc.).

Clearly, the research team for a given survey can never guarantee that the first two conditions for replicability are met. There is no way to insure that others would have access to the same information available at the time that the GIS to support sampling for a given survey was crafted. Information evolves, and sometimes it is lost altogether, there can be issues related to permissions and access, and other parties might not have the same financial resources that the survey investigators did at the time that they acquired information to form a GIS to support sampling. To the extent that information used in the GIS came from fieldwork, it becomes impossible to insure access to the same circumstances as prevailed in the field at the time that the survey investigators conducted their fieldwork to inform the GIS since on the ground conditions are usually constantly evolving. It is not even possible to insure that they would have access to the same software (which also evolves and changes).

Moreover, it seems unlikely to expect that any other party might even try to replicate the spatial patterns within the GIS applied to guide sampling for a particular survey, even if they could do so from the same basis in information and technology. If nothing else, it is hard to see what the incentive to replicate is in the sense that one exists in the laboratory sciences. So replication is probably not a *practical* possibility.

Nevertheless, in the pursuit of scientific rigor there is still great value in thinking about the *hypothetical* possibility of another team working to replicate the construction of one's GIS. The reason is that it is a useful way of thinking about how to thoroughly and effectively insure the one ingredient of the hypothetical replication process that is under the investigator's control.

Specifically, the notional need to promote replicability can sharpen thinking regarding the protocol for crafting a GIS. The protocol can be thought of as the set of operational rules and practices used in crafting a GIS, given the information available to do so. A clear, transparent, and carefully thought out protocol for building a GIS clarifies exactly what the research team is doing, how they are doing it, and in essence why they are doing it.

Developing a clear protocol forces investigators to think very carefully about what their goals are, and the practical approaches to achieving those goals. With a clear sense of goals, it becomes easier to have a sense of whether a proposed practice or procedure for building a GIS really supports or is consistent with that goal. Since the goal in applying GIS to sampling is to inform the sampling process by revealing information about the spatial distribution

of subpopulations of interest, the process of defining goals involves carefully specifying the subpopulation of interest and marshalling the best science about the kind of indicators potentially apparent through a GIS that might reveal something about the spatial distribution of that subpopulation.

A clear protocol helps to insure that consistent practices and standards are applied in crafting a GIS. Consistency can be a surprisingly elusive ideal in some instances—building a GIS can, depending on the circumstances, be a large, complex and, frankly, sometimes tedious task. The actual process of building a GIS can lead to fatigue, a certain degree of drift and a challenge in maintaining focus. The analogy to a long drive across sometimes uninteresting terrain is not unreasonable, and a protocol in some sense can serve the same objective as a map and a clearly planned route on that map—it can keep the team on track.

Given the wide diversity of potential applications of a GIS to sampling, the most one can offer is advice about the ingredients of a useful protocol that truly supports replicability, even if it is only a hypothetical possibility:

1. Define clearly the subpopulation of interest. This seems obvious, but can be surprisingly hard to do in some applications (as we will see in the next chapter). The definition should be theoretically persuasive (it should have real social, economic, or behavioral meaning and be rooted in the goals of the survey the sampling of which the GIS is meant to inform) but also empirically practical: if the definition of the subpopulation does not lend itself to indicators of the presence of the subpopulation that can be operationalized through that GIS, then there is little use to building a GIS in the first place;
2. Define the indicators of the subpopulation to be captured by the GIS. Once the basic overall definition of the subpopulation is in place, one must identify empirically tractable (i.e., can be measured or captured through a GIS at a reasonable cost and with acceptable accuracy) indicators of the varying presence (or intensity of presence) of the subpopulation of interest across the geographic area of focus for the GIS;
3. Define clearly the procedures for obtaining information on each of these indicators. This means clearly describing and defining the information source for the indicator where the information was pre-existing (e.g., maps, official records, etc.) and clearly documenting the process of indicator construction where the indicator was somehow developed by the investigators (e.g., GIS software algorithms or procedures, fieldwork);
4. Document clearly and precisely the analytical procedures by which these indicators were used to determine spatial variation in the prevalence of the subpopulation.

Above all, a protocol carefully designed to meet these criteria will help the investigators obtain the most scientifically persuasive end product possible, given the constraints that they cannot control (budget, pre-existing information, etc.).

Although the focus of this discussion has been on replicability in the sense of insuring that other parties could estimate the spatial distribution of a subpopulation with a high degree of concurrence with the results obtained by building the GIS for the sampling of a survey at hand, one other guideline could be usefully added: clarifying precisely how the information on the spatial variation in the concentration of the subpopulation of interest was translated into strata of sampling units. In cases where the GIS can be reconciled with a pre-existing frame, such as an official census frame, this amounts to documenting how the information in the GIS was merged into that frame in order to stratify the sampling units within it per the information about subpopulation prevalence supplied by the GIS, as well as the rules that drove stratification. In instances where the GIS cannot be reconciled with an existing frame (typically because that existing frame is not sufficiently georeferenced), it typically becomes necessary to craft sampling units from the GIS itself. The rules and procedures guiding the process of doing so and then stratifying the resulting units should be carefully documented.

Source: Shutterstock

Automate, Automate, Automate

We have alluded to the tedium that can attend the actual process of building a GIS. Doing so often involves repetitive tasks. A few examples:

1. Identifying visual "signatures" on satellite images to identify types of structures, infrastructure, crops, etc.;
2. Building polygons that contain or demarcate phenomena or populations of interest;
3. Reconciling discrepancies between different information sources in a GIS (e.g., when road networks on a scale official map do not quite match what is visibly evident from satellite photography);
4. Identifying errors that might require imputation (errors can occur even with highly digital and automatic information capture—for instance, it is common for radar-based precipitation total estimates to require fixes by imputation here and there for transitory radar disturbances);
5. Inputting information either not in digital form or in digital form that cannot be readily merged into the GIS.

All of these quickly become quite boring to perform (on this point the authors speak from personal experience) but can still require a great deal of effort and focus to perform well.

Moreover, depending on the size (e.g., geographic scope, resolution or granularity of information within, etc.) of the GIS tasks such as these might need to be performed thousands or even tens of thousands of times. A general observation the authors have made over the years is that the magnitude of the task of building a GIS can often be underestimated.

Human beings are not particularly well designed for performing repetitive tasks. We tend to grow tired, lose attention, and begin to drift from the strictly defined procedures of the protocol. In other words, we produce a product often subject to tremendous quality variation and, unfortunately, often of substandard overall quality. This can fatally undercut the advantages of building a GIS by introducing measurement error to the final assessment of the spatial distribution of the subpopulation of interest. In other words, it can introduce noise to the stratification process that, at the limit, could actually conceivably make one worse off by applying the information generated by the GIS to sampling.[5]

The solution to the limited human tolerance for repetition in manufacturing has been automation—turning over repetitive tasks to machines, computers, robots, etc. to the greatest extent possible. There is tremendous scope for automation in the processing of the information in a GIS. GIS software such as ArcGIS is a tremendously powerful tool for implementing sophisticated algorithms to perform all sorts of repetitive tasks and analyses.

Sometimes there is a temptation to make the perfect the enemy of the good when contemplating the possibility of automating a GIS task. Often even an algorithm cannot perfectly perform an automated task in all instances. Usually, this is due to aberrant circumstances that are hard to predict (and hence incorporate into an algorithm). Sometimes the algorithm can in essence make a mistake against what the investigators want it to do. For instance there could be cases where the flawless execution of an algorithm results in an error compared with what a human, possessed of the flexibility of all that human judgment provides, might have done.

[5] For instance, the process of stratification frequently involves creating unequal selection probabilities across strata which, all other things being equal, ultimately leads to greater variance in probability weights and hence final estimates of population-level indicators.

Even in instances where algorithms do not complete a repetitive task perfectly in *every* instance, they can still reduce tremendously the scope of what needs to be done. For instance, it can in many instances be much easier to visually quality check the results of hundreds of tasks performed by an algorithm than to perform those tasks one's self. In some sense, there is a tradeoff between the consistency problems of humans in performing repetitive tasks and the rigidity of algorithms in doing so. The fact that an algorithm might not always yield the preferred output is not necessarily a reason for eschewing them altogether in performing a task.

Be (Reasonably) Uncompromising

Source: iQoncept/Shutterstock

This is more of a philosophical point but one that has arisen at some point in much of the GIS work that we have performed. The work surrounding the construction of a GIS can be a bit messy in practice, and subject to a certain unavoidable degree of error. In the face of this inexorable reality, there can sometimes be a subtle danger of becoming too comfortable with error. Think of this as a kind of slippery slope of GIS work.

Consider an example from a recent conversation over a GIS project in which one of the authors participated. In that instance, the subject was a GIS containing the coordinates of health facilities and the center points of clusters from a household survey. The GIS was intended to link households to facilities in order to characterize the health care supply side confronting them. For instance, one might wish to know how many facilities of a certain type are within one kilometer of a given cluster, or the distance to the closest facility of a certain type from the household.

It had become apparent that some of the facility coordinates might not be particularly accurate. They were recorded with older global positioning units (GPSs) for which there can often be some error in the recording of coordinates. Often this error can be corrected by taking multiple recordings from the same position and then using the centroid of those various coordinates as an estimate of the true location of the position. The question was whether to revisit each facility to take multiple coordinate readings. One suggestion briefly entertained was that error in the facility coordinates might not be that big a deal since the center of each cluster was in some sense just a proxy for the true location of each household within it (the interest lay primarily with characterizing the health care market confronting each household).

This reasoning can seem at first glance reasonable. However, think about what it is really being said: the reality of some error makes further error more reasonable. One does not need to go far down this road before the data at the foundation of a GIS contains more noise than signal.

There are many instances in building and working with a GIS where some degree of unavoidable error arises. Because of this further error that can reasonably be avoided should be avoided. At the end of the day, the goal is to improve the efficiency of sampling, thus lowering the costs of fieldwork. It should always be remembered that modest savings

from compromising with error in building the GIS may be more than outweighed at the survey fieldwork phase by higher costs per observation obtained or fewer eventual observations from the subpopulation of interest. To the extent that discretionary fault is tolerated, it should only be in instances where the costs of repair clearly outweigh any savings that the repair might yield in terms of lower unit costs per observation obtained or a larger final sample from the subpopulation of interest.

Be Ready for Challenges with Sampling Units

Under ideal circumstances, one would typically hope that the information regarding the spatial distribution of a subpopulation of interest yielded by a GIS could somehow be merged into an existing official sampling frame so that the information could then guide the stratification of sampling units in that frame according to the likely prevalence of the subpopulation with each unit. For instance, one might merge such information into an official census frame.

The advantage of an official census frame, even a somewhat dated one, as the basis for primary sampling units is that the variation in the true population size between sampling units (which, following convention both in general practice and for this manual, we once again refer to as enumeration areas, or EAs) is typically modest compared with what one might encounter by, say, attempting to delineate primary sampling units from scratch. The reason for this is that census enumeration areas were typically originally defined in such a way that they all contained roughly the same number of households. While there may be some drift over time from this original balance, it is often safe to assume that the variation in the number of households and therefore, in all likelihood, people between census EAs is no larger, and probably much smaller, than what would emerge from any effort by the survey investigators to generate their own primary sampling units for selection. That said, dated official frames can be risky in terms of the actual population parity between EAs, particularly in highly dynamic settings.

Since the probability weight for each unit of observation is the inverse of that unit's overall probability of selection, lower variation in population between enumeration areas (compared with primary sampling units scratch built by the team) translates into lower variation in weights between units of observation. This can become an important consideration because excessive variation in weights can translate eventually into much larger standard errors to estimates of (sub)population parameters of interest. It is therefore advisable whenever possible to apply the information concerning the spatial variation in subpopulations of interest to official census frames.

Unfortunately, this is not always possible. Merging the information from a GIS into a census frame for stratification requires that that census frame be sufficiently georeferenced to provide a basis for merging between the GIS and the frame. Many official census frames, particularly from lower-income nations, lack geographic reference information such as the longitude and latitude of the reference point for each EA (though this is slowly changing). Instead, the identifiers for the EAs in many census frames are essentially string variables that contain the actual names of the administrative units within which the EA is located (these are sometimes augmented with integer variables that are essentially codes for these administrative units). Sometimes it is possible to introduce these identifiers into the GIS, but this can be an awkward exercise and suffer from a "curse of dimensionality" since particularly lower level administrative units can be quite numerous (running to the tens of thousands in larger societies). Bringing them into the geographic coordinate system that is the common frame of reference for a GIS can be a very time consuming exercise.

When merging between an official frame and the GIS is not possible, a common fall back alternative is to craft primary sampling units within the GIS itself by defining polygons that capture the physical area of each primary sampling unit. In the authors' experience, this is can be an extremely tedious process fraught with peril. Per the first concern (tedium) whenever possible the process should be automated, though there may be limits to this option.

The peril arises from the possibility of primary sampling units that prove on closer inspection of selected units during fieldwork to have wildly different population sizes, leading to tremendous variation in eventual probability weights and hence elevated standard errors for the parameter estimates. The authors themselves have been involved in several attempts at crafting primary sampling units from a GIS. In almost all cases, the method was to use satellite photos to identify "reasonable" units that appeared likely to have similar population sizes contained within them. A number of methods have been used to try to accomplish this, though in most cases it amounted essentially to basing

sampling units on the numbers of structures contained within them and perhaps some ancillary census information that gave a sense of likely population density in different areas. Taken against the goal of crafting sampling units with similar populations, these efforts essentially universally failed.

This may reflect a lack of imagination or sophistication in these exercises, but it also seems reasonable that so defining primary sampling units from scratch is a messy process fraught with peril. And it may well be the case that the efforts made to design sampling units with similar population sizes succeeded in the sense that the counterfactual (not taking explicit care to try to craft primary sampling units with similar population sizes) might have resulted in even wilder variation in selection probabilities, and hence weights and standard errors.

Thus, perhaps it seems reasonable to have some plan to deal with larger variation in the population sizes of the primary sampling units built into the sampling plan. There is at least one immediately obvious potential mechanism for mitigating this problem: proportional sampling of households within selected primary sampling units. In other words, once multi-stage selection gets down to a unit within which households can be listed, a proportion of the households listed would be selected rather than a fixed number of them.

Funny, It Didn't Look This Crowded From Outer Space

Source: leungchopan/Shutterstock

To see why this could mitigate the problem of uneven primary sampling units (PSU), let us assume that we wish to select a sample of households. The sampling plan is to select a sample of PSUs. Within each PSU, households will be listed and from that list a sample of households will be selected. Returning to the notation employed in Chapter 2, for the first-stage probability of selection for a given EA (let's say EA *j*) the probability of selection is

$$P_{1j} = \frac{N_1 \cdot S_j}{S}$$

where N_1 is the number of PSUs to be selected in the first stage, S_j is the size of EA number j (as recorded on the frame) and S is the overall size of the frame. P_{1j} simply stands for the first-stage probability of selection for EA j. This has been posed in terms of probability proportional to size sampling. However, if we craft each PSU so that its size (in terms of population size) is intended to be constant as S*, this becomes

$$P_{1j} = \frac{N_1 \cdot S_j}{S} = \frac{N_1 \cdot S^*}{S}$$

Notice that if we assume that there are k primary sampling units this means that they are all the same size on the frame

$$S_1 = S_2 = \cdots = S_K = S^*$$

Therefore the size of the frame is

$$k \cdot S^*$$

and the selection probability becomes

$$P_{1j} = \frac{N_1 \cdot S_j}{S} = \frac{N_1 \cdot S^*}{S} = \frac{N_1 \cdot S^*}{k \cdot S^*} = \frac{N_1}{k}$$

This is simply the first stage probability of selection that would emerge if N_1 PSUs were selected from the frame by epsem selection.[6]

In the second stage of selection, a listing of households occurs in each selected PSU. Suppose that a fixed number of households H is then selected in each PSU. The second stage probability of selection is then

$$P_{2j} = \frac{H}{H_j}$$

Where H_j is the number of households listed in PSU j. The overall probability of selection for a household in PSU j is then $P_j = P_{1j} \cdot P_{2j}$ and the probability weight for that household is $1/P_j$.

The overall probabilities of selection for any two PSUs A and B are then

$$P_{2A} = \frac{N_1}{k} \cdot \frac{H}{H_A} \text{ and } P_B = P_{1B} \cdot P_{2B} = \frac{N_1}{k} \cdot \frac{H}{H_B}$$

The only reason that these can differ is if $H_A \neq H_B$. If the number of households listed in A and B differ, the selection probabilities and relative weights between them will differ. If these differences are large enough in magnitude it will start to significantly increase the eventual standard errors of estimates of population parameters. The problem is thus selection based on first stage equal probabilities of selection when the primary sampling units are in fact of different sizes.

The easy way to get around this is to select a proportion ϕ of households in each selected PSU rather than a fixed number H. The second stage probability of selection then becomes

$$P_{2j} = \frac{\phi \cdot H_j}{H_j} = \phi$$

[6] In general, probability proportional to size selection is equivalent to epsem sampling when all sampling units have the same size.

The overall probabilities of selection for our two hypothetical PSUs then become

$$P_A = P_{1A} \cdot P_{2A} = \frac{N_1}{k} \cdot \phi \text{ and } P_B = P_{1B} \cdot P_{2B} = \frac{N_1}{k} \cdot \phi$$

and

which is the same across PSUs, returning us to a self-weighting sample with equal probability weights across sampling units.

Proportional sampling can be a bit tricky to execute in practice—it is difficult to know exactly what to set ϕ to initially to insure a large enough final household sample size (because you likely will not know perfectly the average cluster size in advance). However, a reasonable first guess can then be iteratively updated as more information about the true average size of PSUs is learned through the listing process. The final sample might not be exactly self-weighting, but the weight variance should be lower than in the case of selecting a fixed number of households *H*.

Try Spotting This Roof Profile from Outer Space

Source: orxy/Shutterstock

Consider Ground Truthing and Testing

In order to inform efficient sampling, it is important that the GIS be reasonably accurate at predicting the spatial distribution of the subpopulation of interest. Wherever possible, the effectiveness of the GIS as a platform for such spatial prediction should be tested. There are two obvious ways of doing this.

First, in some instances it could be useful to "ground truth." This is a process by which the team building the GIS conducts field visits to verify emergent predictions from the frame. The specifics of this process will depend on the particular application, but given finite resources an inherent general tradeoff must be recognized—the investigators must decide whether it is better to inspect fewer areas more thoroughly or more areas in a more cursory fashion. The former course yields more accurate information about field conditions against which to compare the predictions

of the GIS. However, that course also generates a narrower picture of reality. With a broader focus, one can have a more comprehensive picture with which to assess the GIS, though perhaps at the cost of more noise to the field measurements due to the more cursory inspections.

Second, one could contemplate reserving some of the information that might otherwise contribute to the GIS to see how well the GIS makes "out of sample" predictions. Consider the motivating example of predicting risk of dengue outbreaks. Suppose that investigators had at their disposal a decade's worth of information about the conditions portending outbreak and the actual pattern of outbreaks that emerged in the city. It might be worthwhile to withhold a year or two of the information on actual outbreak patterns from the GIS to see how the GIS, informed by the data from the other years about the link between risk factors and actual outbreak patterns and the risk factor information from the withheld years, predicts the pattern of outbreak that actually occurred in those withheld years.

These sorts of tests can give a strong prior indication of the value of the information yielded by the GIS. This can help the team decide whether the GIS really is more likely to help or hinder the most efficient possible sampling.

Be Careful About "Betting the Farm"

Related to this consideration, investigators should be intimately aware of what factors drive the predictions of the spatial patterns to the subpopulation of interest yielded by the GIS. To the extent that the predictions yielded by the GIS rest on narrow and strong assumptions, the credibility of those predictions is somewhat less convincing than if there were weaker assumptions behind the predictions. At the same time, this is not a call for, for instance, adding indicators only weakly associated with the population of interest or indicators that might otherwise add more signal than noise regarding the spatial distribution of the subpopulation. Rather, it is a call to be conscious of, and careful about, lynchpin assumptions—if they prove inappropriate, the entire effort of crafting the GIS could be wasted and the efficiency of sampling could actually be harmed by the application of information from the GIS to it. A specific illustrative example of this will be discussed in the next chapter.

His Next Decision Could Potentially Be Important

Source: pinkypills/Shutterstock

Do Not Let Your Reach Exceed Your Grasp

It has been suggested by psychologists that human beings often labor under what has been referred to as a kind of "optimism bias." We tend to focus more on promise than peril and place too much probability weight on "good" outcomes in forming our expectations. For instance, it is easy to forget now that many of the best minds of the time were initially convinced that World War I would be a short, decisive, ennobling experience that would not inconvenience anyone too seriously. For that matter, optimism bias has probably been an essential ingredient to most military fiascos throughout history.

Perhaps another manifestation of this bias is that we have a tendency to believe that any new and shiny technology will naturally improve our lives and expand our horizons. We often fail to foresee the ways that new technology might generate challenges. We don't envision that it will help to invade privacy, steal identities, undermine security, blur the line between work and private life, actually *increase* the amount of work we have to do, and undermine the final product quality, etc.

Technological progress does hold a seductive promise in the realm of GIS work. The information sources and technological possibilities for building and maintaining a GIS are constantly expanding. The temptation to harness the latest and greatest technological tools in crafting a GIS is a strong and understandable one (the authors are currently frantically searching for virtually *any* justification to purchase and experiment with unmanned aerial vehicles as a tool for gathering information for building GISs).

However, one must be careful not to get carried away with this possibility. Newer technologies are often essentially untested. And application of theses technologies to actual real world work often reveals logistical or technical complications to their actual successful use that are not immediately apparent. In short, while a newer technology or technical possibility often seems quite promising, the details of real execution can prove to be more than the investigators anticipated.

This having been said, it is difficult to dismiss the possibility altogether of employing an emergent technology when that technology might either create scope for a feature of the GIS not otherwise possible or potentially dramatically lower the cost of building the GIS. A healthy balance can perhaps be struck by careful testing of the new technology before committing fully to its use in building a GIS. This will allow investigators to have a clearer idea of the complexities involved in utilizing the new technology or technique and, assuming that they feel the use of the technology is still justified in light of those complexities, have a better sense of how to plan realistically in the protocol for the application of them.

Time Might Not Be On Your Side

Source: liseykina/Shutterstock

Another area where "optimism bias" frequently rears its dangerous head is in planning for the timing of an activity. It has been our near universal experience that the time required for successful completion of high quality survey work is almost always underestimated. We assume that for many readers this resonates (after all, our readership is probably skewed toward those with a history of and interest in survey research).

What is true of surveys in this sense can be doubly so of GIS work. It all seems so easy: just merge a bunch of information along geospatial lines and it practically creates itself. What could possibly go wrong?

Unfortunately, the building of a GIS is a process, one that can take some time and involve unexpected twists. We have already alluded to the possibility that it can involve a large number of repetitive, time-consuming tasks. For instance, creating a single polygon in a GIS software such as ArcGIS is typically fun and quick. Generating several thousand of them is typically neither fun nor quick, even with harnessing the power of automation (at a minimum they still need to be quality checked). One should not plan optimistically when thinking about a timeframe for a GIS but instead do so conservatively. This will allow for a careful process (as opposed to a frantic mess) that does not threaten to derail the timing of the survey, the successful sampling of which depends in part on the completion of a quality GIS.

Aside from budgeting appropriately for what you anticipate, allow some time for the unforeseen. There are numerous other ways that the building of a GIS can turn out to be a more time-consuming process than even conservative planning anticipated. One cannot take for granted that all of the information sources intended to inform the GIS will be immediately available (regardless of promising initial indications) in a useful format—it may take time to gain access to the information, tedious work might be necessary to modify it into a form readily mergeable into the GIS, and so forth. The initial plan for the GIS may also come to require midcourse modification. It may turn out that some planned information sources were not of the anticipated quality or in some other way did not live up to expectations. This could require a search for alternative sources or even recourse to fieldwork. It might turn out that the GIS as originally conceived did not have the predictive power in terms of the spatial patterns for subpopulations of interest, forcing a similar rethink of the GIS design.

Any of these possibilities can render even an initially conservative schedule unrealistic. In light of this, perhaps the final message is to plan conservatively and even then allow some flexibility for the possibility of an unexpected twist. If it turns out that events proceed quickly and smoothly and the GIS is completed ahead of schedule (and, dare we hope, under budget) then consider the luxury of the position compared with the frantic crisis-driven process too many survey preparations become.

A GIS Can Be A Tool That Does Many Things

Source: Volodymyr Krasyuk/Shutterstock

Harness the Power of Leveragability

Building a GIS can be an expensive and time-consuming business. This is justified either when the savings from more efficient sampling from building a GIS still justify these outlays of time and funds, or in the limiting case when the sampling simply could not occur without crafting a GIS to inform it. Whatever the case, a GIS can represent a large fixed cost in the survey process.

Whenever possible, investigators should consider the possibilities for leveraging the GIS. In some cases this involves looking for partners to assist in the construction of the GIS with financing, technical inputs or some other contribution. This helps to spread the cost of the building of the GIS across various parties, making it easier to justify the expense for any one party given the particular purpose to which they will put the GIS.

Even if partnership opportunities are not available, survey investigators can still serve a greater social good by, wherever possible, crafting a GIS that has the widest possible usefulness beyond their own sampling application. For instance, given that they are going to the trouble of building the GIS anyway, the marginal cost of obtaining additional information that might render the GIS far more widely useful to many other parties might be comparatively modest. Where opportunities like this present themselves, they should be seriously considered.

Source: Wikimedia Commons, GNU1.2

An Approach to the Famous but Highly Secret "Area 51" in the Nevada Desert

Be Mindful of Local Sensitivities and Laws

The authors of this document are citizens of the United States. As such, we can assure the reader that there are both more and less sensitive places in the U.S. from the standpoint of observation. Above, we provide an image of an approach to the "Area 51" facility, a remote and secluded detachment of Edwards Air Force Base in Nevada, the purpose of which remains largely secret (according to the mainstream media it is a clandestine testing facility for new weapons; to legions of conspiracy theorists it is where the U.S. government keeps materials captured from alien visitors to our planet). Whatever its use (and we are personally betting that the true reality is probably rather disappointingly boring in the grand scheme of things), displaying an inordinate interest in Area 51 might lead the U.S. authorities to start to take an inordinate interest in you.

The U.S. is not alone in this. Most societies have spatial areas that are for one reason or another quite officially sensitive. Often, as in the case of Area 51, these are military facilities, the seclusion and privacy of which are an article of national law. However, they can sometimes be what seem to be rather mundane areas that are even in plain sight in the midst of busy urban areas.

A nuanced understanding of local governmental circumstances is quite important in such cases, because even official sensitivities are not necessarily codified into law. As a high school student on a school trip, one of the authors sat at the base of the Washington Monument in Washington, D.C. and tested a pair of binoculars by looking at the White House, the viewing of which is not officially prohibited. As the minutes passed, it became clear that a group of security officers had formed on the roof looking back at him. Before long, some sort of law enforcement approached and briefly (and politely) inquired after the author's purpose in looking at the White House through binoculars. Satisfied with the author's response, the officers left and the group on the roof dispersed, but the anecdote does illustrate that not every official sensitivity regarding observation within a society is codified into clear and apparent legal code.

Aside from official sensitivities, one must also be aware of cultural ones. Some groups within society are quite sensitive about being observed. Sometimes this sensitivity can be to observation by means of modern technology. For instance, there are communities in the U.S. that essentially eschew modern technology, often for religious reasons. A good example are the Amish, who live in communities clustered throughout the United States, with a particular concentration in the vicinity of Lancaster County, Pennsylvania.

Whatever the case, investigators should learn something about the society in the context of which they are contemplating building a GIS. Doing so can avoid legal friction and uncomfortable cultural circumstances that could either end or complicate the present GIS activity or do the same to future projects. In building a GIS, a useful guideline is to adhere to the kind of sensitivities regarding, for instance, privacy that inform most modern survey work.

Chapter 5. Case Study: A Survey of Urban Health in Bangladesh

In the last chapter we provided some motivating examples of how GIS could inform sampling. From these it was clear that the potential applications of GIS to sampling are wide ranging. Indeed, it is impossible to offer truly specific guidance about how to create a GIS for every situation in which one could inform sampling. However, there was a common thread that ran through all of the examples, and almost by definition must be a feature of virtually every instance where GIS can inform sampling: there must be a subpopulation of interest to the survey that has a meaningful spatial distribution that can be predicted, at least to a degree, through the GIS (or at least predicted better through a GIS than through whatever pre-existing frame is available to support sample selection).

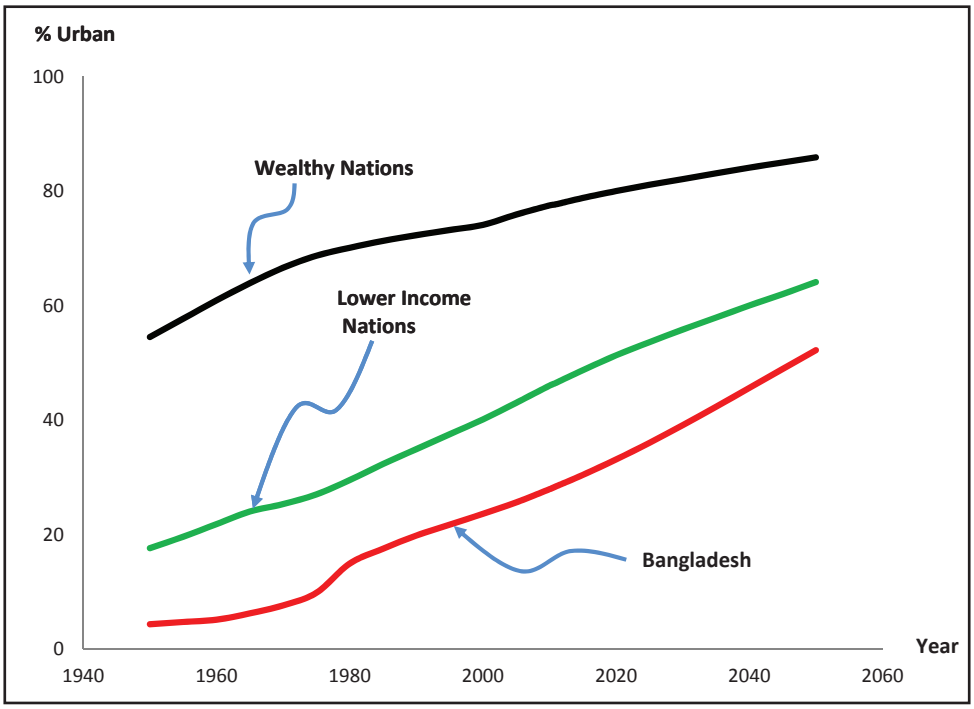
Starting from this basis, we then outlined a series of principles that should inform best practice for the application of a GIS to sampling. These are intended to motivate the questions that should be asked when thinking about building a GIS to support sampling, as well as provide some framework for thinking about the standards to which the construction and application of a GIS should adhere.

In this chapter we offer a detailed case study of one instance where a GIS was used to inform sampling for a specific survey, the Survey of Urban Health conducted in 2006 in Bangladesh. While this is one specific example (and hence might not in its specifics be particularly relevant to potential applications of GIS to sampling that the reader might encounter in their own work), our hope is to provide a more concrete sense of how to work through the process of building and applying a GIS to sampling and how the principles laid out in the last chapter can inform the process.

Our example involves a survey of urban health that was conducted in Bangladesh in 2006, with thinking about the process of sampling for the survey beginning in late 2004 and early 2005. The goal of the survey was to develop a broad health profile for the urban population of Bangladesh. The motivation for doing this was the prospect of rapid urbanization in Bangladesh—the percentage of the population of Bangladesh estimated to live in urban environments is expected to rise from around 25 percent in 2005 to just over half by 2050. This is a process that is playing out globally, but is most pronounced in less developed nations such as Bangladesh, as illustrated in the figure on the following page.

The urbanization of Bangladesh means that urban spaces in that country will be a particularly important focus for programming to improve health and other dimensions of human welfare. There was thus a special urgency to learning more about human welfare in the cities of Bangladesh. Toward that end, the USAID Mission in Dhaka and the Government of Bangladesh (as represented by NIPORT, the National Institute of Population Research and Training) asked the MEASURE Evaluation project and the International Centre for Diarrhoeal Disease Research, Bangladesh (icddr,b) to partner with them on a survey to develop a detailed health profile for the urban population of Bangladesh. Such a profile would help to identify emergent programmatic priorities, allowing for more efficient and effective use of scarce resources for health promotion programming in the urban setting.

Figure 28. Urbanization in Wealthy and Lower Income Nations and Bangladesh



As part of this, one population that was identified as being of particular interest were urban residents living in concentrated, densely settled pockets of urban poverty and environmental vulnerability. In short, the population living in slums was of special interest. This was quite sensible—they were likely among the most vulnerable urban residents from a human welfare standpoint. The formal goal was thus to conduct a survey capable of generating estimates of key health indicators representative of slum and, for comparison, non-slum populations.

This immediately presented a potential problem: there really were no reliable separate frames for slum and non-slum populations in Bangladesh. There had been several earlier censuses of slums conducted by the Dhaka-based Center for Urban Studies (CUS), the most recent being in Dhaka in 1996. The 1996 effort, like those before it, seemed to support two pieces of conventional wisdom about the slums of urban Bangladesh or, at the least, those in Dhaka, the largest city in Bangladesh:

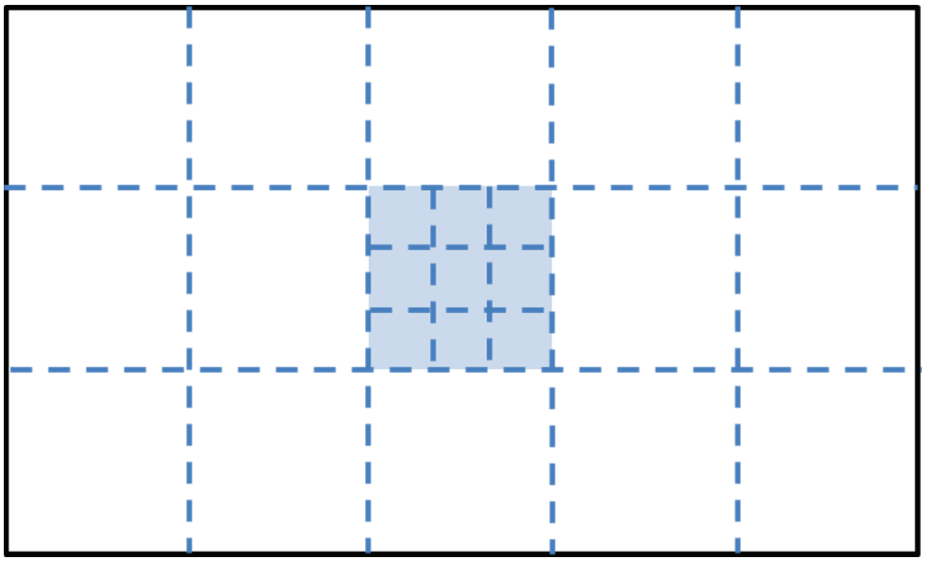
1. The slum populations were largely concentrated in "squatter settlements," which referred to informal slum clusters with no tenure rights, typically illegally sited on publicly owned tracts of land;
2. The slum population represented a significant minority of the urban population, but the physical space in which they resided represented a tiny proportion of the land area of a city.

This second stylized fact of slums was in some sense a reflection of the very high population densities typically found in Bangladeshi slums—this virtually guaranteed that the proportion of physical space occupied by slums in any given city was likely much smaller than the proportion of that city's population residing in slums.

For now we focus on this second piece of received wisdom regarding Bangladeshi slums, because it represented both a challenge to traditional sampling methods as well as a potential opportunity to harness the power of a GIS to overcome that challenge. It was clear that the subpopulation living in slums as of 2005 was likely not uniformly distributed spatially, but instead settled in dense slum clusters. In other words, there were likely spatial patterns to the distribution of slum populations.

One possibility for obtaining samples of slum and non-slum populations would simply have been to perform straightforward selection of a sample of primary sampling units (PSUs) from the official census frame for the study cities and then, in selected PSUs, list and select households. As originally designed, the official, census-based sampling frame for Bangladesh was intended to contain enumeration areas (EAs) that were roughly the same size in terms of the number of households within each of them. An immediate general implication of them is that a sample of EAs could be selected from the frame by epsem (i.e. equal probability) sampling. A more particularly important implication for this application was that, if the EAs were in fact truly the same size (by household population), then there should be more EAs per square mile (or square kilometer, but at any rate per unit of area) where the household population was dense than where it was sparser. Thus, if there should have been twice as many households located in one square mile tract as the other, there should have been twice as many EAs in that tract as the other.

Figure 29. Slum Sampling from a Census Frame: EAs with Equal Population Sizes



This is the situation captured in the preceding figure. It shows a simplified hypothetical[1] example of a representative or typical example of a tract of a city. The white background areas are less densely populated non-slum areas while the blue shaded area represents a much more densely populated slum. The dashed areas represent the boundaries of census EAs. Notice that there are 14 EAs in the non-slum areas and 9 in the slum, for a total of 23 EAs. If the same number of households were in each EA, we would conclude that slum households are 9/23 (39 percent) of those in the tract (and, since this hypothetical tract is mean to be "representative," that would mean that the slum households would be approximately 39 percent of the total number of households in the city).

Epsem sampling from a frame where the EAs indeed each had about the same number of households should yield a sample where the slum populations were represented in proportion to their actual presence in each city.[2] Epsem selection from a frame where the EAs all had the same number of households in them would yield a sample of EAs where approximately 9/23 of which were slum EAs. In other words, straightforward selection from the census frame would yield a sample of slum EAs that was "reasonable" in terms of the presence of slum dwellers in the city.

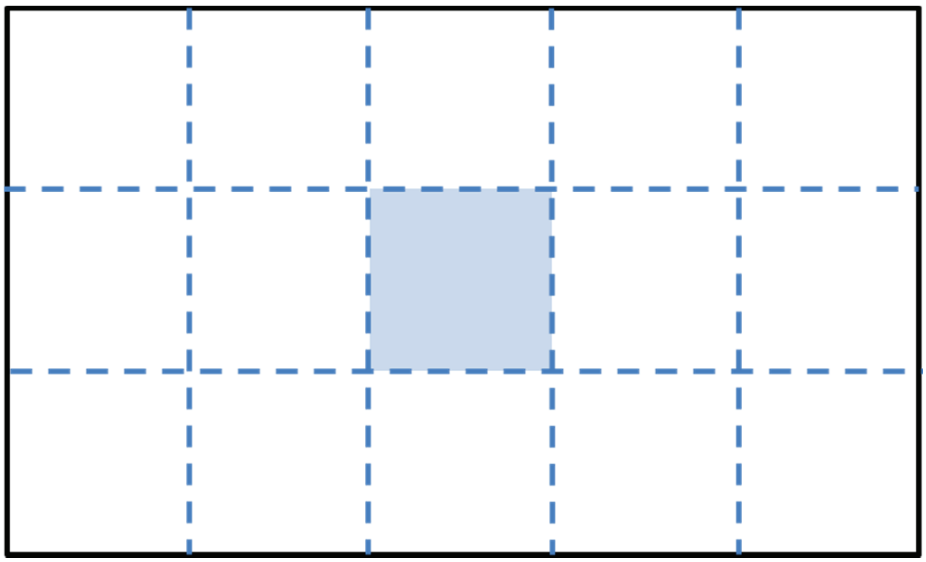
[1] In actuality, census EAs are typically not perfect squares. Rather, they come in all sorts of shapes, with boundaries associated with distinct physical features of the landscape, such as roads, walls, canals, etc. The need to base boundaries on recognizable local features means that it is impossible in practice to have exactly the same number of households in each EA. The emphasis instead was on establishing EAs with *nearly* the same number of households.

[2] Here we assume that the population within an EA is proportional to the number of households within it, and that the proportional relationship is the same across slum and non-slum areas. That is actually not exactly true, but it is sufficiently close to being true that we can ignore the complication.

Moreover, to the extent that similar sized samples were required in slum and non-slum areas for the purpose of yielding sufficiently precise estimates of health indicators for the respective populations within them, a sufficiently large sample size for slums could be obtained without absurdly overshooting the EA sample size required simply to meet the target for non-slum sample size. For instance, suppose that the final sample needed to be 50 slum and 50 non-slum EAs. A possible strategy would be to select enough EAs overall to insure (or at least be highly confident) that 50 slums were selected. Then, one would need to select around 128 EAs to insure a final sample of 50 slum EAs. (Field screening could identify the 50 slum and non-slum EAs from the 128 selected.) This is overshooting the 100 EAs that would be required if slum dwellers were half of the population of each city, but not wildly.

Significant trouble might arise if the EAs of the official census frame did not in fact have similar numbers of households in them (regardless of whether the stated household populations for them in the frame suggest that they do have, or they originally had, similar household populations). Consider the next figure. Suppose that this represents a tract of the city that was entirely less densely settled non-slum areas at the time that the EAs (the boundaries of which are again indicated with dashed blue lines) were demarcated. Then, at some point after demarcation, a slum was formed in the light blue shaded square. The slum contains 1/15th (.0667%) of the EAs in this area but, continuing our assumption about the relative density of the household population in slums compared with non-slums, 39% of its household population. So epsem sampling from the frame would yield a sample where around 6.7% of EAs were slum EAs, against 39% households of the city actually located in slums.

Figure 30. Slum Sampling from a Census Frame: EAs with Unequal Population Sizes



For the purpose of calculating city-level health indicators (aside from separate slum and non-slum population-level health indicators for each city, the study team was also tasked with producing city-level indicator estimates) this would lead, other things being equal, to incredibly uneven probability weights. More importantly, assuming roughly similar final sample sizes required for slum and non-slum populations, the study team would have to sample far more EAs to fulfill the slum sample size requirement than in the case where each EA had had the same number of households in them. If the required sample sizes were 50 slum and 50 non-slum EAs, it would now be necessary to select around 750 EAs to obtain an eventual sample size of 50 slum EAs.

The point of these two simplified examples is that the official census frame is not an efficient vehicle for obtaining slum samples to the extent that slums are relatively densely populated and the census frame EAs do not have a roughly similar number of households in each of them. Unfortunately, while the EAs of the census frame may have indeed had roughly similar numbers of households in them at the time that they were demarcated, all indications suggested that the incredible dynamism of urban Bangladesh (particularly Dhaka and Chittagong) had likely conspired to

undo that original balance. Anecdotal evidence, examination of the sampling experiences of earlier surveys (such as the recurrent Demographic and Health Surveys), and field spot checks of the census EAs suggested that there was great variation between them in terms of how many households they contained. This portended the possibility that the team would need to overshoot, perhaps tremendously, on census EAs to achieve a given sample of slum EAs.

In practice, the prospects for sampling from the census frame were even messier. For instance, in the hypothetical examples presented the EAs were internally homogeneous—either entirely slum or entirely non-slum. In reality, it was thought that by 2005 small slums were becoming increasingly prevalent, and there could potentially be several of these per EA interspersed among the rest of the EA, which might be non-slum. Even in the comparatively simpler times of the 1996 census (where the distribution of slum typologies was more slanted toward larger squatter settlements) it was evident that there was the problem of EAs that were partly slum and partly non-slum. All of these complications made it very hard to form a reasonably confidence inspiring prior estimate of how many EAs would need to be selected to obtain given sized eventual samples of slum and non-slum residents.

The presumption in these two examples is that slums would represent distinct clusters within each study city, and in some sense it had to be the case that they would in fact present as such—most definitions of slums involve the idea that they are distinct, densely settled pockets within the urban environment. So in some sense it already seemed likely that there was a spatial pattern to the distribution of the slum population: they were concentrated in slum clusters.

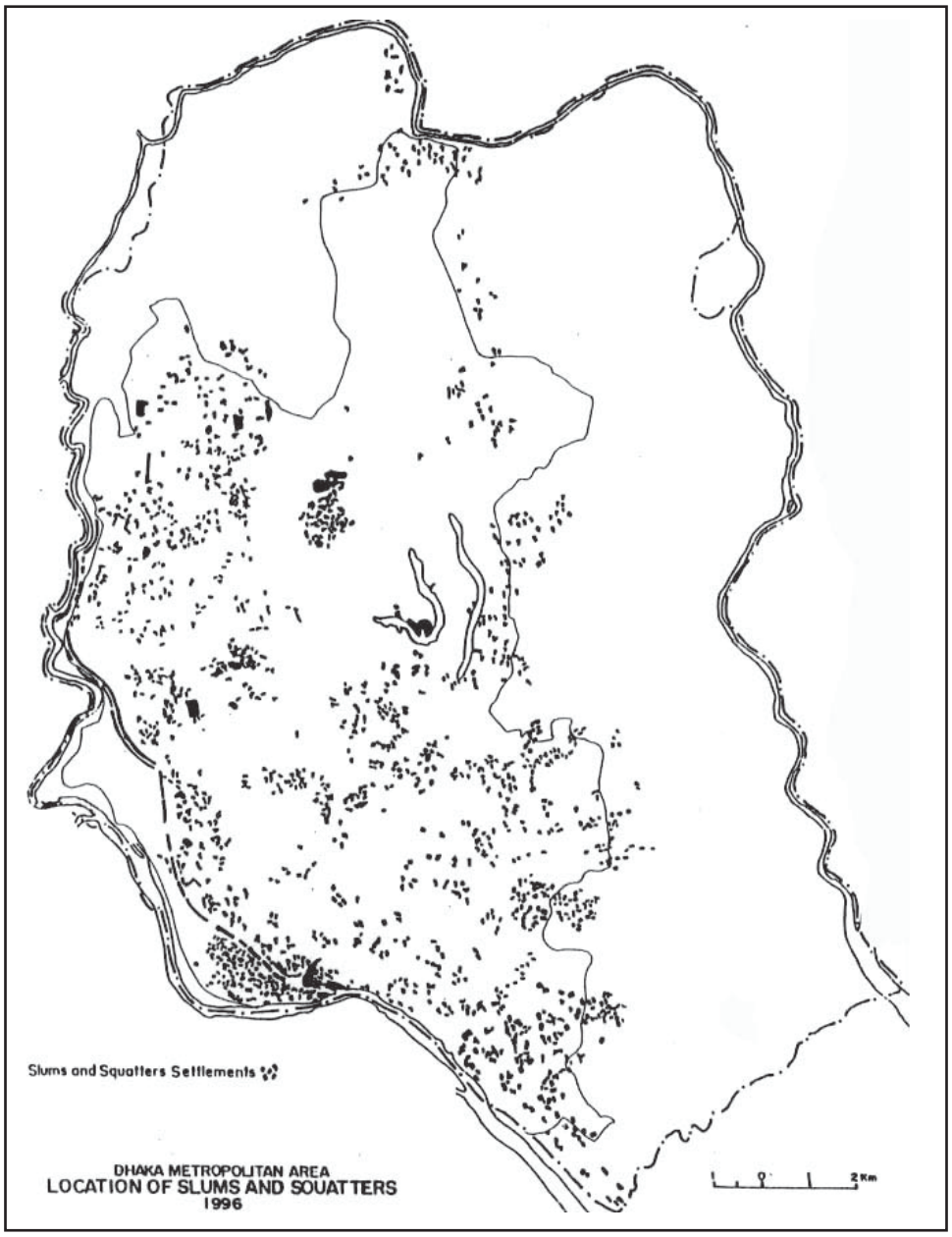
However, available evidence suggested that even then the clusters were unlikely to be evenly spatially distributed. For one thing, the 1996 Census had produced a picture of uneven distribution of slum clusters across Dhaka. The figure on the following page shows a map of the slums of Dhaka from that 1996 Census. Notice that there are large pockets of Dhaka where there were essentially no slum settlements/clusters, while in other areas there were many of them. Although the cities of Bangladesh are subject to a great deal of dynamism, and it was possible that the distribution had become more even by early 2005, even a casual drive around Dhaka or Chittagong in early 2005 revealed that the general pattern of unevenness likely hadn't changed since 1996: in some areas slum-like concentrations seemed to be everywhere, while in other places within these cities they were quite rare.

The situation that was emerging was that straightforward selection from the census frame might not lead to efficient sampling, and even then the complicated realties of the field (many EAs were only partly slum) would make estimation of an appropriate number of EAs to select very difficult. At the same time, there appeared to be pronounced spatial variation to the distribution of slum populations. This appeared to be a circumstance where GIS might inform sampling.

It was clear that it would be difficult to merge the official census frame to a GIS. Most of the identifiers in the census frame were essentially administrative units, the lowest level of which might contain numerous EAs. The maps for the EAs were generally hard copy, hand drawn and not necessarily to scale. These complications would make it difficult to meaningfully merge the EA list into a GIS where the main basis for merging would be geographical coordinates.[3] It was therefore decided that the entire sampling process might be conducted from a GIS.

[3] One could potentially build the administrative units into the GIS, a step that actually was eventually taken, but this would not necessarily allow the EAs, several of which could be in the lowest level administrative unit, to be meaningfully merged into the GIS.

Figure 31. The Slums of Dhaka in 1996



What remained was simply to determine how a GIS could usefully capture the spatial pattern of slum settlements. To facilitate this discussion, the Center for Urban Studies (CUS), a premier urban geography organization, was invited to join the team. This was critical because the CUS researchers brought with them an incredibly rich knowledge of slum life in Bangladesh. Among other things, CUS had conducted the earlier censuses and mappings of slums in Bangladesh, such as the 1996 effort in Dhaka.

With CUS on the team, the design of the protocol for developing a GIS to support slum mapping began in earnest. There was much to discuss. A first surprise came from CUS itself. At the time, conventional wisdom, largely based on the earlier work of CUS, held that most slums were in fact informal squatter settlements on public land. The classic example of such a slum was Korail, a massive informal settlement in Dhaka located on land originally owned by the publicly owned Bangladesh Telephone and Telegraph company (ownership had passed between several public entities since the founding of the settlement).

Figure 32. The Korail Slum

In some sense, this provided one possibly convenient way of locating slums—they should be located on public land. Our CUS colleagues informed us, however, that they felt anecdotal evidence suggested a profound shift, with far more rigorous enforcement of property rights by the state and state entities (in other words, there was a trend toward clearance of squatter settlements). Casual inspections by CUS researchers also suggested that there appeared to have been a great deal of slum formation on private land in recent years. It would thus have been risky to rely on a search restricted to public land.

A more general slum search approach was in order. It was decided that satellite imagery from the study cities would be used to identify likely slum clusters throughout those cities. Bangladeshi slum settlements frequently exhibit what might be called typical roof typologies, which can vary from city to city and even within cities, but still provide a means for identifying possible slum concentrations from satellite photos. For instance, particularly in larger cities such as Dhaka or Chittagong, the roofs in slum settlements are frequently corrugated steel. Concentrations of corrugated steel roofs present a very distinct visual pattern on the photos. Thus, simply by visually analyzing the photos one could identify many likely slum settlements.

The Corrugated Steel Roofing Typical of Bangladeshi Slums…and Other Establishments

Satellite imagery clearly had other benefits as well. Satellite imagery could also be quite useful for detecting errors in spatial information sources like official maps, which are often quickly rendered obsolete by the evolving circumstances of the highly dynamic Bangladeshi urban environment. This was thought to be potentially quite important given anecdotal evidence of the formation of impromptu settlements, particularly on the outskirts of the larger cities such as Dhaka. It was useful for capturing recent road and infrastructure formation as well (i.e., physical features that might, for instance, be useful as reference landmarks for eventual survey field teams).

However, satellite photos also presented limitations. Relying on just satellite photographs was likely to miss many slum clusters and misidentify some clusters as slums. Though larger squatter settlements had been particularly prominent components of the overall profile of slums in earlier CUS mapping efforts (such as the 1996 Dhaka mapping), the researchers at CUS speculated, based on their own informal observations, that circumstances had changed since 1996 and smaller slums on private land had become far more important. This was a surprising possibility since most of the slum literature on Bangladesh at the time seemed to emphasize the squatter settlements, in part owing to the historical findings by CUS. Satellite photos would be likely to miss some smaller slum settlements. The CUS researchers were also somewhat concerned about the possibility of missing those slum clusters subject to the most dynamism (i.e., rapid formation and dissolution). Finally, the CUS researchers warned that many slum clusters identified from satellite imagery (for instance by visual signatures that focused on factors like corrugated steel roofing) might prove not to be slums—they pointed out that many light industrial and food processing facilities at the time would exhibit a similar size and roof typology.

Clearly, the satellite photographs alone would likely prove insufficient for effectively identifying slum and non-slum clusters in the study cities. Additional information was needed. The prospects for merging into the GIS other, pre-existing information sources that might predict the presence of slums in a given area were rather slim. The information regarding potential slum related indicators (poverty rates, population density, etc.) was often no more georeferenced than the census frame. In cases where the indicator was derived from a sample such as the DHS, it did not represent

a true aggregate for an area but instead essentially a proxy based on the one or two survey clusters that might be available within a given geographic area. There were often large geospatial gaps in the information on possible indicators or no information at all. Finally, and perhaps most importantly, the granularity of what information on spatial variation in slum related indicators was available was not fine enough to be useful; urban Bangladesh is incredibly densely settled and slum circumstances can vary tremendously even within very small geographic spaces. Information aggregated to much larger administrative units is thus of limited value since it misses the considerable variation in slum circumstances likely to present *within* those administrative units.

Given the limitations of satellite photograph analysis and the lack of compelling pre-existing ancillary information sources for detecting slum concentrations, it was decided that the slum mapping effort would require fieldwork. The fieldwork was referred to as "ground truthing," the purpose of which was to confirm that possible slums identified through analysis of the satellite photos were indeed slums and to find any slums missed by visual inspection of the satellite photographs.

As the prospect of fieldwork was contemplated, a critical consideration quickly became important: once one moves beyond mere identification of visual signature patterns on satellite photographs (such as concentrations of corrugated steel roofs), a more compelling definition of a slum is required. Slum definitions are a tricky business, and it was clear to the team that the extent of the slum population identified by the mapping would depend heavily on the slum definition developed. Moreover, the slum definition needed to be empirically tractable in the sense that it would hinge on characteristics that could be readily assessed in the course of fieldwork to identify them. Above all, the definition had to be transparent and sufficiently specific to support a scientific standard. The goal was to insure that another mapping team, with the same resources, time frame, and slum definition would arrive at a set of identified slum clusters that would overlap heavily with those found by the Urban Health Survey team.

Slums were defined as settlements with a minimum of 10 households or a mess unit with a minimum of 25 members and:

- predominantly very poor housing;
- very high population density and room crowding;
- very poor environmental services, especially water and sanitation;
- very low socio-economic status;
- lack of security of tenure.

The final criteria was motivated by the historic importance of squatter settlements sited illegally (i.e., without legally secure tenure rights) on (typically) publicly owned land. To qualify as a slum, an urban community had to meet at least four of these criteria. For each of these criteria, a series of specific empirical thresholds were developed. For instance, very high population density and room crowding in practice required settling on threshold figures for each. The thresholds for density used were 300 persons per acre (or 751 persons per hectare) for overall settlement density and three or more adults of mixed sexes per room, 37 sq. ft (or 4 sq. m.) per person floor space, and predominantly (over 75 percent of units) single room family occupancy. If more than 50 percent of the dwellings in the cluster had three or more persons to a room, it was regarded as being characterized by room crowding.

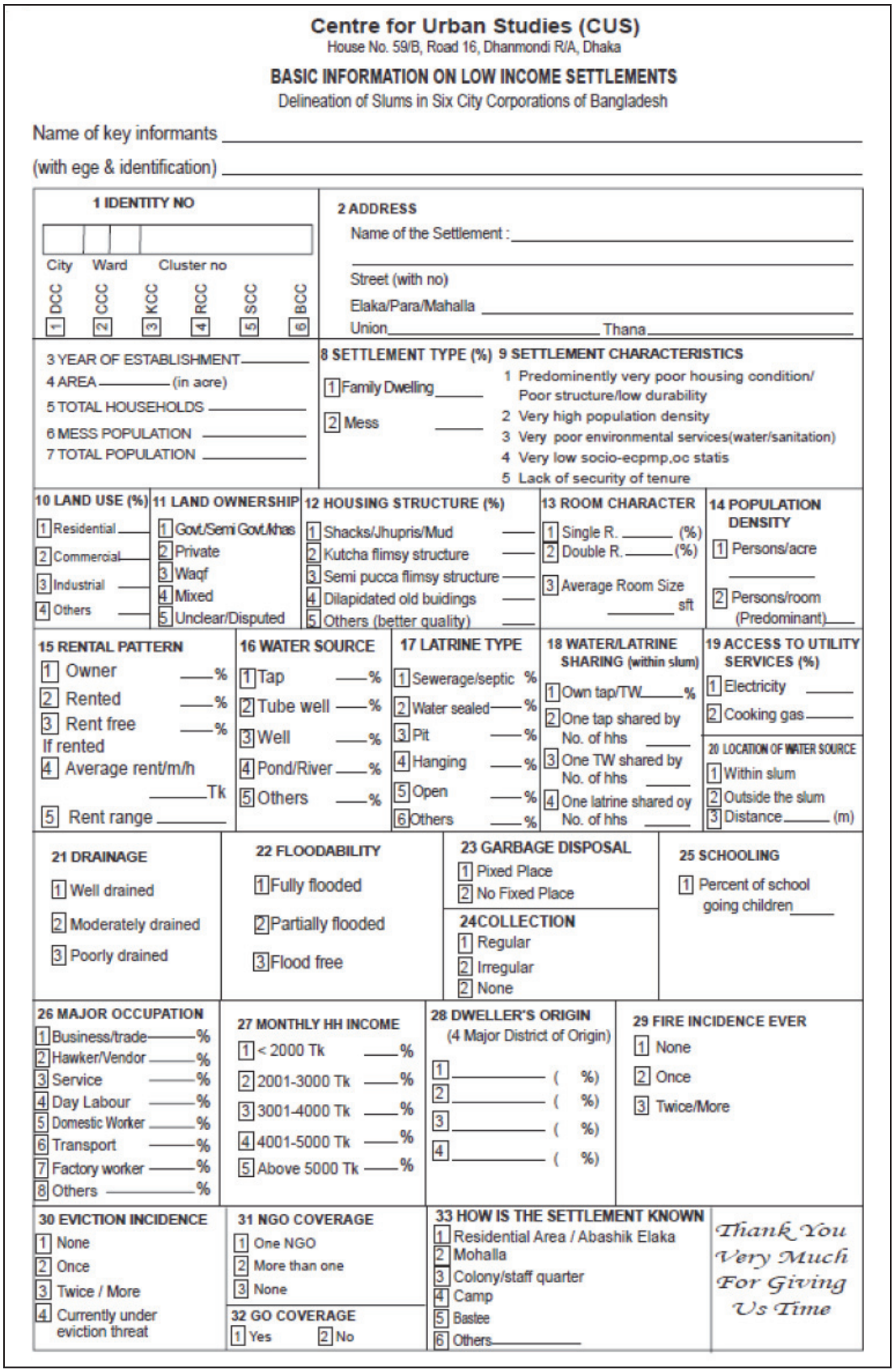
The criteria and specific thresholds were developed with appeal to the literature (international and Bangladeshi) and recent field visits to assess evolving conditions, but also with a consciousness of what could reasonably be expected to be quickly and reasonably accurately assessed by field teams. In many cases these thresholds were actually tested through exploratory field visits. One of the major lessons of this exercise was that a definition that is practical (i.e., one that can be operationalized in the field quickly and accurately) is as important as one that it theoretically perfect (which no one would have been able to agree on in any case). As Voltaire noted, the perfect can be the enemy of the good.

The actual instrument of slum characteristics to be collected for each slum was somewhat more elaborate than the bare minimum required to operationalize the slum definition adopted for this study. Relatively early in the planning process it became clear that the identification and mapping of slums in the study cities was going to be a fairly

involved process that would consume a great deal of time and involve a large financial outlay. Indeed, the sheer scope of the challenge in a mega-city such as Dhaka is simply breathtaking. However, the material point was that challenge *would* be met, creating an occasion on which each slum in the study cities would be visited by a field team. In that sense, a visit to each slum was akin to a fixed cost from the Urban Health Survey study team's perspective.

It was therefore decided that, wherever possible, the instrument for the mapping of slums would be augmented by questions that were easy to collect but might have some usefulness beyond the motivation for the study itself (i.e., to craft a sampling frame of slums). There were two obvious potential alternative uses for the outputs from the mapping exercise. First, the information collected in the slum mapping could be used to provide a brief but highly useful community-level profile of slums in the study cities. This information revealed by such a profile could prove quite useful to, for instance, programmatic planning and policy. Second, the slum lists and maps so generated could allow for far more precise targeting of programs to slum communities. This could save a tremendous amount of resources for a wide range of human welfare related programs that would otherwise have to be committed to the task of simply finding the slum communities they intended to serve. Indeed, the outputs from the slum mapping held the potential for programs to rapidly isolate only certain *types* of slums in which they wished to conduct their operations. Where possible, instruments were therefore added that did not increase significantly the marginal cost of fieldwork but did enhance the value of the information retrieved from the standpoint of these two objectives. By making these additions, it was hoped that the outputs from the slum census and mapping would be as widely useful as possible, enhancing its social benefit in light of the high time and financial costs of conducting it at all.

Figure 33. The Slum Survey Instrument



Centre for Urban Studies (CUS)
House No. 59/B, Road 16, Dhanmondi R/A, Dhaka

BASIC INFORMATION ON LOW INCOME SETTLEMENTS
Delineation of Slums in Six City Corporations of Bangladesh

Name of key informants _____

(with ege & identification) _____

1 IDENTITY NO
City Ward Cluster no
1 DCC 2 CCC 3 KCC 4 RCC 5 SCC 6 BCC

2 ADDRESS
Name of the Settlement : _____
Street (with no)
Elaka/Para/Mahalla _____
Union _____ Thana _____

3 YEAR OF ESTABLISHMENT _____
4 AREA _____ (in acre)
5 TOTAL HOUSEHOLDS _____
6 MESS POPULATION _____
7 TOTAL POPULATION _____

8 SETTLEMENT TYPE (%)
1 Family Dwelling _____
2 Mess _____

9 SETTLEMENT CHARACTERISTICS
1 Predominently very poor housing condition/ Poor structure/low durability
2 Very high population density
3 Very poor environmental services(water/sanitation)
4 Very low socio-ecpmp,oc statis
5 Lack of security of tenure

10 LAND USE (%)
1 Residential _____
2 Commercial _____
3 Industrial _____
4 Others _____

11 LAND OWNERSHIP
1 Govt/Semi Govt./khas
2 Private
3 Waqf
4 Mixed
5 Unclear/Disputed

12 HOUSING STRUCTURE (%)
1 Shacks/Jhupris/Mud _____
2 Kutcha flimsy structure _____
3 Semi pucca flimsy structure _____
4 Dilapidated old buidings _____
5 Others (better quality) _____

13 ROOM CHARACTER
1 Single R. _____ (%)
2 Double R. _____ (%)
3 Average Room Size _____ sft

14 POPULATION DENSITY
1 Persons/acre _____
2 Persons/room (Predominant) _____

15 RENTAL PATTERN
1 Owner ____%
2 Rented ____%
3 Rent free ____%
If rented
4 Average rent/m/h _____Tk
5 Rent range _____

16 WATER SOURCE
1 Tap ____%
2 Tube well ____%
3 Well ____%
4 Pond/River ____%
5 Others ____%

17 LATRINE TYPE
1 Sewerage/septic %
2 Water sealed ____%
3 Pit ____%
4 Hanging ____%
5 Open ____%
6 Others ____%

18 WATER/LATRINE SHARING (within slum)
1 Own tap/TW ____%
2 One tap shared by No. of hhs _____
3 One TW shared by No. of hhs _____
4 One latrine shared oy No. of hhs _____

19 ACCESS TO UTILITY SERVICES (%)
1 Electricity _____
2 Cooking gas _____

20 LOCATION OF WATER SOURCE
1 Within slum
2 Outside the slum
3 Distance _____ (m)

21 DRAINAGE
1 Well drained
2 Moderately drained
3 Poorly drained

22 FLOODABILITY
1 Fully flooded
2 Partially flooded
3 Flood free

23 GARBAGE DISPOSAL
1 Pixed Place
2 No Fixed Place

24COLLECTION
1 Regular
2 Irregular
2 None

25 SCHOOLING
1 Percent of school going children _____

26 MAJOR OCCUPATION
1 Business/trade ____%
2 Hawker/Vendor ____%
3 Service ____%
4 Day Labour ____%
5 Domestic Worker ____%
6 Transport ____%
7 Factory worker ____%
8 Others ____%

27 MONTHLY HH INCOME
1 < 2000 Tk ____%
2 2001-3000 Tk ____%
3 3001-4000 Tk ____%
4 4001-5000 Tk ____%
5 Above 5000 Tk ____%

28 DWELLER'S ORIGIN
(4 Major District of Origin)
1 _____ (%)
2 _____ (%)
3 _____ (%)
4 _____ (%)

29 FIRE INCIDENCE EVER
1 None
2 Once
3 Twice/More

30 EVICTION INCIDENCE
1 None
2 Once
3 Twice / More
4 Currently under eviction threat

31 NGO COVERAGE
1 One NGO
2 More than one
3 None

32 GO COVERAGE
1 Yes 2 No

33 HOW IS THE SETTLEMENT KNOWN
1 Residential Area / Abashik Elaka
2 Mohalla
3 Colony/staff quarter
4 Camp
5 Bastee
6 Others _____

Thank You Very Much For Giving Us Time

The basic strategy behind ground truthing was then to use trained teams to scour a given area of a study city looking for urban concentrations that appeared to be suitable candidates to be a slum. On finding such a concentration, the team would enter that urban concentration and attempt to identify key informants who might provide information regarding the characteristics of the slum to be collected. These key informants were typically community leaders and organizers, personnel of NGOs and programs operating in the suspected slum cluster, landlords, etc. Through discussions with the key informants, the field team then decided whether the cluster was a slum and, if so, mapped the slum cluster on paper.

With this framework established, the slum mapping proceeded in three stages. These three stages refer to successive phases of work that had to occur with respect to any given area of a study city, but at any given moment in the slum census and mapping process all three phases were being conducted as work began in new areas and continued or was completed in others. These three phases were:

1. **The Preparation of Base Maps** — Official scale maps of the study cities providing information about road networks, physical infrastructure and key features were obtained and georeferenced. Basically, this involved scanning the maps into image files which were then loaded into GIS software and rendered in terms of geographical coordinates.[4] Satellite images of the study cities were then obtained. These were used to first correct any inaccuracies to the official maps (such inaccuracies were usually related to urban development that had occurred since the maps were made). The images were then carefully analyzed, both visually and by algorithm, to identify possible slum clusters, usually based on degree of structural concentration and roof typology. The result of this phase was highly accurate base maps for the census and mapping exercise that could then serve as an organizing framework for fieldwork while also providing information about the location of many potential slum clusters.
2. **Ground Truthing** —With base maps complete, the study cities were divided into operational areas for ground truthing. Teams were trained at the outset of the study (including many practice field visits) and then sent into the operational areas of each city. They visited all suspected slum clusters from phase one and administered the instrument to ascertain whether they were indeed slums. The teams also systematically scoured their operational area looking for any slum clusters that had been missed and likewise administered the instrument to any that were. There were many cases where suspected slum settlements from phase one turned out to be something else, and visual examination of the satellite images had missed many, particularly smaller, slums;
3. **Final map and database preparation** — The completed questionnaires and slum cluster maps from field operations were returned to CUS headquarters. There, the field maps were used to create polygons in the digital map of the relevant study city that were designed to capture the borders of the slum cluster. Via a unique identification number assigned to each slum (the design and maintenance of the identification number system was a small but absolutely crucial aspect of the slum census and mapping exercise), the polygon on the map was linked to a record in a database containing all of the information about the slum collected by field teams as they administered the instrument during their visit.

The result was a highly detailed digitized, georeferenced map of each city that showed the location of each slum in that city and a database that captured the characteristics of each slum.

[4] To georeference them, highly accurate readings were taken at points of reference throughout the study cities with GPS devices. The coordinates so obtained then provided reference points by which the GIS software could then assign coordinates to each point on the images of the scanned maps.

Figure 34. The Slum Map of Dhaka

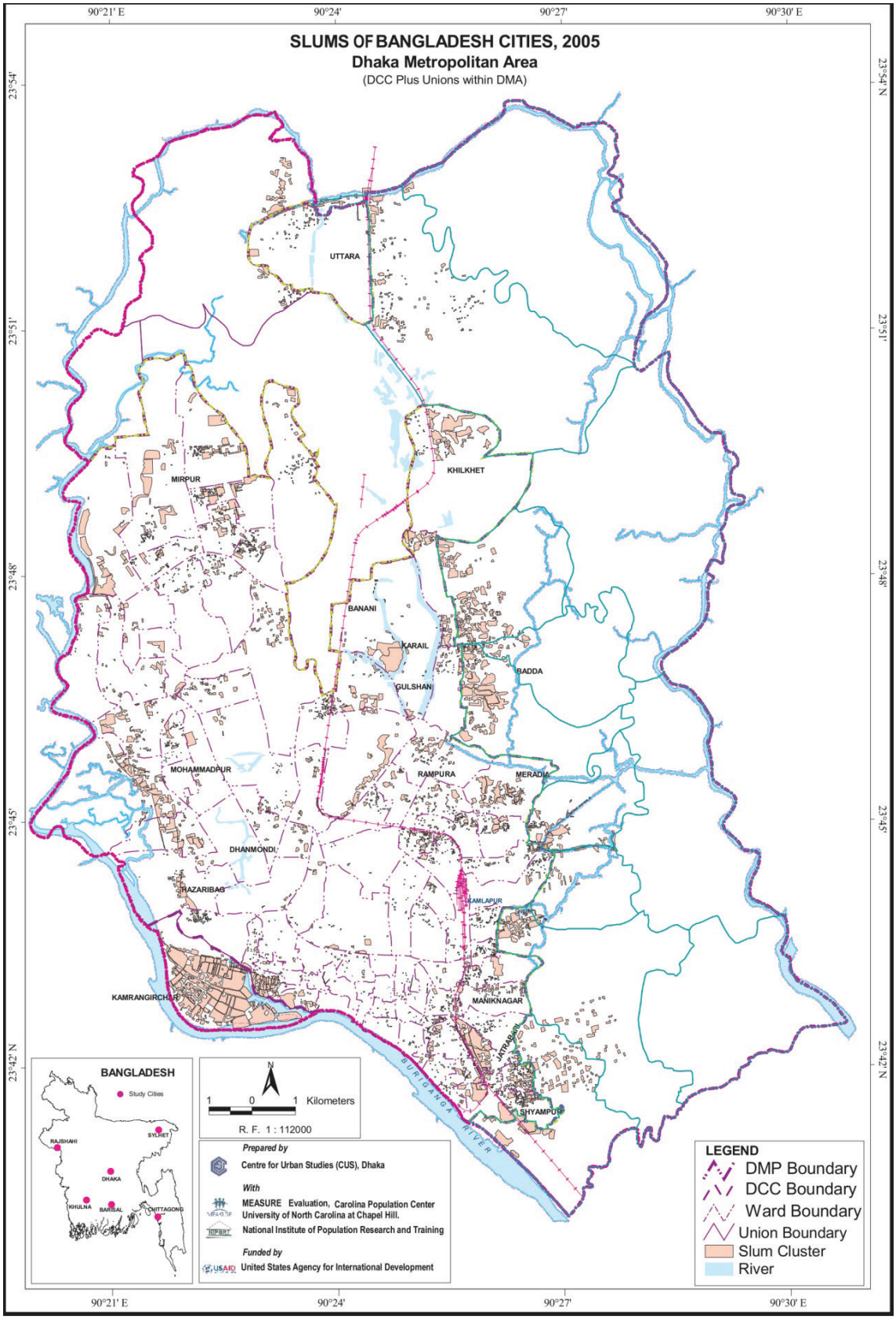
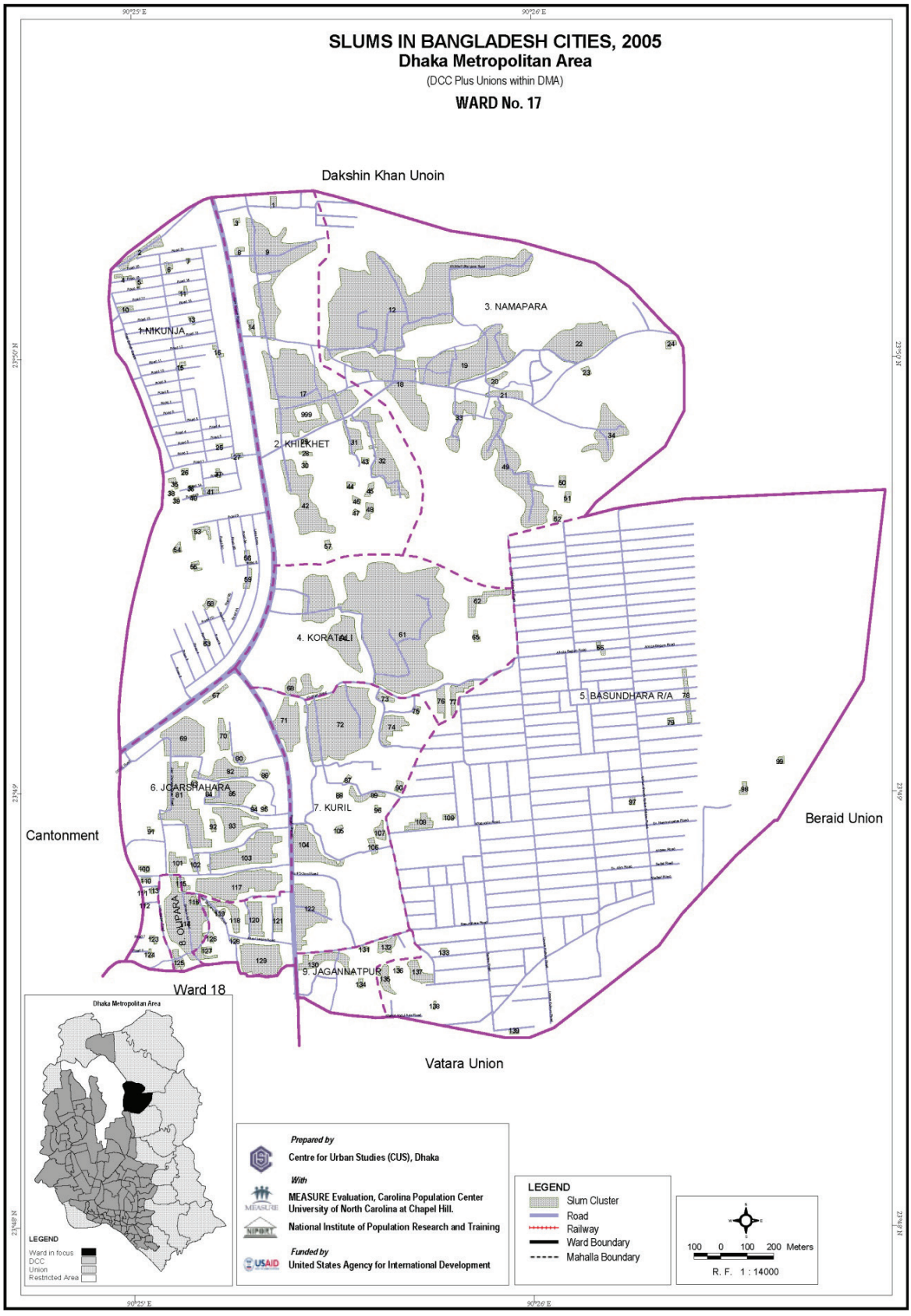


Figure 35. Dhaka, Ward 17



The preceding two figures illustrate the slum maps generated. The first provides the overall slum map for Dhaka. Plainly, slums were not evenly distributed across the physical space of Dhaka. A dense band of slums stretched from central to southern Dhaka, and there was a particularly intense pocket of slums in Southwest Dhaka. This uneven spatial distribution also emerged over smaller subspaces of Dhaka. For instance, the second of these figures shows the slum map for one particular ward of Dhaka (ward 17). The slums of ward 17 were concentrated in a band stretching from the southwest to northeast corners of the ward, with far fewer slums in the southeastern and northwestern corners of the ward.

One important result obvious even before the conclusion of the slum census and mapping exercise is that CUS's suspicion was in fact correct: private slums operated on a rental basis had become far more important. The slum census and mapping revealed that by 2005, 87 percent of the slum clusters, containing 67 percent of the slum populations of the study cities, were cited on private land (the figures for Dhaka were 90 and 70 percent, respectively). This represented a dramatic change from, for instance, the 1996 census of slums in Dhaka.

The goal of the mapping was to locate and list all slums in the study cities in order that a sample of them could be selected from that list. The maps of slums and linked database of their characteristics provided all that was required to form primary sampling units to support multi-stage sampling from the slum populations in the study cities. The idea behind this was that, to the greatest extent possible, each slum would serve as a slum primary sampling unit. The only exceptions were slums too small (by population) to meet the within-cluster household sampling targets for the Urban Health Survey and slums so large (again, by population) that, in the event that they were selected, listing the households within them would be an impractical prospect. In the former case small slums were combined, sometimes with other small slums and sometimes with a modest size slum, to form a new, combined sampling unit (with combinations driven by geographic proximity above all). The largest slums were usually divided into several smaller sampling units of roughly equal size, with division boundaries based on ground features apparent from the satellite photographs that would, in the event that one was selected, make the task of finding their boundaries easy for Urban Health Survey field teams.

The planning and preparation for the slum mapping had been a major exercise that stretched over many months. Having completed the preliminaries and launched the initial map preparation and fieldwork for the slum mapping, a huge oversight became apparent: very little attention had been paid to the sampling of non-slum populations. The question of what should serve as the primary sampling unit for the non-slum populations proved surprisingly hard to answer.

The eventual solution was to base the non-slum primary sampling units on the lowest level administrative unit in urban areas, called the mohalla. Specifically, the non-slum area of each mohalla was to serve as the primary sampling unit for non-slum populations. To operationalize this, mohallas were mapped as part of the slum census and mapping fieldwork through consultation with local ward and mohalla officials. The non-slum area of each mohalla was mapped by simply laying the slum map on top of the mohalla map in the GIS. Population estimates for each mohalla were based on census figures. The population of the non-slum area of each mohalla was estimated as the census total population figure for that mohalla minus the total population of the slums within that mohalla (where a slum straddled multiple mohallas, its population was apportioned between them for this purpose according to the proportion of its surface area in each mohalla). Where the resulting non-slum area mohalla proved either too large (or small) in terms of population a remedy similar to that applied to large and small slum clusters was employed.

All of this work to develop final lists of slum and non-slum primary sampling units was relatively easy because it was done within the context of a well-crafted GIS. Nearly always the tasks involved were automated via algorithms that, for instance, implicitly layed the slum and mohalla maps on top of each other and referenced the databases of slum characteristics and mohalla populations to identify the slum and non-slum areas of each mohalla, estimate the total slum population within that mohalla, produce an estimate of the non-slum population of the mohalla and, finally, deal with small and large slum and non-slum primary sampling units as indicated (to be sure, the last task required a great deal of quality checking).

With work complete, sampling progressed and the 2006 Urban Health Survey proceeded. This was the original motivation for conducting the census and mapping of slums, and the successful stratified sampling from slum and non-slum populations justified the enormous effort. That said, in the months and years after work was completed, the census and mapping of slums found wide application as a tool to guide program placement in Bangladesh. A few examples:

1. The Bangladesh Rural Advancement Committee (BRAC) used the maps to place birthing huts in slums and to target expansion of their health program for the urban poor;
2. The Bangladesh AIDS program used the information to plan the future location of counseling and treatment centers for most-at-risk populations (e.g., sex workers, drug users, truck drivers, rickshaw pullers, etc.), who disproportionately reside in slums;
3. The United States Agency for International Development (USAID) funded Non-Governmental Organization (NGO) Service Delivery Program (administered by Pathfinder) used the maps to place main and satellite clinics for its next five-year phase;
4. The United Nations Development Programme (UNDP) funded Local Partnership for Urban Poverty Alleviation requested the maps for use in targeting their efforts in their next seven-year phase;
5. The World Bank, which was working with the water and sewer authorities of the City Corporations to extend water and sanitation to the urban poor, requested the maps so that they could identify those slums not serviced by the present grid;
6. Family Health International (FHI)/Bangladesh used the maps to track their own intervention sites and areas of program coverage for various programs.

Such leveraging of the census and mapping outputs made it a far more cost-efficient exercise from a social welfare perspective.

There are a number of important lessons from this exercise,[5] and a variety of ways that the design and execution of the census and mapping was informed by the principles of practice outlined in the preceding chapter. We conclude by summarizing how the principles were put into action.

- **Know When GIS Can Inform Sampling:** Slums are urban concentrations that are typically unevenly distributed across cities. Moreover, they are typically strongly associated with circumstances, both in terms of visual signatures apparent on spatial imagery and readily obtainable indicators (in this case on field visit, but in some societies such information might exist down to the census EA or block). This situation thus presented a perfect circumstance for GIS to inform sampling—the subpopulation of interest had a highly uneven geographic distribution, the locations of which could in principle be accurately predicted through a GIS.
- **Apply a Scientific Standard:** Once again, the scientific concept behind the census and mapping hinged above all on replicability—another research team with the same resources and slum definition, and operating at the same time, would likely have been able to recover a list of slums and locations for them with a high degree of concurrence with those actually produced. A key to this was the transparent protocols for basemap preparation, as well as a carefully designed field instrument and extensive training of field teams to execute that instrument (since the fieldwork could not be automated, it was the portion of the census and mapping exercise most prone to human error).
- **Automate, Automate, Automate:** Wherever possible, the power of the GIS as a tool hosted by computers was harnessed. For instance, the tasks of locating potential slum clusters through visual signature patterns on satellite images, the determination of non-slum areas for each mohalla, estimation of total slum, and therefore non-slum, populations, for each mohalla, and the building of final primary sampling units were all automated per algorithms with the results checked by humans. This yielded tremendous labor and time

[5] In case the reader has not already guessed, the authors were involved in the census and mapping of slums exercise.

savings (indeed, the planned start date for Urban Health Survey fieldwork would not have been possible without it) and likely reduced the error rate tremendously. Not all of the GIS work could be automated (e.g., slum polygons themselves were generally inputted by hand), but the automation that was implemented made the census and mapping a far more manageable exercise with higher quality outputs. That said, human checks on automated work was still essential—no matter how carefully the algorithms were designed, they still occasionally yielded strange results compared with what a human possessed of judgment would have done.

- **Be (Reasonably) Uncompromising:** The slum census and mapping was marked by extremely detailed planning, precise and transparent definition of instruments, careful field tests of those instruments, intense training of field workers, overlapping quality control mechanisms, etc. That said, some initial ambitions had to be abandoned. For instance, the original notion was to map precisely the boundaries of slums with GPS devices. This proved unworkable, as even the most precise available devices required multiple readings for each point (and the organized processing of all of these coordinates was clearly going to be a challenge), more complicated slum borders required excessive readings, the devices were clearly going to be a pain to maintain in the field though such a massive exercise, etc. In the end, after trial and error, it was determined that the quality gained from the devices (as opposed to clear hand-drawn maps) was so modest that it did not justify all of the logistical risks associated with their use. To cite another example, slum population numbers were approximate—a precise count would have been prohibitively time consuming for little gain in sampling efficiency or effective program targeting. If the study team had tried to do everything "perfectly" (whatever that means) the census and mapping would still not be complete even now and the quality of the final product would ironically have worsened as headquarters drowned in a deluge of information (e.g., millions of GPS coordinate readings) and field teams were overwhelmed by labor-intensive, tedious tasks. Don't let the perfect become the enemy of the good.
- **Be Careful Designing Sampling Units … and Be Ready for Problems Even Then:** A great deal of thought and effort was put to the task of preparing final primary sampling units. Even then, however, problems were encountered conducting the Urban Health Survey in the field, particularly with overly large selected PSUs. This was usually due to the census population number for the mohalla being too small, a problem more common to the mohallas in the more dynamic, particularly peripheral areas of cities experiencing rapid population growth. In such cases, a segmentation scheme (whereby the primary sampling unit was divided into segments roughly equal in population size and one was randomly selected) was already in place.
- **Consider Ground Truthing and Testing:** This study was massive in scope and as such in some sense can be viewed as a sort of proof of concept exercise—it established that large-scale fieldwork could be conducted as part of the building of the GIS. The fieldwork for the census and mapping uncovered 9,048 slums and involved visits to a couple thousand other urban formations that turned out for one reason or another not to be slums. What this establishes is that large-scale fieldwork is possible to support a GIS and within reasonable means—the census and mapping of slums still cost far less than the actual Urban Health Survey itself. Even if there is no prior, linkable (i.e. georeferenced) information, one can still build a GIS in many circumstances by gathering the information themselves. What lay behind the success of the census and mapping of slums was careful planning, robust organization and pre-testing of procedures—virtually every instrument, field search technique, community engagement method, etc. was tested with frequent experimental visits to the field before the start of actual fieldwork.
- **Be Careful About "Betting the Farm":** The generally received wisdom on the eve of the slum mapping was that slums in urban Bangladesh were overwhelmingly squatter settlements illegally sited on land usually owned by the government or some sort of parastatal. On the eve of the 2005 census and mapping, there was thus a strong argument for simply sweeping publicly owned lands to look for squatter settlements. This would have involved effectively betting that the received wisdom was correct. Of course, the 2005 census and mapping involved a more general search approach robust to a violation of this conventional wisdom, and what was found was that squatter settlements had become somewhat less important (since, for instance, the 1996 Dhaka census and mapping) and that most of the tremendous growth in slum populations in Dhaka from 1996-2005 had been absorbed by private, largely rental slums. When building a GIS, it is

important to understand the implicit and explicit assumptions behind a proposed design for the GIS, and to recognize the consequences of violations of these assumptions. By doing so one can minimize reliance on assumptions that could serve to undermine the value of the GIS if they prove unreliable.

- **Do Not Let Your Reach Exceed Your Grasp:** The 2005 census and mapping of slums was seen at the time as a path-breaking application of GIS technologies to the subject at hand. Indeed, the GIS design and reliance on detailed satellite photographs was unprecedented for a slum mapping in Bangladesh (or, to our knowledge, any other society as of that time). At the same time, not every technological possibility was brought to bear. For instance, as discussed previously, the decision was made to follow a decidedly low-tech approach to the mapping of slum boundaries—to do otherwise would have introduced a complexity to communications between the field and headquarters, as well as a recording burden in the field, that was felt to introduce risk for only slight reward.
- **Time Might Not Be On Your Side:** The census and mapping of slums took roughly a year from initial discussions to final delivery of primary sampling units for the Urban Health Survey. The research team knew from the outset that it would not be an overnight process, but technically correct execution to generate high quality outputs ended up taking far more time than was originally expected. This is probably more the norm that the exception in GIS work, particularly in cases where one needs to design and execute their own protocols for gathering information not readily mergeable in to the GIS, especially through fieldwork. One should not enter this process thinking that it will be quick and easy, as it will require effort to produce quality outputs and probably take longer than expected.
- **Harness the Power of Leveragability:** The census and mapping of slums was a massive effort. The research team was in the fortunate position of being able to carry it out with the funding already programmed in for the Urban Health Survey. That said, it was still an enormous, time-consuming and expensive undertaking. Finding ways to leverage it in terms of making it widely useful beyond the original motivation for conducting it was a major achievement. Where the marginal cost of doing so was manageable, the research team augmented the protocol (e.g., by adding instruments of likely particular value to human welfare programs). The result was a tool of use not only to the Urban Health Survey but also the allocation of numerous slum-oriented programs at the time.
- **Be Mindful of Local Sensitivities and Laws:** The census and mapping of slums could not have been completed without the help of thousands of program officers, local community leaders, government officials, and residents of the slums themselves. They made the survey possible, and they are the source of information on the indicators, mohalla boundaries, local socioeconomic setting, etc. that made high quality slum maps and lists possible. Developing relationships with these individuals was perhaps the most important ingredient to the success of the census and mapping. The census and mapping was also conducted in a fashion sensitive to local laws and considerations. For instance, slums are not found in enclosed military installations, and hence they were not scrutinized for this study.

While the reader might have a very different application of GIS to sampling in mind, the hope is that this discussion of a particular application of GIS to sampling will help to render the principles of practice suggested in the last paragraph more concrete. Nonetheless, careful thought would be required to understand how they manifest themselves in other circumstances.

Source: Wikimedia Commons, CC-PD-Mark

Chapter 6. Conclusion

One rarely recalls all of the details of any technical document that they read. Such a goal is essentially impossible in the age of information overload—there are just too many often comparatively arcane technical details in a manual of this sort competing for too little space in the human memory. It is perhaps often more useful to think instead of the core messages that one hopes would continue to resonate from a resource such as this. This manual has essentially two such messages.

First, it is important to recognize the circumstances in which a GIS can inform sampling. At the most essential level, it is necessary that the object of sampling be a subpopulation with a distinctive spatial pattern of settlement predictable through a GIS. This allows one to stratify the geographic space within which a survey is to be conducted into areas where that subpopulation is more and less concentrated, allowing for more focused sampling of the subpopulation by concentration of sampling efforts in those areas where that subpopulation is thicker on the ground. Whatever the subpopulation of interest or information goal of the survey, it is hard to see how a GIS can inform sampling if there is not a predictable pattern to the spatial distribution of the subpopulation of interest.

Second, regardless of the subpopulation of interest, information goals of the survey, or resources available (immediately or potentially with time and effort), there are some reasonable principles that should guide the application of GIS to sampling. Aside from knowing when a GIS can inform sampling (discussed above, this is perhaps the most important principle) these principles include:

- **Apply a Scientific Standard:** The foundation of laboratory science is the standard of independent replicability. The key to this is typically carefully crafted, complete and transparent protocols. Even when it seems unlikely that anyone will ever attempt to replicate your GIS, it is still useful to think in these terms when designing and building it. This will help to insure the highest level of quality and coherence in the resulting GIS database, and hence the most efficient sampling possible. It is important to remember as well that, since the ultimate purpose of applying a GIS to sampling is to increase sampling efficiency, you are really cheating yourself when you cut corners scientifically when building that GIS.
- **Automate, Automate, Automate:** Humans are not very good at repetitive tasks, which are often the lion's share of the work in building a GIS. For instance, the authors' experiences at the actual work of using GIS for sampling have more often than not been dominated by tedious tasks such as identifying specific material patterns to structures (especially their roofs), delineating polygons, correcting discrepancies between GIS information sources (such as official maps and satellite images), etc. Repetitive execution of such tasks inevitably leads to mistakes as the mind wanders. Wherever possible, such work should be automated, through the application of carefully constructed algorithms. At the same time, it seems unlikely that an algorithm, no matter how well designed, can foresee every circumstance that might arise in the real world. It is therefore important that the application of algorithms be accompanied by careful checks by humans possessed of a capacity for judgment that no rigid rule can offer.
- **Be (Reasonably) Uncompromising:** Building a GIS is never a perfect exercise. For one thing, it is typically not possible to obtain GIS information sources that reflect perfectly actual circumstances on the ground. Even satellite imagery, however recent, reflects past circumstances. There is thus a limit to how accurate a GIS can ever be. This suggests that one should perhaps be a bit shy about the pursuit of absolute perfection, particularly in instances where that pursuit can come at great cost in terms of time, resources, flexibility, etc. At the same time, this kind of philosophical approach can quickly become a kind of slippery

slope to shoddy quality. One should compromise as rarely as is feasible—the practitioner would do well to remind themselves once again that in the end it is only themselves that they cheat when they degrade the quality of their GIS.

- **Be Careful Designing Sampling Units … and Be Ready for Problems Even Then:** The real world rarely fully cooperates with even the best laid plans. While it is essential that one develops sampling units as carefully as possible through their GIS (again, when they fail to do so it is primarily their own survey research lives that they are complicating), it is unlikely that ground conditions will fully cooperate with the frame construction process. For instance, sampling units might fall outside of the manageable population size range. Practitioners should assume that problems along these lines will arise, and be prepared to deal with them.
- **Consider Ground Truthing and Testing:** Ground truthing and testing is essentially about bridging the gap between the picture of ground conditions offered by the information readily available for building the GIS (i.e., the information available outside of ground truthing) and actual conditions on the ground. Ground truthing can provide an information source in its own right and offer a test of the accuracy of the other information sources behind a GIS. In any event, it is always a good idea to get one's head out of the sand (or, in this case, GIS software) and actually look at the world around them.
- **Be Careful About "Betting the Farm":** In life we constantly make choices. Usually, there are assumptions driving those choices. Rarely, however, do we think explicitly about those assumptions. In building a GIS database, it is important to think carefully about the assumptions, explicit and implicit, driving the predictions about the spatial distribution of the subpopulation of interest that that GIS will provide. By doing so, one clarifies how the predictions depend on the assumptions and, hence, how a poor assumption can lead prediction awry. Avoid reliance on assumptions that are risky or somehow not necessarily very convincing or credible—they may undermine the end product of the GIS-building exercise.
- **Do Not Let Your Reach Exceed Your Grasp:** The capabilities of GIS software are constantly expanding. To a significant extent, this reflects the fact that the technological possibilities for gathering information to inform a GIS are multiplying at a stunning rate (more on this below). The upshot is that at this moment of history crafting a GIS database is something of a dream for an "early adopter" (i.e., one prone to rapidly adopt new technologies). There is, however, some degree of risk in this since new technologies often involve complications or limitations not apparent until the moment of execution. While one should never ignore an emergent possibility, they should also make sure to adopt technology in a fashion designed to minimize the risks of unforeseen complications. If nothing else, any emergent technology to be considered should be thoroughly tested during the design phase for a GIS.
- **Time Might Not Be On Your Side:** Virtually every instance where the authors have been involved in mobilizing a GIS to inform sampling took far longer than was originally envisioned. There is almost always some unforeseen twist or complication in building a GIS database. Sometimes this is a good thing (as when a new and powerful information source becomes available mid-course). Whatever the case, one should always begin as early as possible the process of building a GIS database to inform sampling.
- **Harness the Power of Leveragability:** Building a GIS database to inform sampling can be a massive task. Moreover, slight modifications to a GIS designed for sampling (including simple extensions to the protocol to insure that the final product is more readily available and usable) can have a tremendous impact on the range of purposes to which the information in the GIS can be put. At the most basic level, a GIS that identifies spatial patterns to vulnerable subpopulations for sampling purposes can do the same for program targeting purposes. Look for opportunities to make a GIS as widely useful as possible, helping to justify further what can be the considerable costs (in money, time and effort) of developing one.
- **Be Mindful of Local Sensitivities and Laws:** As with any survey research work, it is important to conduct the work in a fashion consistent with the laws and cultural sensibilities of the society in which work is to be done.

These principles have been introduced in the most general terms. However, it seems reasonable to expect that each of them could in some fashion or another usefully inform virtually any application of sampling to GIS.

There was a purpose to the rather generalist approach behind framing the discussion in terms of these principles. Laying down a set of specific instructions for applying GIS to sampling would have been a fool's errand—it is impossible to predict all of the ways that the tools, techniques, and technologies of GIS might be applied to sampling for population surveys. The simple reason for this is that the potential scope for applying GIS to sampling is a moving target in at least two senses, which can be thought of roughly in terms of the "demand for" and "supply of" applications of GIS to sampling.

To begin with, there is the "demand side," by which we mean all of the instances in which researchers might wish to apply GIS to sampling for a particular survey. One cannot predict all of the sampling challenges that might arise for which GIS might improve the efficiency of sample selection.

The focus in this manual, and in much applied work in the real world, is population survey research to supply information that is somehow programmatically relevant. Programming in the human welfare arena is rapidly evolving. It is growing ever more sophisticated in delivery, conceptualization of participation, and other key programmatic parameters. For instance, a traditional approach to participation in more rural, less technologically sophisticated societies might simply involve placement of clinics in villages thought to have more poor people. Today, programming in rapidly urbanizing societies that are beginning to experience more penetration of information technologies increasingly focuses on encouraging participation in some fashion or another at the individual level. Above all the reason for this is efficiency, in that it allows for the more precise targeting of scarce resources to the subpopulations that a program strives to serve in a given operation area.

Human welfare programs are also striving to serve ever more finely defined subpopulations, often based on complex (but behaviorally persuasive) conceptualizations of vulnerability. In some cases, these are not novel conceptualizations of subpopulations, but contemporary events have placed a greater focus on them. For instance, the global HIV/AIDS epidemic has resulted in an emphasis on programming oriented toward those most at risk of infection (e.g. sex workers) or those most vulnerable in the wake of the epidemic (e.g., those orphaned by the disease). At the same time, there are still many traditional programming challenges for which GIS could usefully inform sampling (e.g., sampling from nomad populations).

What we are driving at with this line of thought is that the number of questions that can be asked in the context of survey research to inform programming is vast, and growing. Appropriate sampling to answer these questions presents countless potential circumstances in which GIS might inform sampling. From the "demand" side perspective, general principles would hence be more useful than specific instructions.

At the same time, the "supply side" for GIS applications to sampling is also growing. By this we mean that the possibilities for applying GIS to sampling are expanding rapidly. In short, with each passing year there are more types of sampling challenges that can usefully be informed by GIS.

The primary driving force behind this is a rapid expansion in information sources available to inform a GIS. To some extent, this reflects the continuous, inexorable trend toward more transparent and digitized official data, as well as the expansion in available private sector data sources. With each passing year far more public and private information is available (the latter for generally falling real prices), often in pre-digitized, geo-referenced form.

To cite an example, one of the authors of this manual was involved in both the 2005 census and mapping of slums in Bangladesh discussed in the preceding chapter and a more recent (2010) slum census and mapping in select cities of Uttar Pradesh, India designed to inform sampling. In the former case, it was necessary to scan (from hard copy paper to a digital image) and georeference (i.e., render the points in the resulting digital image in geospatial coordinate terms) official street maps of the study cities in Bangladesh to provide a framework to the satellite images employed. By the time of the 2010 effort in Uttar Pradesh, a commercial firm offered a fairly polished finished product that integrated such digitized, geopreferenced street maps with administrative boundaries and other official statistical information. It was our understanding that by that time similar products were available for the 2005 study cities in Bangladesh.

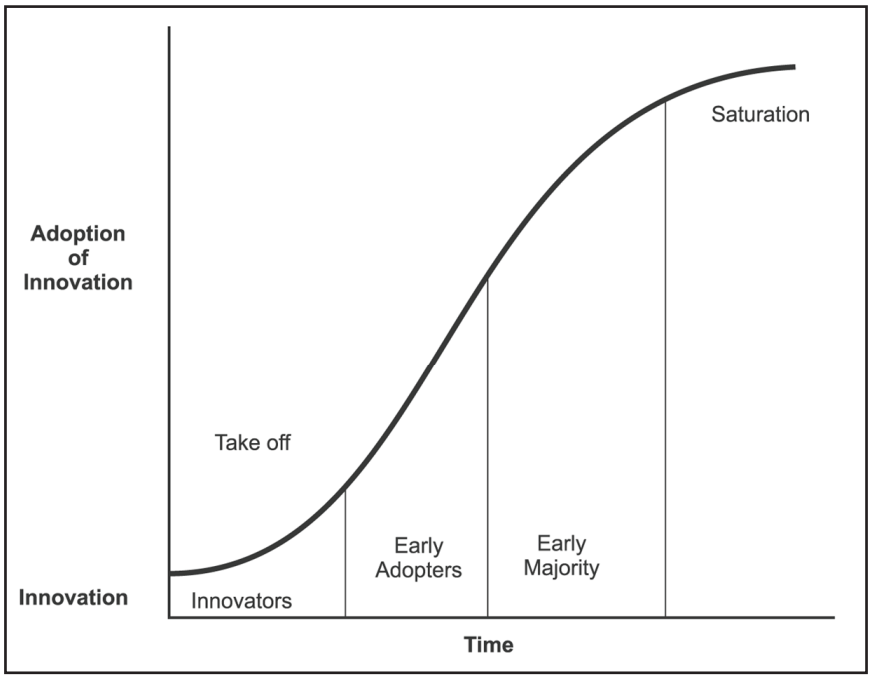
Aside from this general trend toward the digitization and georeferencing of traditional data sources for a GIS, there has also been an explosion in the technologies available for gathering data to inform a GIS. Essentially, there are far more kinds of sensors becoming available, and they are becoming far cheaper and more ubiquitous.

In some sense, we are actually well into one revolution in terms of a technology for collecting information to inform a GIS—the widespread emergence of satellites. The first launch of a commercial satellite was the Telstar I satellite in 1962. It was a communication satellite, and most of the satellites (public and private) launched in the decade or so to follow were more or less geared toward telecommunications objectives. For instance, comparatively risky and expensive spy plane flights were a mainstay of aerial intelligence gathering well into the 1970s owing to a paucity of satellites gathering aerial imagery and the limited quality of that imagery.

Beginning perhaps in the late 1980s, satellite imagery became increasingly available, from a wider range of public and private satellites, and with increasing precision and clarity. This presented a framework for GIS work not really available until that time. As time passed, the number and sophistication of satellites offering ground imagery grew, and economies of scale and competition reduced the cost (say, per square meter) of providing high resolution images. The integration of satellite imagery into GIS, at first very limited, is now common place. In some sense the process followed what is often referred to as an innovation curve.

Diffusion of technologies often follows an innovation curve. According to this theory once a new technology is created it is used by a small number of people known as early adopters. Early adopters often pay more for technology and face challenges and bugs as the innovation matures, but they provide the opportunity for manufacturers to improve functionality and production methods. As technology matures it becomes cheaper and as a result more accessible until it is widely used.

Another example of a technology relevant to GIS now riding an innovation curve is mobile phone technology, which is beginning to become a potentially powerful tool for GIS. Mobile phones are currently generating many innovative information sources that might inform a GIS, such as the Google traffic maps discussed earlier. Mobile phones were large and expensive devices with limited functionality when initially introduced in the 1980's. Over time technology improvements reduced the cost of the devices and as the costs went down, more people purchased them. The increasing demand led to economies of scale that decreased costs further. The cycle repeated until the devices were in wide use around the world.



The basic dynamic of the innovation curve is presented in the previous figure, which illustrates the process. Essentially, the initial phase is dominated by innovators who make the new technology work. Early adopters then use the technology, in the process serving as experimental subjects for the purpose of working out remaining kinks in the technology. Traditionally, they often pay dearly for the new technology as well, since factors such as competition and economies of scale have not been brought fully to bear to lower costs. At some point the technology is adopted on a widespread scale on the road to a point where it has achieved saturation levels of adoption and is a common feature of everyday life.

At this point in history, there are many technologies somewhere along the earlier track of this curve that will likely soon revolutionize GIS by opening up many new information sources. The emergent possibilities from these technologies often represent a confluence of developments in materials science, micro-computing and ever cheaper and more precise and sensitive monitoring technologies.

An excellent and exciting example of this is the emergence of unmanned aerial vehicles (UAVs). UAVs have been around for some time, but were typically used only on an experimental basis or put toward some sort of specialized mission for which their high costs and limited capabilities were acceptable. Then, in the past two decades, they became integral to military operations of wealthier nations (particularly the United States) and found more widespread use for particular purposes such as weather monitoring.

Now UAVs are increasingly available to civilians. Indeed, many consumer-oriented UAVs are available for easy purchase on websites such as Amazon.com. The image below illustrates a typical moderately priced civilian (i.e. consumer) UAV at this point in time. These kinds of devices were made possible by several factors. First, many of the civilian UAVs for sale today are most like the small hand-held military drones that became increasingly common in the past decade or so. The military can be thought of as early adopters of this technology, and their experiences allowed for many performance kinks (e.g., avionics and stability issues) with these early drones to be resolved.

Source: Tyler Olson/Shutterstock

In some sense, these consumer UAVs are also natural heirs to the remote control aircraft that hobbyists have flown for decades. However, in their capabilities, the new UAVs offer far great potential. In short, they can do far more than the old remote controlled aircraft of the hobbyist world, and the range of possibilities that they afford are set to explode.

These consumer UAVs have, like their military cousins, benefitted from a quiet revolution in materials science that has steadily increased their payload, flexibility and range. An extremely important example of this is increasingly widespread adoption of carbon fiber, which combines immense strength with low weight. A likely next area for huge technological improvement is battery science. This could increase tremendously the stored energy per unit of weight of UAVs, thus extending their range, operational time and capacity.

In a development that could have implications well beyond UAVs, many monitoring and recording technologies are becoming increasingly sensitive and precise, even as their physical profiles grow smaller and lighter. A tremendous impetus for the current consumer demand for UAVs has been the arrival of lightweight, rugged (remember, UAVs sometimes crash, just like all aircraft) high resolution digital cameras that can be easily mounted onto a UAV. An example of this is the lightweight, inexpensive, tough, and currently wildly popular GoPro cameras, an example of which is illustrated in the next image.

Photo by Wayne Hoover

The truth, however, is that even this information capture technology is at its dawn. These cameras will only grow smaller, lighter and more capable. Different types of imagery are likely to become more ubiquitous. For instance, an increasing number of consumer digital cameras offer some sort of night vision capabilities, and it seems likely that the most sophisticated current infrared imagery (i.e., imagery that essentially captures heat signatures) will be available in digital cameras in a short while.

The increasing range and sophistication of images available could open up powerful new possibilities for GIS. For instance, in a not too distant future, population size measures might be based on night time (i.e., while most are sleeping) recordings of infrared imagery, the digital analysis of which will allow for cheap and far more current (and thus accurate) estimates of the number of individuals living in a sampling unit.

Moreover, the information capture possibilities are unlikely to be long confined to imagery. Similar improvements are occurring to air quality measurement technology, sound detection, mobile phone monitoring, etc. Within a short while the combination of these technologies with UAVs could make the measurement of virtually any feature of a physical space cheap and trivial.

If this description of UAVs has seemed somewhat rambling and unstructured, it is because so much innovation is occurring, and along so many fronts, that it is difficult to present a really organized, deliberate summary of current and emerging developments. A similar narrative could have been applied to any number of other lines of technological developments.

Whatever possibilities GIS offered for sampling in the past, it seems safe to say that the "supply side" will offer many more in the near future. What can be captured, and at reasonable cost, continues to expand dramatically. This means that the range of things that can be characterized, and hence predicted, through a GIS continues to expand just as dramatically.

MEASURE Evaluation
Carolina Population Center
The University of North Carolina at Chapel Hill, CB 3446
Chapel Hill, NC 27516 USA
www.cpc.unc.edu/measure