Sampling Guide for Displacement Situations & Practical Examples

June, 2020

WE ARE HERE!

Initiating the profiling exercise

Establishing the collaborative process

Designing the methodology

Implementing data collection

Processing and analysing the data

Validation, reporting and dissemination
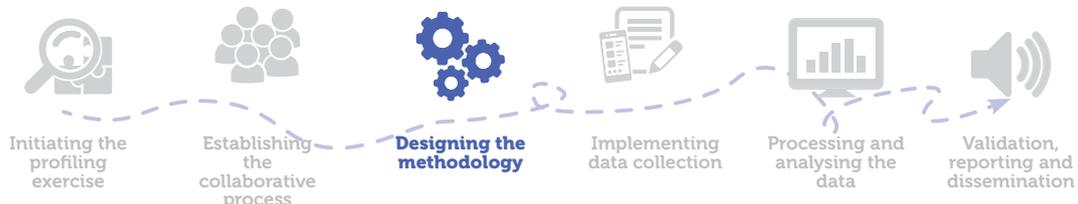
JET Sampling Guide Part I&II

Introduction

With the *JIPS Essential Toolkit* (JET), we embarked on a journey to create and disseminate accessible tools and methodology guides that empower our colleagues and partners around the world to create a shared understanding around displacement situations. By producing an agreed-upon evidence-base for designing programs and policies, a profiling exercise can support the achievement of durable solutions for displaced populations. In an effort to share our collective experiences in this area and make sure that no one starts from scratch, we broke down the profiling process into six phases for a step-by-step and accessible approach. As such, the JET is a collection of templates, guidance documents and checklists that are designed to guide people in carrying out profiling exercises.

As part of phase 3 which walks practitioners through the methodology design of a profiling exercise, the Sampling Guide for Displacement Situations represents a critical element in this process. Indeed, the objective behind designing a methodology is to ensure that the approach taken, and the analysis that results, will actually address the contexts' data needs and that the results of analysis are consistent with the objectives of the profiling exercise. With that in mind, a properly designed sampling methodology ensures that the characteristics of the sample are representative or as closely representative as possible of the larger population, using a comparative approach. Therefore, choosing and planning for the right sampling approach is the cornerstone to collecting robust and reliable data.

That being said, the Sampling Guide for Displacement Situations aims to only briefly introduce key concepts on sampling (e.g. target population, sample frame or bias) and focuses on being a reference for practitioners in their process of selecting a sample in complex situations such as those of protracted displacement. This guide seeks to be as practicable as possible by providing concrete examples of sampling approaches directly taken from JIPS' experience in supporting profiling exercises over the years and by including helpful and go-to resources.

Part I of the guide focuses on sampling as a methodology by answering key questions with regards to sampling selection, terminology, available options and documentation. It provides basic definitions and facts around sampling, before diving into more detailed technical aspects of the process. It breaks it down step-by-step and reviews all sampling options available to a researcher with a detailed description of how it is to be conducted, how to decide which one to choose and what errors one can expect to come across. Part I is therefore a go-to guidance document to understand the theoretical and technical aspects of sampling. It does not intend to cover all possible aspects of sampling in depth – as there is a wealth of literature on that, rather it should serve as a refresher and an introduction to key concepts.

Part II comes as a very practical complement to Part I as it presents a selection of profiling exercises supported by JIPS. Here, practitioners will find a detailed description of the approaches selected, why they were chosen and what their limitations were in six different displacement contexts (El Salvador, Greece, Iraq, Kosovo, Somalia, and Sudan). Researchers interested in carrying out profiling exercises, can thus have a glimpse of the operational reality and are invited to formulate assumptions for similar contexts accordingly. By sharing our experience in the field, we hope to encourage and inspire a community of practice to lead the way and set high standards on responsible data processes.

J
E | JIPS
T | Essential
Toolkit

⊗ Who is JIPS?

This guide was developed by the team at the Joint IDP Profiling Service (JIPS). JIPS is an inter-agency service established in 2009 to help governments and humanitarian and development organisations design and implement profiling exercises of displacement situations. Our primary mission is to provide support, either on-site or remotely, in the form of technical assistance, capacity building and training, tools and guidance.

Acknowledgements

Table of contents

Part I

Sampling Guide for Displacement Situations

JET | JIPS
Essential
Toolkit

Introduction

When we want to collect information about a group of people, we have the choice between collecting information on all members of the group, or only from a sample of them. The first alternative is called a census. In theory, a census will give the most accurate data. In displacement-related contexts, however, conducting a census can often be particularly vulnerable to bias, due to a number of reasons related to the specific context, as well as the size and operational complexity of the exercise. Conducting a census is generally also an expensive task and often requires years of planning and preparation.

The second option is called a sample-based survey. With a survey, we select a sample (a subset of the population) that represents the entire group of interest to the study as accurately as possible. With the right design, a sample-based survey can give us information of sufficient accuracy to produce reliable results. It can, in many displacement contexts, be less vulnerable to certain biases[1], and is most of the time more cost-efficient than a census. Thus, conducting sample surveys is often the preferred way of collecting information in displacement situations or situations with specific thematic focus.

The aim of this guide is to introduce key concepts on sampling and serve as a practical reference in the process of selecting a sample that can represent the population of interest in a displacement context and, where suitable, provide examples of sampling approaches taken in JIPS-supported profiling exercises[2]. This guide is not looking to provide an exhaustive review of the theory behind the statistical concepts of sample selection in general, but will point to helpful resources for this where possible.

> ## (ℹ) Profiling
>
> A collaborative exercise that seeks to **establish a shared understanding of displacement situations** and the circumstances and characteristics of those affected. It uses **mixed-method approaches, which often entail sample surveys** to collect and analyse data on displaced populations, their host communities and others, and situates this in broader considerations of the economic, political and social backdrop of displacement. The overall aim is to create a comprehensive and mutually **agreed evidence base to inform more effective humanitarian and development interventions, advocacy efforts and the development of national policies** to support the achievement of durable solutions for displaced populations.

We start by looking at why choosing the right sampling approach for your sample survey is essential for obtaining reliable results. Then we will go through some essential terminology that will be useful to keep in mind when reading the rest of the guide, as well as when communicating with potential partners or colleagues about sampling methodology. Finally, we will dive into the process of selecting a sampling approach.

---

1   Such as inclusion and exclusion biases, which will be explained later in the document.

2   For more practical examples of sampling approaches in JIPS-supported profiling exercises, see Part II: Examples of sampling approaches in displacement profiling exercises.

Why is choosing the right sampling methodology essential?

Choosing the right approach for how to select the sample for a sample survey is a critical part of the process of designing the overall methodology in a profiling exercise. A properly designed sampling approach enables you to assume, with a given certainty, that the characteristics of the sample hold for the larger population it was selected from (to generalise) and is therefore an essential prerequisite for a sound and rigorous analysis.

On the other hand, a poorly designed sampling approach can severely limit the collected data's usefulness in answering the research questions. In short, the effort invested in choosing and planning an adequate sampling approach will be rewarded with reduced risk of errors and reduced risk of potential costs later on in the profiling exercise.

What is the most important terminology?

Target population

This is the population group that you want the results from the sample survey to represent.

Example

"Refugees and asylum seekers who arrived in Thessaloniki after January 2015"

Sample frame

An actual list of the individuals or households in the target population, from which you will select the sample. This could for instance be a list of addresses. In an ideal situation we would have access to a complete list of the whole target population. However, in displacement contexts this is rarely the case, and we have to find other ways to go about in order to select a representative sample.

Representative sample

When deciding on a sampling approach, the main goal is always that the sample should be representative of the target population. This means selecting a sample that closely ressembles the characteristics of the target population and thus reflects the target population as good as possible. If, for example, 50 percent of the target population is female, the sample should ideally also consist of 50 percent women. Drawing the sample at random will generally, with a known level of certainty, produce a representative sample. Further measures can also be taken to strengthen the likelihood of obtaining a sample that is representative of the target population by introducing strata. Stratification will be explained in detail further below in the document.

JET | JIPS
Essential
Toolkit

Level of confidence and margin of error

These two variables describe the level of accuracy of the results of your sample survey: if you have a confidence level of 95 percent with a margin of error of 5 percent, this means that if the survey is conducted repeatedly (by drawing a new sample from the same sample frame each time the survey is repeated), the results will be within 5 percentage points of the actual population values in 95 percent of the time.

Sample size

Refers to the number of individuals or households in the sample. The sample size should be determined on the basis of certain variables, such as the desired level of confidence, desired level of margin of error and expected response-rate. The amount of funding often influences the feasible sample size.

Response-rate

The percentage share of individuals/households in the sample that complete the survey. The higher the response-rate, the more accurate will the results be. Hence, it is crucial to have a high response-rate to get correct figures. If the response-rate is low, one can try to correct for it by introducing weights. It is important to note that non-response cannot simply be solved by including new persons or households in the sample who were not included initially, because this might introduce bias to your results and you will no longer be able to claim that the results of your survey are accurate with a certain level of confidence and margin of error.

Bias

Refers to a situation where the results from a sample survey systematically over- or underestimate a population parameter (a value that describes the target population). Bias can occur due to different factors and you will find a description of different causes of bias later in this guide.

Weight

A weight is a variable that is sometimes used to provide value for each observation in a dataset. A weight is given to a data variable to increase or decrease that variable's importance relative to the other variables in the dataset. In the context of sample surveys, weights are sometimes used to make samples best match the target population, either in the case of disproportionate stratified samples or in the case of high non-response-rates, which will be touched upon later in this guide. Weights indicate that an observation in a survey represents a certain number of people in the target population.

J E T JIPS Essential Toolkit

The process of choosing a sampling methodology

What do we need to know before choosing the right sampling methodology?

There are several key matters and contextual factors that you need to map out and consider when deciding what sampling approach to choose in a displacement situation. We provide here a list of the most important ones:

- *Establishing what type of analysis is needed to answer objectives*

- *Establishing the target population*

- *Establishing the sample frame*

- *Keeping in mind the availability of resources*

≡+ Example

In a sample survey conducted as part of a *profiling exercise in Sudan in 2018*, IDP households living in two different camps were one of two target populations of the survey. When drawing the sample for the survey, IDP households living in some areas of the camps were oversampled, or in other words had a higher probability of being included in the sample. The reason for the oversampling of certain areas was related to practical measures that had to be taken on the ground. Weights were then used to adjust for the fact that some of the households had had a higher probability of being included in the sample. The weights reduced the importance of the answers from the oversampled households and increased the importance of the answers of the non-oversampled households.

Deciding on the desired analysis outcomes

Clearly establishing and defining the type of analysis that the researchers are seeking to yield from the sample survey in order to answer the objectives of the study, is one of the first matters that needs to be addressed when planning for a sample survey in a profiling exercise. Prior to choosing the type of sampling approach, you need to address questions such as: what do I want the survey to be able to answer? Whom do I want the survey to say something about?

What type of analysis you want needs to be discussed in detail, because it will be decisive for the type of sampling approach that is most suitable. In addition to clearly defining the target population(s) of the study (to be further discussed in detail below), you will need to decide on whether you wish to be able to disaggregate the results by different characteristics of the target population and what characteristics you wish to disaggregate by, implying that you will need to ensure representative results for different sub-groups of your target population. For example, do you want to look closer at the results from a subgroup defined by location or nationality? This might suggest that you will need to create a stratified sample, to ensure representative results for the sub-groups that you wish to study, which will be discussed more in detail below.

Important to note already now, however, is that introducing many strata can drastically increase the sample size needed to produce representative results for each stratum, which will again require more resources when conducting the sample survey. For this reason, careful consideration must be given to what types of analysis levels are necessary for answering the objectives of the survey.

Establishing the target population

The target population is the population group that you want the findings of your sample survey to represent. It is essential that the target population is clearly defined, both in space and time. An example of a target population that is well defined can be found in a JIPS-supported profiling exercise in Thessaloniki in Greece in 2017:

> ≡+ Example
>
> "Refugees and asylum seekers who arrived in Thessaloniki after January 2015"

This target population is defined in space – Thessaloniki – and in time – after January 2015 and up until the time of the survey. In this example it is also important to clarify who is considered a refugee or an asylum seeker. In this case, refugees and asylum seekers were defined as persons who were in possession of asylum seeker pre-registration cards, asylum seeker full registration cards, decision papers on granted asylum, residence permit cards or asylum applications under administrative appeal.

A core component of profiling is a comparative approach: to compare the situation of different population groups. That is, for example: IDPs, refugees, returnees, economic migrants and host populations. Even if the main objective of the profiling exercise is to assess the situation of only one of these groups, it is often necessary to look at the situation of other groups in the same area (e.g. same municipality, region or country), to understand if the situation and characteristics are specific to that population group or shared by all.

If you decide to conduct a sample survey for more than one population group, you will thus need to define several target populations, which should in turn all be clearly defined, as described above. Having said that, defining target populations in a displacement context is not always an easy task; it might, for instance, be challenging for partners to agree on the definitions of target populations. In some cases, there are internationally agreed-upon definitions that one can lean on (e.g. refugees[3] and IDPs[4]). In other cases, however, there are no such guidelines on how to define certain target populations in displacements contexts (e.g. how to define a "host population"), and you will need to develop and agree on a technical definition with the interested partners for each specific study.

Even though clearly defining the target populations might at times seem like a challenging task, we wish to stress the **importance of working towards establishing clear definitions of the target populations in question. It is an essential prerequisite for conducting a sample survey** and will make your work easier when selecting the sample, conducting the survey and analyzing the results.

3   Expert Group on Refugee and Internally Displaced Persons Statistics (2018). International Recommendations on Refugee Statistics. *https://www.jips.org/jips-publication/international-recommendations-on-refugee-statistics/*

4   Expert Group on Refugee and Internally Displaced Persons Statistics (2018). International Recommendations on IDP Statistics. *https://unstats.un.org/unsd/statcom/51st-session/documents/BG-item-3n-international-recommendations-on-IDP-statistics-E.pdf*

J E T | JIPS Essential Toolkit

Establishing the sampling frame

The next step in the process of selecting a sample is to establish a sample frame, i.e. list of individuals or households you can draw a sample from. This sample frame should ideally **include all the persons or households in the target population**. Establishing or getting access to an exhaustive and up-to-date list of the target population is the ideal scenario in any sample survey exercise and the best point of departure for designing a sampling approach that will enable you to select a representative sample.

In order to investigate whether such a list exists, or if it is possible to establish one by compiling lists from several sources, one needs to look at secondary data. Relevant sources can for instance be population registers identifying displaced persons or households, recent census data, lists of residents in camps or lists of beneficiaries for displacement-related programs.

In displacement contexts, the availability of lists identifying displaced persons or households will, however, vary from context to context, and sometimes one will not be able to establish a reliable sampling frame based on secondary data. This can be for example due to data sharing sensitivities. Even with access to a list of the target population one might not get access to information of how the list was compiled, and therefore one cannot always know whether it is of sufficient quality.

In the absence of an already existing list of the target population you will need to create one. However, in some displacement contexts you might not be able to create a complete and accurate list of the whole target population. This can be due to a variety of reasons, for example because you have to seek out hard-to-reach populations living in places with limited access, or are highly mobile ones, dispersed across large areas and/or generally trying to remain "invisible", or sometimes simply because you don't have enough funds. The cost of creating such a list through enumeration of the target population is resource-intensive and funding is thus often a limiting factor.

Whether one is able to establish or get access to an accurate and complete sample frame or not, has implications for the design of the sample selection approach, as well as for the interpretation of the findings, i.e. **what and who can we actually say something about based on the survey findings? These implications will be discussed further later on in the guide.**

Below, we present and illustrate possible scenarios you might face when establishing a sample frame for a profiling exercise:

Ideal scenario

One up-to-date and exhaustive register is available

During a profiling exercise conducted in Kosovo in 2018, Serb IDPs residing in specified areas of Kosovo were one of several target populations. The profiling partners were able to access an updated list of all Serb IDP households from a database managed by a public organisation in Serbia.

Possible to establish an up-to-date and exhaustive list by combining different sources

In a profiling exercise in Thessaloniki in 2017, refugees arriving in Thessaloniki after January 2015 were one of the target populations. The profilers combined data from a UN database on refugees (UNHCR ProGress registration database) with a list of residents in a refugee camp to construct a sample frame for this target population. However, it was found that the information in the database was outdated, which leads us to the next scenario.

Less ideal, but often prevailing scenario

Registers are biased and/or not up-to-date

In displacement contexts, it is rare to find registration lists that are completely accurate. This is often because the population in question is very mobile, which in turn makes it difficult to keep a list of addresses or other contact information updated. In the profiling exercise in Thessaloniki, it became clear that the profilers would need to find a way to update the outdated sample frame. Typical approaches to update a sample frame can entail enumeration or consolidation of secondary sources, such as program beneficiary lists or lists held by local authorities.

Need to construct sampling frame

A profiling exercise conducted in El Salvador in 2016 with the target population being IDP households in El Salvador displaced because of violence between 2006 and 2016, is an example of a scenario where no list of the target population was available or possible to construct, combining already existing sources. The profilers thus had to find an alternative way of establishing a sampling frame and selecting a representative sample. A stratified cluster sampling approach including enumeration was chosen. This approach will be described below.

Keeping in mind the availability of resources

Lastly, an important consideration to take when choosing a sampling methodology is the amount of time and funds available for implementing the survey. For instance, doing a full enumeration of a large area can be a costly undertaking, due to the cost of hiring, training and paying enumerators. Similarly, traveling across large areas to access the sampled target population can impose high travel costs, and the amount of funding available will often also affect the possible size of the sample. Time constraints may also affect which sampling approach is feasible.

Therefore, the availability of time and funding needs to be taken into consideration throughout the process of choosing the sampling approach. The amount of resources available might particularly affect the type of analysis possible (disaggregation levels) and sometimes even the number of target populations or the definition of the target population.

What are the different methodology options?

In this section we describe different approaches to drawing a sample, aspects you need to consider when deciding which approaches to choose, their limitations, as well as some examples on how the different approaches can be used depending on the context. We will go through each approach separately; however, in many cases a combination of several of the approaches can be used, as illustrated in the examples.

Sampling methods are categorised as either probability or non-probability approaches. In *probability samples*, each member of the target population has a known probability of being selected. This means that you can calculate how likely it is that a member will be included in the final sample, and **how closely the information from your sample represents the target population** (see definition of confidence interval and margin of error above).

In *non-probability sampling*, members are selected from the target population in some non-random manner. Unfortunately, **non-probability sampling does not allow you to know for sure just how much the information from the sample differs from the target population; this is quite simply unknown**. There are, however, ways to overcome some of the uncertainty, which will be touched upon further below.

Selecting a probability sampling approach is always the preferred option, although not always feasible in displacement contexts. In case you have no other option but to resort to a non-probability sampling approach, you should work towards minimising the expected bias resulting from using a non-probability approach and we will describe an approach that can help you do so.

Probability sampling

When each member of the target population has a known probability of being included in the sample, it is also possible to say something about the uncertainty (i.e. confidence level and margin of error) of your findings, which makes it possible to assess to what extent you can apply the findings of your survey to the whole target population. You will also be able to communicate this to the audience/potential users of your findings, increasing trust and credibility of your findings. **In order to use a probability sampling approach, you need a sample frame**.

When you have access to an already existing sampling frame

The following approaches all require access to a sampling frame that has a good coverage of the target population and that is up to date and accurate.

Simple random sample
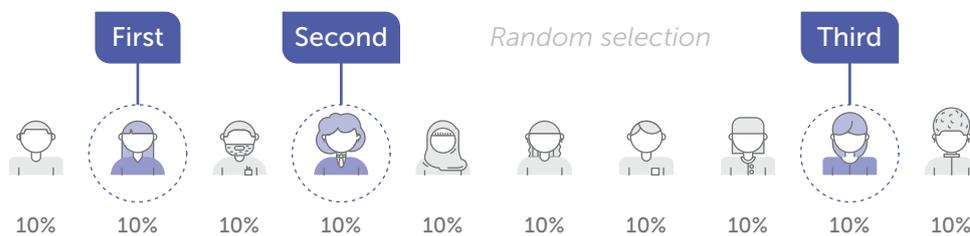
In a simple random sampling approach, one randomly selects persons from the entire sample frame, and each person in the sample frame has an equal probability of being selected. This approach is viewed as the most basic form of random sampling and is often used in combination with another approach, e.g. stratified sampling. Performing a simple random sample selection can be done by using Excel or some other computer software, picking a random selection of persons or households for your sample.

Systematic sample

This sampling approach is similar to the simple random sampling approach, in the way that all persons in the target population have the same probability of being selected. In systematic sampling, however, you select 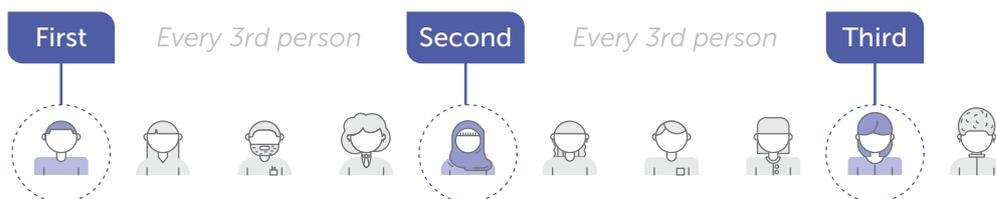your sample from a list with a fixed interval. Let's say you have a paper list of the addresses of all households in your target population, and you select every 12th address on that list to take part in your sample. This method is an easy way to select the sample but should be used with caution if there is some cyclical pattern in the list of households. Let's say for instance, that the list of addresses is ordered by street. Every 12th household selected may have something in common such as always being situated on a bottom flat, or a corner house. These characteristics may in turn affect the answers of the respondents in your sample. Thus, if there is some kind of systematic ordering of the list, you should preferably try to rearrange it before you sample.

The systematic sampling approach is a commonly used technique but should be used with caution to ensure it gives equally good results as simple random sample. The systematic sampling approach is also often used in combination with another approach such as stratified sampling.

Cluster sample

Street 1   *Street 2*   *Street 3*
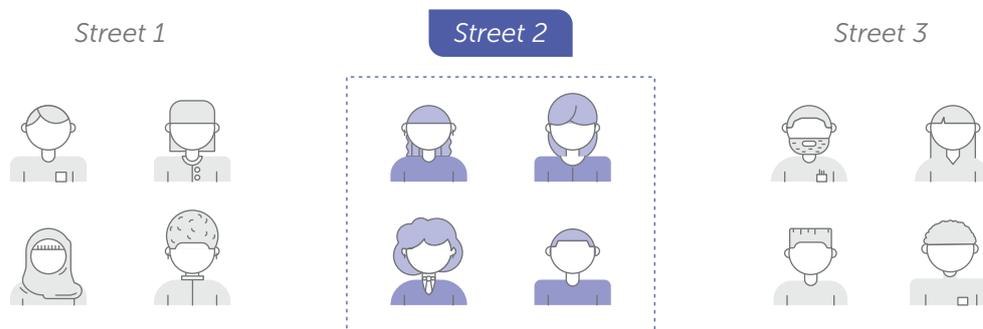
Grouping according to location and choosing a group

In a cluster sample, groups of people are selected rather than persons, such as settlements, schools or districts, etc. The advantage of choosing a cluster sampling approach is that it can often be more cost efficient and administratively easier to conduct. For example, if you were to conduct a sample survey for school pupils in a country, you would save both time and funds by randomly selecting school classes and interviewing everyone in one class, rather than randomly selecting single school pupils dispersed across a possibly large number of schools and classes in a country.
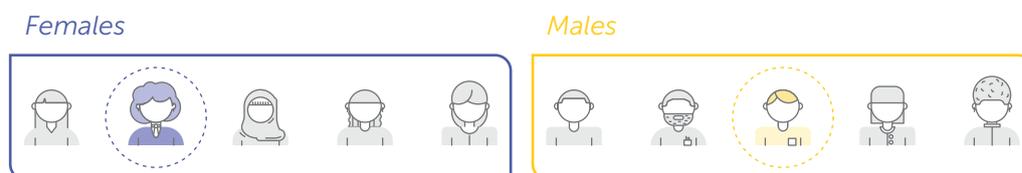
A limitation of the cluster sampling approach is that the persons within the group chosen can be more alike. This can reduce the representativity of the sample and thus increase the uncertainty of the results, compared to simple random sampling or systematic sampling.

Stratified sample

Females   *Males*

In a stratified sample the target population is divided into non-overlapping subgroups, which are called strata, for better control of the sample. The divisions into strata are made based on auxiliary information about some characteristics of the target population, such as the place of residence (e.g. rural/urban areas, regions or municipalities), socioeconomic background, or demographic variables. In stratified sampling, each subgroup is treated as a separate target population, and a separate sample is selected for each stratum. Dividing the target population into different strata often has several advantages.

Firstly, a stratified sample can improve efficiency and the representativeness of a sample. With a simple random sample, we will get a representative sample on average, but we do run a greater risk of having over- or underrepresentation of subgroups in the sample. Stratifying a sample is a way of trying to avoid bias and obtaining more precise estimates of the target population. Furthermore, if each of the subgroups are homogenous when it comes to the variables we wish to study, we can obtain results of desired quality with a smaller sample than with a non-stratified sample.

Secondly, the stratified sampling approach can guarantee representation of small subgroups of your target population. By treating a smaller subgroup as a separate target population, you are able to include a representative sample of this subgroup in your study.

The allocation of the sample size between the strata can be either proportional or disproportional to the size of the target population in each stratum. In the latter case, the probability of being included in the sample differs across the strata. When estimating a population mean based on the survey data, this uneven representation must be factored in by the use of weights. **The weight is equal to the number of persons or households each sample person or household represents. Weights are not applied when conducting analyses that involve comparing the strata to each other; they are only used when calculating figures that cover the whole sample**.

≡₊ Example

In a profiling exercise conducted in Kosovo in 2015, one of the target populations were Serb IDP households displaced between January 1998 and March 2004. A complete list (sample frame) of the target population was available to the profilers. The majority of the target population was residing in the northern municipalities of Kosovo, and a smaller proportion in the southern municipalities.

The partners in the profiling exercise had information indicating that the situation of the IDPs living in the south was different from the situation of the IDPs residing in the northern municipalities. It was therefore desirable to be able to disaggregate the results by location (north and south) to be able to compare the situations of the target population in the two locations. It was thus decided to stratify the target population into northern and southern municipalities to make sure that the overall sample of IDPs would be representative of the target population residing in both the northern and the southern municipalities. The sample size was equally distributed between the southern and northern municipalities, i.e. disproportional to the size of the target population in each stratum, which implied that the use of weights was necessary in the analysis of the findings whenever results were calculated on the national level.

The profilers also wanted to be able to compare the situations of the IDPs residing in urban areas and in rural areas. They therefore stratified the sample further into urban and rural areas, proportionately to the size of the target population in urban and rural areas. Thus, use of weights was not necessary when compiling findings for these strata together.

A simple random sampling approach was used within each of the strata to select the sample.

When you do not have access to an already existing sampling frame?

In displacement contexts, it is more often than not, the case that we do not have access to a complete and up to date sampling frame. This means that we often need to create one ourselves. But identifying and listing all persons of a specific target population of a study is most of the time a too demanding and costly task. This in turn, means that we will not be able to rely solely on the approaches described above. A common way of working around this challenge is to choose a two-stage sampling approach.

J E T | JIPS
Essential
Toolkit

Two-stage random sample

Primary Sampling Unit

Geographic Area 1   *Geographic Area 2*

*List of the Target
Population per PSU*
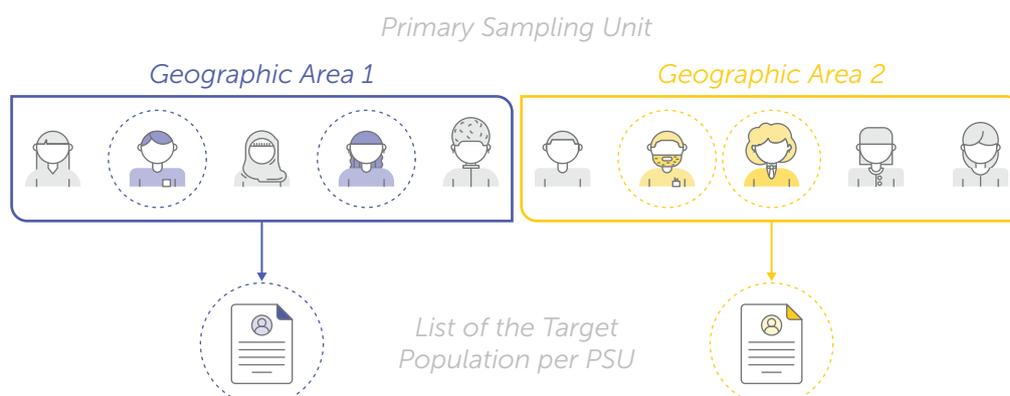
The approach is, as its name suggests, conducted in two stages.

In the first stage, you divide the areas your population lives in into smaller areas – so-called Primary Sampling Units (PSUs), which can be for example municipalities, neighbourhoods, blocks or census segments[5] – and select a sample of these units. How many PSUs you need to select depends on the total number of households to be included in your sample and the size of the PSUs. When you have made a list of geographic areas to draw a sample from, you can either draw randomly or systematically.

If you know how many of your target population live in each unit, draw the sample with a probability proportional to the number of the target population living in the area. This means areas with many members of your target population will have a higher probability of being selected: if you have twice as many target population members in area A than area B, the chance for an individual or household of being selected in area A should be twice as high as for those in area B.

In the second stage, you make a list (a sample frame) of the target population (e.g. IDP households) living in the selected PSUs, often by conducting an enumeration of all households within the selected units, or by consulting local administration or local organisations. You then draw a random sample from each PSU based on these lists. The households are called the secondary sampling units (SSUs). If the primary sampling units are of a similar size, the same number of households should be selected in each one. Conversely, if the primary sampling units are of different sizes, the number of secondary sampling units selected from each PSU should be proportionately distributed between the PSUs according to the number of SSUs in each PSU.

One of the benefits of drawing a sample in two stages is having a sample of households located in the same area, thereby reducing travel costs for enumerators. However, the fact that the sample is selected in two stages increases the uncertainty about the findings and will impose something called a design effect. This implies that you will need to increase your sample size compared to the size you would have chosen if you were to use simple random sampling.

---

5   Geographically defined area established by the national statistical office for the purpose of conducting a census.

≡₊ Example

In a profiling exercise in Mogadishu, Somalia, in 2015, the target populations were IDPs, economic migrants and host communities in Mogadishu, living in informal settlements across three districts of the city. A mapping of all settlements within these districts was carried out, followed by an enumeration of households in all the identified settlements to group them into different target populations.

Following this, a two-stage sampling approach was conducted for the IDP target population. The settlements were stratified into the three districts. In the first stage, settlements (primary sampling units) were randomly selected within each of the districts, based on probability proportional to size of the IDP population in each settlement. In the second stage, a random sample of households (secondary sampling units) was drawn within each of the selected settlements based on the lists made during the enumeration. The households selected were interviewed for the sample survey.

Non-probability sampling

In contexts where there is no existing sampling frame available, a two-stage approach is the preferred option. However, if it is particularly difficult to find the displaced households despite mapping the populations, or if you lack resources for conducting a full enumeration of primary sampling units, you may have to turn to non-probability sampling. Using non-probability techniques when establishing your sample will probably give you more biased results compared to when using probability techniques, as some groups are often not represented in the sample. Furthermore, it is difficult to assess if a non-probability sample is representative. In other words, it is difficult, if not impossible given existing tools at our disposal, to calculate the uncertainty of any estimates you make. If you decide to use non-probability sampling, you should always make sure to state these limitations clearly when communicating the results of the study.

Used with caution and with measures taken to increase representativity, the technique has however proven to provide reasonably sound results, despite the fact that results are more uncertain and that the sample size needs to be increased, compared to systematic or random sampling.

Snowball sampling
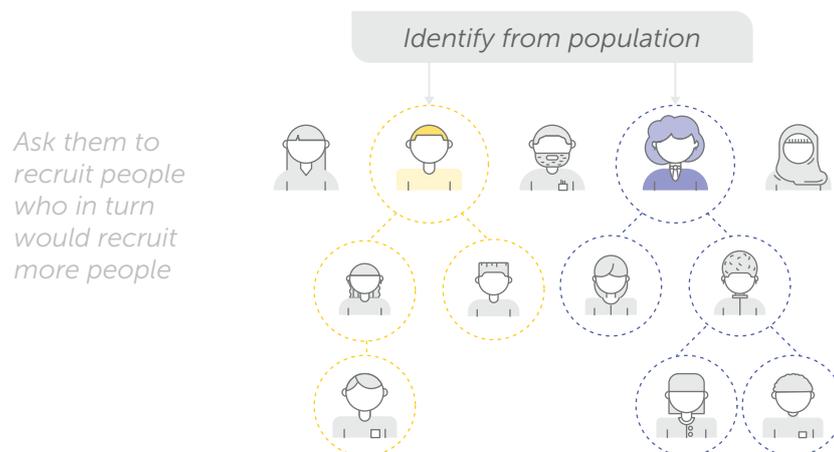
One of the most "basic" and most well-known forms of non-probability sampling is the so-called snowballing approach. Snowball sampling relies on the network structure of the target population. This sampling method involves identifying some persons that belong to the target population and ask those if they can recruit or refer other persons belonging to the same target population. The persons recruited by the initial subjects are in turn asked to recruit or refer other persons, and so this process goes on and the sample grows. You continue this process until you reach the sample size needed (which should be larger than in a simple random sampling) or until you reach the point of saturation. Reaching the point of saturation means reaching a point where you are certain that you have identified all persons in the target population, and no new referrals are made. This is the only way of being sure that you have reached a representative sample (in practice a full count) because the whole target population is included. In such cases, snowball sampling can also be used to create a sample frame, from which you can select a simple random sample. However, it can be difficult to know whether you have actually reached the point of saturation even if no new referrals are made. This could be due to the fact that in displacement contexts people might want to stay invisible or you might be missing out on people with no or small networks. Reaching saturation can be possible in cases with relatively small target populations, but with a large target population this is rarely the case.

An important limitation of the snowball sampling approach is that respondents are likely to recruit persons that share similar characteristics as themselves, which reduces representativity. Furthermore, persons with large networks are more likely to be included in the sample and they in turn might refer relatively more persons compared with persons with smaller networks. With this in mind, a way of increasing the chance of obtaining a representative sample is to make sure that the initial set of respondents differ when it comes to important characteristics, such as certain demographic and socioeconomic variables.

Another limitation of the snowballing approach is that in certain contexts the respondents might not wish to refer to other persons in the target population in order to protect them from a potentially perceived threat related to being identified as part of the target population. This is referred to as "masking". A way of avoiding masking is to ask the respondents to recruit other respondents themselves, instead of only providing names and contact information to the profilers. This way, potential respondents have the possibility to choose whether they want to participate or not without first being identified by the profilers.

Respondent-driven sample

Respondent-driven sampling is similar to the snowball sampling but has been developed as an attempt to compensate for the fact that the sample was selected in a non-random way. The approach combines snowball sampling with a mathematical model and aims to generate samples that are independent of the initial respondents from where the sampling begins, as well as correcting for the differences in people's network sizes[6].

The approach tries to limit the bias related to differences in network sizes, firstly by limiting the number of referrals from each respondent (usually to three referrals). This is done in practice by providing a limited number of coupons to each respondent for him or her to pass on to other members of the target population. Each coupon has a unique ID number, and also contains the ID number of the coupon of the person recruiting that respondent. This way one is also able to keep track of who was referred by whom, which may also help to avoid people being referred more than once. Secondly, the approach tries to correct for differences in network size by applying weights to the data of the respondents, according to their network size. The data provided by persons with smaller network sizes is weight more heavily than the data provided by persons with larger network sizes. To be able to do this, one needs to collect information about each respondent's network size during data collection (e.g. adding a question about how many persons within the specific target population each respondent knows).

The approach also suggests providing incentives (e.g. small amount of cash) to the respondents, both for participating, and for recruiting new respondents. Providing a small compensation for each new recruitment can result in a higher willingness to participate in the survey and therefore increase the sample.

Recruitment of new respondents continues until one has reached the required sample size, if the target population size is known, or until one has reached a so-called equilibrium. Equilibrium is reached when the characteristics of the initial respondents no longer influence the sample, for instance when the ratio of men and women in the sample remains stable among recruitments.

Limitations of the respondent-driven sampling approach includes the fact that a number of assumptions need to be fulfilled in order for the approach to produce representative samples. For instance, the assumption that people with larger networks tend to be oversampled. In addition, one needs to use specialised data analysis tools to perform the analysis of the results. There are, however, free tools available online, see for instance RDSAT[7].

---

6   Heckathorn, Douglas D. (1997). Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations. *http://www.respondentdrivensampling.org/*

7   *http://www.respondentdrivensampling.org/*

How to decide which sampling approach is best?

To sum up the above discussion of the different sampling methodology approaches, we have mapped out a simple "decision tree" below, to try to point you in the right direction when deciding which sampling approach to use.

Already existing sample frame?

Yes
- Simple random sample
- Systematic sample
- Cluster sample
- Stratified sample

No → **Possibility to create sample frame?**

Yes
- Two-stage random sample

No
- Snowball sample
- Respondent-driven sample
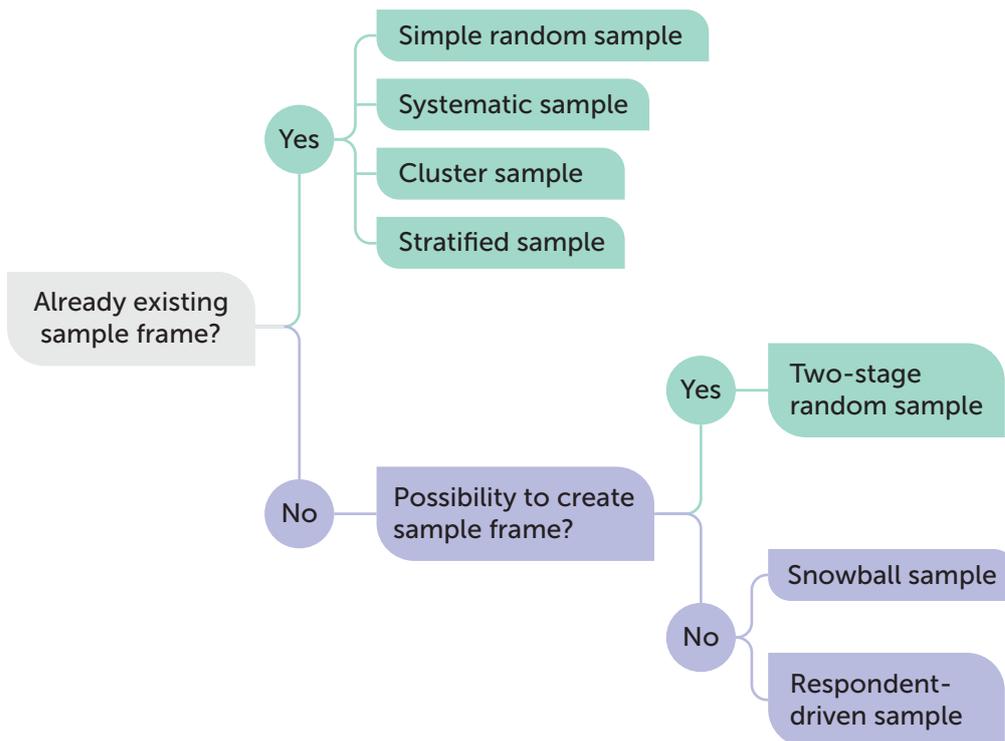
Deciding on the sample size: How many should we ask?

To determine the sample size for a purely random sample, the context needs to define the level of uncertainty one is willing to accept. Usually the decision on the level of acceptable uncertainty is influenced by the cost implications and the plans for how the final results will be used. For a profiling exercise, a confidence level of 95% with a margin of error of +/- 5% is sufficient to analyse broad trends within a population. The size of the overall population also affects the sample size needed, but generally this does not change much if the population is 20 000 or above.

Online sample size calculators (for example Raosoft)[8] can help to determine the sample size needed for a survey. Below is an example of the results for different sized populations, levels of confidence and margins of error.

≡+ Example

For population of 20,000	5% error margin	2% error margin	1% error margin
90% Confidence level	270	1560	5060
95% Confidence level	380	2140	6490
99% Confidence level	640	3440	9070

For population of 150,000	5% error margin	2% error margin	1% error margin
90% Confidence level	270	1670	6470
95% Confidence level	380	2360	9030
99% Confidence level	660	4040	14940

The size of the sample also needs to take into account the complexity of the sample design, described by the design effect. In short, if conducting a two-stage sample approach, you need to increase your sample size due to the complexity of the approach. In this case, it is recommended to use a design effect of 1,8.

The response-rate is also relevant for the sample size. It is recommended to base the response-rate estimate on previous surveys in the same context or other information that can be used to make an educated guess for the expected response-rate. If no other information exists a response-rate of 50% is typically assumed when calculating the sample size.

It is important to note that the sample size calculators are generally based on random or systematic sampling; if using Respondent Driven Sampling, the sample size should be increased above the recommendation of the calculator. How large the increase ought to be is difficult to calculate, and largely depends on how much resources are available.

---

8   *http://www.raosoft.com/samplesize.html*

What are the possible errors that might occur?

Findings of a sample survey will never completely reflect the actual characteristics of a target population – there will always be some error, although, as we have seen above, some sampling approaches generally produce less errors than others. Errors in sample surveys can be divided into *sampling errors and non-sampling errors*[9].

Sampling errors are errors resulting from the design of the sampling and occur because the findings are based on a sample rather than the whole target population. The *estimation error* is the difference between the values found in the sample and the actual values of the target population (see definition of *margin of error* in the terminology section). Another type of sampling error is called specification error, which is given when the true probability of being included in the sample differs from the probability intended according to the sampling design. We would face a specification error if, for example, the same individual or household is listed twice or more in the sample frame.

Non-sampling errors are errors that occur in the process of obtaining answers to the questions asked in the questionnaire. Non-sampling errors include situations where persons who do not meet the criteria of the target population are included (*inclusion error*) and where people who meet the criteria of the target population are left out (*exclusion error*). Profilers in displacement situations often risk facing inclusion and exclusion errors and need to be particularly aware of this risk. Another type of non-sampling error is measurement error, which occurs if the respondent misunderstands the question, does not want to provide the right answer, or if interviewers make mistakes when recording the answer. A *measurement error* can also occur if for instance the interviewers ask leading questions or if the respondent has trouble remembering the correct answer, e.g. when asking for recall periods. A last type of non-sampling error is *non-response*: when persons in the sample refuse to respond to the questionnaire all together or abstain from answering some of the questions in the questionnaire.

---

9   Bethlehem, J. (2010). Selection bias in web surveys. *https://pure.uva.nl/ws/files/1040458/92124_330481.pdf*

The importance of documenting

How the results of the sample survey should be interpreted and used depends a lot on the design of the sampling approach, as we have seen above. This makes it highly important to be transparent towards partners and the general audience about how the sample approach was designed and how the sample was actually selected in the end, as well as limitations to the approach chosen. Profiling in displacement situations can often entail several unexpected turns of events, also when it comes to selecting the sample. This implies that you might not always be able to take the sampling approached initially planned. In such cases it is important to remember that what should be communicated to the audience should reflect the approach actually applied when selecting the sample. Even if you had planned for a two-stage sample approach with enumerations and random selection of sampling units, but due to some unexpected event you ended up having to do snowball sampling, it is the snowball sampling approach which was actually undertaken that needs to be communicated to the audience.

In addition to transparency towards partners and the general audience, it is strongly advised to thoroughly document the outlined sampling methodology as well as all the steps taken to reach a sample for your survey, for internal future reference. This includes explaining challenges encountered throughout the process and how those were solved. If you document the steps towards selecting your sample, this can serve you or your colleagues well, when seeking to develop a sampling methodology for future data collection exercises.

Where can I learn more?

The Department of Economic and Social Affairs of the United Nations Statistics Division has produced a handbook on household survey sampling called "Designing Household Survey Samples: Practical Guidelines". It is available in Russian, Spanish, French, Arabic and English:

http://unstats.un.org/unsd/publication/seriesf/Seriesf_98e.pdf

A lot of useful information can also be found in the handbook on "Household Sample Surveys in Developing and Transition Countries". It presents best practices on several aspects of conducting household surveys in developing and transition countries, including sample design, survey implementation, non-sampling errors, survey costs, and analysis of survey data. The publications are available in Arabic, Chinese, English, French, Russian and Spanish.

http://unstats.un.org/unsd/publication/seriesf/Seriesf_96e.pdf

The "Integrated Framework for Household Survey" developed by UNHCR is a toolkit to facilitate design, collection and analysis, among other topics on sampling:

https://unhcr.github.io/Integrated-framework-household-survey/Sampling.html

UNHCR also provides guidance with the "Standardised Expanded Nutrition Survey (SENS)" that can be useful for implementing household surveys among IDPs and refugees:

http://sens.unhcr.org/

UNICEF provides several tools with their "Multi-Indicator Cluster Survey (MICS)", including a tool on mapping and household listing, which can be useful information if you are considering a two-stage sampling approach with enumeration:

http://mics.unicef.org/tools?round=mics5

See also US Aid's "Demographic and Health Surveys (DHS)" manuals and training material:

http://www.dhsprogram.com/publications/publication-dhsm4-dhs-questionnaires-and-manuals.cfm

http://www.dhsprogram.com/What-We-Do/Student-Resources.cfm

Eurostat's publication "Survey Sampling Reference Guidelines" gives a good overview of probability sampling approaches:
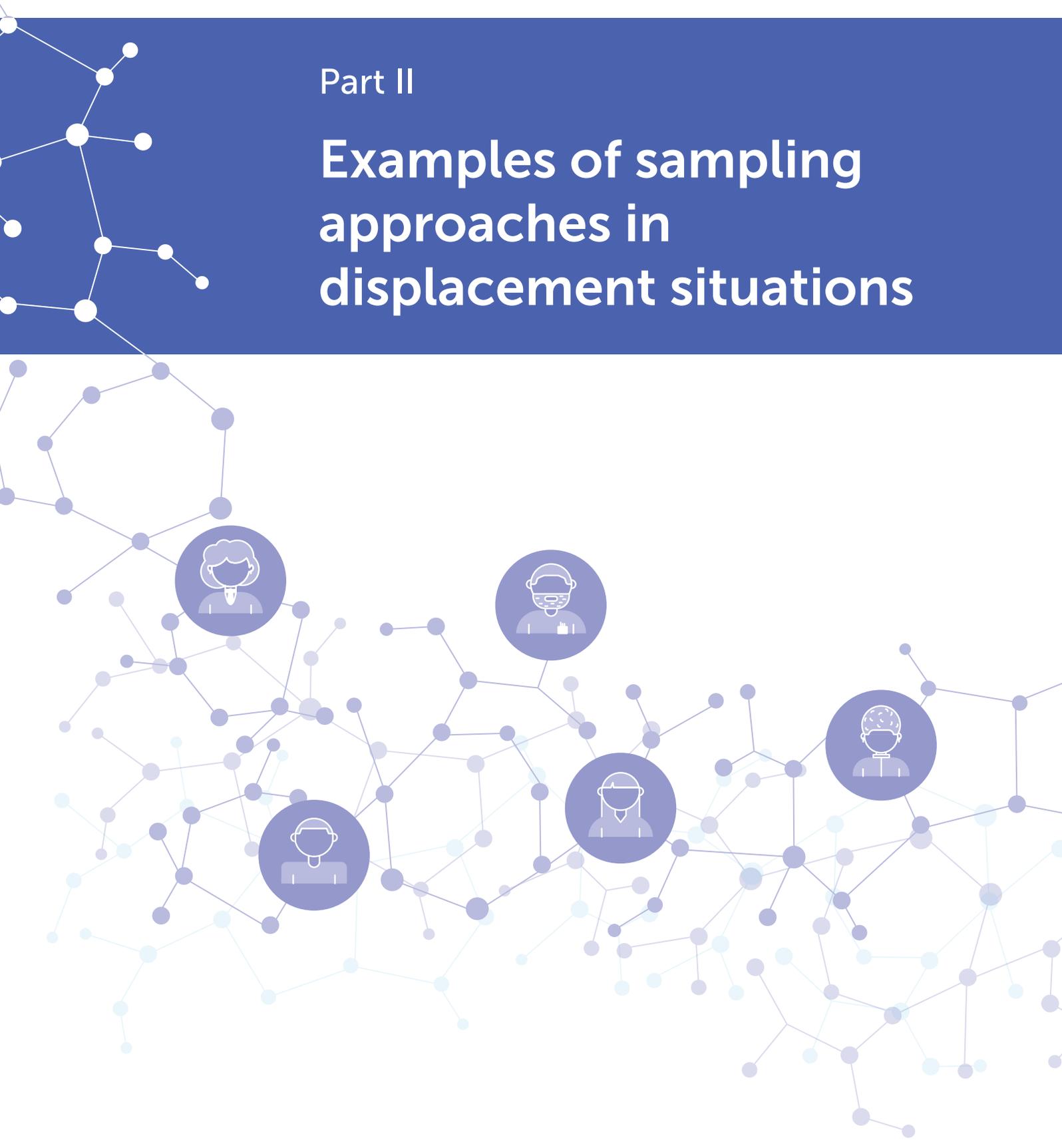
https://ec.europa.eu/eurostat/documents/3859598/5901961/KS-RA-08-003-EN.PDF/833f7740-0589-47e1-99a5-c14878a2c1a8

See also this website on respondent driven sampling:

http://www.respondentdrivensampling.org/

Part II

Examples of sampling approaches in displacement situations

Introduction and general remarks

The objective of this document is to review and consolidate various approaches to sampling that have been used in a selection of JIPS-supported profiling exercises in recent years. The document gives a detailed description of the different approaches used, explains why they were chosen and identifies their possible limitations.

In displacement contexts the operational reality is one of a variety of constraints that influence the choice of sampling approaches. This may sometimes result in "second best" probability sample approaches or even non-probability sample approaches. However, these may still provide sample surveys with useful results, if transparency regarding how the sample was selected is ensured, and if the process is openly and thoroughly documented as part of the methodology description in the profiling report.

This document seeks to provide examples as to how sampling was conducted within the limitations of the respective contexts to ensure a "good enough" approach. More specifically, the document goes through the sampling methodology of six profiling exercises, conducted in El Salvador, Greece, Iraq, Kosovo, Somalia and Sudan. We aim for this to be a living document and to add further examples to it as profiling exercises are completed. We hope to inspire and foster a community of practice which capitalises on transparency and leads the way in terms of responsible data processes.

Key take-aways from the review include:

Importance of documenting sampling methodology and limitations:

A general lack of documentation of the sampling methodology **limits the usefulness of the survey results** to others than the profiling partners themselves. Where the methodology and limitations are shared, it becomes easier not only to understand and use the generated data but also to share it – ensuring a responsible data management process is in place.

The lack of documentation and transparency around the sampling limitations can result in the **misinterpretation of results** (i.e. the extent to which they can be generalised); further contributing to the expectation that fully probabilistic sample surveys are the norm even in displacement contexts. However, all empirics show that this is rather the exception.

Sampling limitations and the objectives of the exercise need to be viewed in conjunction and their interdependence clearly communicated in order to manage expectations as to the analysis possible:

The specific objectives of the exercise and the analysis expected set the requirements for the sampling approach. Careful consideration is therefore needed to formulate the objectives of the exercise as concretely as possible. Sometimes, the sampling approach will be limited by what is operationally feasible. In such cases, it is key to revisit the objectives and ensure that these reflect the possible analysis approach.

Designing the sampling approach requires thorough review of the available population data, the population definitions and a thorough understanding of the type of analysis that best will address the objectives:

To create a sampling frame, information on the target population size and its distribution across locations is required. In most cases, however, up-to-date registers are not available for a variety of reasons. In practice, other sources may be good enough to establish a sampling frame. This necessitates several steps of preparation prior to carrying out the survey (e.g. multi-stage sampling). Often an enumeration exercise will be required.

The profiling of displacement situations will typically sample displaced and non-displaced populations to allow for a comparative analysis. These target populations will require clear, operational and agreed upon definitions. Additionally, the analysis may aim for a comparison of subgroups, e.g. based on geographic location, accommodation characteristics, etc. This will require additional distinct sampling frames and approaches. Identifying the relevant sub-groups to sample requires a very good understanding of the displacement context and the programming or policy which the results will inform.

Different target groups in many cases require different sampling approaches, since the availability of sampling frames may differ by group. This may entail limitations in the extent to which the results of sample-based surveys are suitable for comparison across target groups.

Operational realities during data collection may influence the final sampling design; it is paramount that such changes are captured transparently in the final sampling description:

Common challenges during data collection include outdated personal information such as addresses, phone numbers, etc. This influences the validity of the sampling approach chosen and need to be factored in to attain the necessary sampling size. Dynamic factors of the operational reality can also influence the sampling approach and the overall exercise: those include a changing security situation, access to the populations and areas to be assessed, movement of the target population, population groups not wanting to be identified, etc. As a result, often the sampling design may need to be adjusted or sometimes even entirely changed. It is key to reflect such changes in the original sampling description.

Before choosing a sampling approach, there are a minimum of three things that (almost [1]) always need to be in place: **(1)** the objective(s) of the research, **(2)** the target population(s) which one wants to study, and **(3)** an established sampling frame or at least a plan how this will be obtained. Consequently, these three points are also included in the description of each of the six use cases, before diving into the sampling approaches. At the end of each example, main limitations and challenges to the approach chosen are also identified.

In the next section a summary table is presented to provide a brief overview of different key aspects of the sampling approaches in the different exercises. This table can serve as a point of reference for looking up examples with different characteristics. The rest of the document is an overview of the sampling approaches in the different exercises. For an explanation of different terms and sampling methodology approaches, refer to the document Sampling Guide to Displacement Situations [2].

Profiling

A collaborative data collection approach/system that seeks to **establish a shared understanding of displacement situations** and the circumstances and characteristics of those affected. It uses **mixed-method approaches, which often entail sample surveys**, to collect and analyse data on displaced populations, their host communities and others, and situates this in broader considerations of the economic, political and social backdrop of displacement. The overall aim is to create a comprehensive and mutually **agreed evidence base to inform more effective humanitarian and development interventions, advocacy efforts and the development of national policies** to support the achievement of durable solutions for displaced populations.

---

1 The exception to this rule is with snowball sampling or respondent-driven sampling, when a reliable sampling frame is unavailable and (3) is thus not fulfilled.

2 See Sampling Guide Part I

Summary table

	Target population(s)	Sample frame	Sample methodology
El Salvador, 2016-2018	IDP population in El Salvador displaced because of violence between 2006 and 2016	No full list of IDP population available. Whole population of El Salvador served as sample frame	Stratified random sample
	Population in El Salvador not displaced because of violence between 2006 and 2016	Whole population of El Salvador served as sample frame	Two-stage stratified sample
Thessaloniki, Greece, 2017-2019	Refugees and asylum seekers who arrived to Thessaloniki after January 2015	Constructed combining two different but outdated sources	Census and snowball sample (Stratified random sample attempted)
	Persons without asylum service documentation	No list available, estimate of total number of households available	Snowball sample
Erbil, Kurdistan Region of Iraq, 2015-2016	Syrian refugees residing in specified areas in Erbil Governorate	List available, unclear whether complete	Simple random sample
	IDP population displaced since December 2013 residing in specified areas in Erbil Governorate	List available, unclear whether complete	Two-stage stratified sample
	Host community residing in specified areas in Erbil Governorate	List available, unclear whether complete	Two-stage stratified sample
Kosovo, 2015-2016	IDP Albanians displaced between January 1998 and the end of March 2004, residing in specified municipalities	Constructed combining sources, but outdated	Snowball sample (Stratified random sample attempted)
	IDP Serbs displaced between January 1998 and the end of March 2004, residing in private accommodation in specified municipalities	List available	Stratified random sample
	IDP Serbs displaced between January 1998 and the end of March 2004, residing in Collective Centres in specified municipalities	List available	Census
	IDP Roma/Ashkali/ Egyptians (RAE) displaced between January 1998 and the end of March 2004, residing in specified municipalities	Compiled based on two different outdated and incomplete sources	Census and snowball sample

J E T JIPS
Essential
Toolkit

	Target population(s)	Sample frame	Sample methodology
Mogadishu, Somalia, 2015-2016	IDPs residing in informal settlements in Mogadishu	No list available, created through mapping and enumeration	Two-stage stratified sample
	Economic migrants residing in informal settlements in Mogadishu	No list available, created through mapping and enumeration	Simple random sample
	Host communities residing in informal settlements in Mogadishu	No list available, created through mapping and enumeration	Simple random sample
Sudan, 2017-2019	IDPs residing in the two camps of Abu Shouk and El Fasher	No list available, created through mapping and enumeration	Two-stage cluster sample
	Non-displaced population residing in peri-urban and urban El Fasher	No list available, created through mapping and enumeration	Two-stage cluster sample

Profiling exercises and sampling methodology

El Salvador profiling exercise, 2016-2018

⊙ **El Salvador**

Objectives

The objectives of the exercise were to identify the number of families displaced by violence in El Salvador in recent years, and to better understand the living conditions and specific needs faced by the victims of displacement[3]. Specifically, the study sought to

(1) Obtain statistical estimates of the **number of families and individuals** who were forced to move internally because of criminal violence, with estimates representative at the national level;

(2) Identify the **causes and specific impact** of internal mobility due to criminal violence;

(3) Describe the **profile and condition** of the population that was forced to move internally because of criminal violence.

To obtain the required information, the study adopted a mixed-methods approach using qualitative and quantitative data collection techniques, including a sample survey which is the focus here.

Target population

The profiling exercise had two target populations:

(1) The main target population of the study were **members of households[4] in El Salvador where at least one member had been displaced** because of violence between 2006 and 2016.

(2) In addition, another target population was also defined for the purpose of comparing the situations of the IDP households to that of non-IDP households. The comparison target population was determined to be members of households in El Salvador where no members had been displaced because of violence between 2006 and 2016.

---

3 Profiling study on internal mobility due to violence in El Salvador (2018). *https://www.jips.org/uploads/2018/03/El-Salvador-profiling-report-EN.pdf*

4 Defined in the study as one or several people who live together in the same home and share expenses to provide and meet their dietary needs.

Sampling frame

There was no register over internally displaced persons or households available to the profiling partners. The absence of such a register was also part of the rationale for this profiling exercise, as one of the objectives of the study was to provide an estimate of the number of families displaced by violence. It was, however, not possible to construct a complete list of all households displaced due to violence due to time and resource constraints. One therefore had to explore alternative ways of creating a sample frame that would allow for the selection of a representative sample and an estimation of the number of persons displaced.
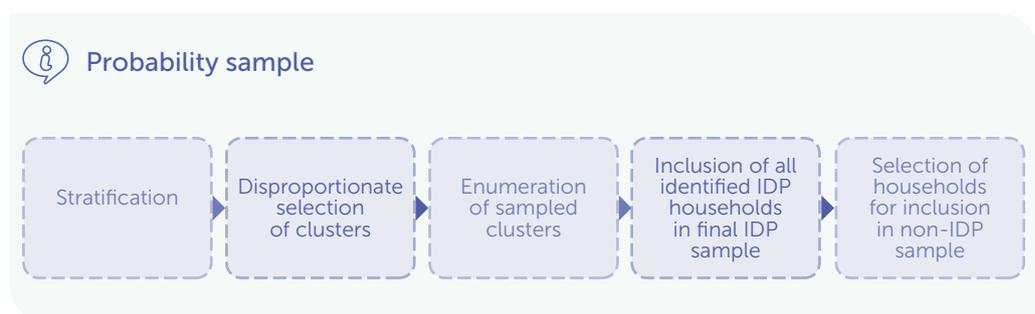
Unit of sampling and analysis

Primary sampling unit	Secondary sampling unit	Unit of analysis
census segments[5]	households[6]	families and individuals

Sampling approach

In this profiling exercise, the objectives entailed both being able to give an estimate of the number of households displaced due to violence on a national level (objective **(1)**), and to provide information about characteristics and conditions of this group (objective **(2)** and **(3)**). Since there were no lists of the main target population available and doing a full enumeration of the whole country was not possible, profiling partners needed to design a sampling approach that would allow them to estimate the entire IDP population on a country level, based on a sample. Furthermore, profiling partners needed to ensure that the sample was representative of the target population, to be able to meet the objectives concerning characteristics and conditions of the displaced households.

For the displaced target population, a stratified random cluster sampling approach was chosen. For the non-displaced target population, a two-stage sampling approach was chosen. The process of selecting the two samples can be summed up in the following way:

Probability sample

Stratification → Disproportionate selection of clusters → Enumeration of sampled clusters → Inclusion of all identified IDP households in final IDP sample → Selection of households for inclusion in non-IDP sample

---

5 Geographically defined area established for the purpose of conducting a census. Such areas can be useful ways of geographically divide larger areas (e.g. municipalities) into smaller ones for the purpose of selecting a sample in profiling.

6 One or several people who live together in the same home and share expenses to provide and meet their dietary needs.

In stratified sampling the researcher divides the target population into different strata (groups), and selects a sample from each stratum. Disproportionate stratified sampling means that the size of the sample drawn from a particular stratum is not proportional to the relative size of the stratum, but rather decided based on another characteristic of the stratum. In a cluster sample, groups of people are selected rather than persons. In cases where there is no sample frame available, selecting a sample of clusters for enumeration to identify members of the target populations is a much-used approach. Below follows a detailed description of how the sampling was carried out.

Step 1 : Stratification

Selecting a number of municipalities at random in El Salvador to enumerate, would have been one option to proceed. However, while this approach might have provided sufficient information to estimate the number of IDP households in the country, the sample size of displaced households would have been too small to address the other objectives listed above i.e. to better understand the displaced households' different situations and needs. To compensate for this, strata were introduced to oversample in areas where sources agreed there were likely to be higher proportions of displaced households.

Starting out, profiling partners gathered information about how they would expect IDP households to be spread across the country to inform their stratification. Through a review of existing sources, it was found that 20 municipalities within El Salvador were expected to have a greater number of households displaced because of violence than other municipalities. The sources reviewed consisted of statistics on demography and migration, as well as on incidents of violence at the municipality level.

A list of census segments[7] in each municipality within the country, which was made available to the profiling partners by the El Salvadoran General Directorate of Statistics and Censuses, allowed for a further breakdown of municipalities. The census segments served as clusters and primary sampling units. Interviews and participatory workshops held with key informants from the 20 prioritized municipalities, resulted in a list of 648 out of 4,427 census segments which were expected to have a higher concentration of IDP households.

Based on this information, the census segments of the country were divided into three strata:

(1) 648 prioritised census segments within the 20 prioritised municipalities

(2) Remaining 3,779 census segments within the 20 prioritised municipalities

(3) 7,996 remaining census segments in the country

Step 2 : Deciding on the sample size

Within each of the three strata, a random selection of census segments was carried out. The size of the sample of census segments was determined based on technical criteria for accuracy (95% confidence) for estimations of low incidence[8], with an assumed high rate of non-response[9], and considering the high design effect[10] caused by stratification. The sample size was determined to be 542 census segments which was equivalent to an estimated goal of 40,650 households.

---

7 Geographic area smaller than municipalities, provided by the El Salvadoran General Directorate of Statistics and Censuses.

8 In a similar study in Honduras, the average proportion of IDP households was 4%.

9 The rate of non-response in the Honduras study amounted to about 50%.

10 Implies increased uncertainty in the findings due to the complexity of the sampling approach.

Step 3 : Selecting a sample of clusters (census segments)

The sample was disproportionately distributed among strata to guarantee having a minimum number of the target population in the sample – assigning a higher probability of being selected to the sample for census segments in the two prioritised strata, than for census segments in the remaining strata. The segments were then drawn randomly in each stratum. The below table shows the distribution of the sampled census segments in the three strata.

Strata	# of segments in El Salvador	Distribution of segments in El Salvador	# of segments in sample	Distribution of segments in sample
1	648	5%	184	34%
2	3,779	30%	125	23%
3	7,996	64%	233	43%
Total	12,423	100%	542	100%

Because of the disproportionate distribution of census segments between the strata, weights needed to be applied in the analysis of the results, when considering the findings on the country level, to compensate for the uneven representation of the segments in the sample. These types of weights are called sampling weights, or alternatively inverse selection probability, and are usually calculated by looking at how many households each of the households in the sample represents (sample weight = total number of households in strata/number of households in strata sample). To obtain the number of households within each stratum, the population figure in each strata can be divided by the average household size. In this exercise, profiling partners didn't document how the weights were actually calculated, but it can be assumed that this common approach to calculating sampling weights was applied.

Step 4 : Enumeration of selected clusters (census segments)

Following the selection of the sample of 542 census segments, an enumeration of all households in 501 out of the 542 sampled census segments was conducted. The remaining 41 census segments were inaccessible to the enumerators due to security restrictions. Profiling partners also faced some non-response during enumeration, although the rate was lower than expected (37 %). During enumeration, households were identified as either displaced or non-displaced. Based on the resulting number of displaced households in each of the segments, numbers could be extrapolated to the overall country level[11], in order to fulfil objective **(1)**. The findings of the enumeration also resulted in a sample of IDP households, as well as a sample frame for the selection of non-IDP households for the sample survey that would address the two remaining objectives.

_____

[11] For explanation of the how the extrapolation was conducted, see Annex 1 of the profiling report (only in Spanish): *https://www.jips.org/jips-publication/profiling-report-el-salvador-2018/*

Step 5 : Resulting sample of IDP households

The enumeration resulted in 466 households being identified as displaced because of violence, which amounted to 1,811 persons. These households were all included in the sample and were asked to participate in the household survey. 431 of these households, equivalent to 1,668 persons, agreed and completed the survey.

Step 6 : Selecting sample of non-IDP households

A sample for the comparison target population was also drawn within the sampled census segments, using a simple random sample. For every three surveys administered to a displaced household, one survey was administered to a randomly selected non-displaced household in the same census segment, using the lists resulting from the enumeration of households in the selected census segments. Because selection of the sample of non-IDP households were done in two steps (i.e. first drawing a sample of census segments, and then drawing a sample of households), this approach can be referred to as a "two-stage" sampling approach.

Challenges and limitations

The profiling exercise in El Salvador illustrates an often-prevailing scenario in displacement contexts: the lack of a register of the target populations in question, that can serve as a sample frame. This has implications for the design of the sampling approach as it often results in more complexity, which in turn can increase the uncertainty about the findings. This was also the case for the El Salvador exercise, where profiling partners chose to use a stratified cluster sample and a two-stage sampling approach.

Another common challenge when conducting sample surveys is **non-response**, i.e. simply not managing to get in contact with the persons in the sample, or respondents not wanting to participate in the survey. This exercise also faced a significant non-response-rate during enumeration, as seen above, although the rate was lower than expected. Furthermore, profiling partners faced security restrictions during enumeration, which meant that they were unable to enumerate all census segments in the sample. These are both factors that can lead to bias in the results.

Finally, the **distribution of gender** was not considered when administering the household survey to the sampled IDP- and non-IDP households. This might also have introduced bias in the results.

Thessaloniki profiling exercise, Greece, 2017-2019

Objectives

The profiling exercise aimed to provide the Municipality of Thessaloniki and other relevant actors with an evidence base that could inform the design of advocacy initiatives, policy, service provision, and integration programs, that were better tailored to meet the needs of persons affected by displacement in the Metropolitan area of Thessaloniki.

The specific agreed-upon objectives of the exercise were as follows:

(1) To produce a **demographic profile** of the refugees, asylum seekers and spontaneous arrivals disaggregated by sex, age and other relevant diversity criteria;

(2) To **analyse the capacities, vulnerabilities and coping mechanisms** of the target populations;

(3) To **assess the degree of integration** of the target populations, with a focus on access to services;

(4) To produce a set of indicators for measuring the degree of integration.

Target population

The profiling exercise had two main target populations:

(1) **Refugees and asylum seekers who arrived in Thessaloniki after January 2015**, residing in housing provided by the ESTIA[12] affiliated accommodation scheme, in the open reception facilities (ORF) of Diavata[13], or who were self-accommodated. This target population included persons with asylum seeker pre-registration card, asylum seeker full registration card, decision paper on granted asylum, residence card permit or asylum application under administrative appeal.

---

[12] The Emergency Support to Integration & Accommodation (ESTIA) scheme provides urban accommodation to refugees and asylum seekers in apartments. Run by UNHCR in coordination with 12 partner organisastions throughout Northern and Central Greece.

[13] Provisory accommodation sites created to house refugees and asylum seekers stranded on the Greek side of the Greek-North Macedonian border after the closing of the border in early 2016. The sites were not suited for long-term accommodation.

(2) **Third country nationals not registered with the Greek Asylum Service.** This included persons with police notes, or persons whose documentation issued by the Greek state had expired, or persons who had not been issued documentation by the Greek state. The study included persons residing unofficially in the ORF of Diavata, persons who were hosted by individuals in Thessaloniki as well as persons living in a situation of homelessness.

The geographic scope of the profiling more specifically included the six boroughs of the Thessaloniki Municipality and the surrounding Municipalities of Kalamaria, Neapolis-Sikeon, Pavlou Mela, Kordeliou-Evosmos, Ampelokipon-Menemenis, Delta, Thermis and Oreokastro. Residents of the open reception facility of Diavata were also included in the exercise.

Refugees and asylum seekers without a valid home address were not included in the target population of the study, due to the lack of available information on this group. Furthermore, refugees and asylum seekers of Turkish nationality were excluded from the sample survey because access to their exact addresses was not available. However, focus group discussions were conducted with the Turkish population.

Sample frame

Refugees and asylum seekers: a sample frame was constructed combining two different sources – UNHCR's ProGres database (a full list of persons accommodated by the ESTIA urban accommodation scheme and of persons who were self-accommodated, including contact information) and a site population list over all the residents in the ORF of Diavata.

Third country nationals not registered with the Asylum Service: due to the lack of complete lists or registries of the persons of concern, profiling partners had to look for other ways to design an adequate sampling approach. Different service providers in-country did have information about this target population, but profiling partners were not granted access to them due to protection considerations. However, through comparing aggregated information from multiple service providers, it was estimated that approximately 200 households belonged to this target population.

Unit of sampling and analysis

Unit of sampling	Unit of analysis
individual	household and individual

Sampling approach

The exercise objectives – to better understand the characteristics of the target populations and their situations – implied that the results from the sample survey should ideally apply to the two target populations. This meant that a sample selection based on probability would be the preferred approach, to ensure a representative sample with a known confidence level and margin of error. As noted above, however, this was not possible given that a sampling frame was only available to profiling partners for one of the two target populations.
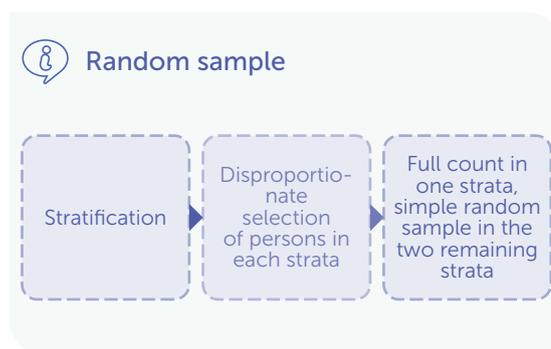
The availability of a complete list of the refugee and asylum seeker target population meant that the sample could in theory be drawn based on probability, whereas the lack of a registry of persons without asylum service documentation necessitated the use of a non-probability sampling strategy, which in this case involved snowball

sampling. While this wasn't ideal and meant that the outcomes of the sample surveys for the two population groups couldn't be directly compared, profiling partners had no feasible alternatives. Although the degree of representativity could not be assessed and the findings could not be applied with certainty to the entire target population[14], the approach would still provide useful data on a population group on which little information existed.

In the following sections, a closer description of the sampling approach chosen is provided separately for each of the two different target populations.

Refugees and asylum seekers

The methodology that was chosen initially for this target population was stratified random sampling. However, during the execution of the exercise, this approach had to be modified, the reasons for which will be explained thoroughly below. The approach that was initially planned can be summed up in this way:

Random sample

Stratification → Disproportionate selection of persons in each strata → Full count in one strata, simple random sample in the two remaining strata

Step 1 : Stratification

The target population was divided into three strata based on their different accommodation types. The three strata were chosen because the situations of the persons in each accommodation type was expected to differ significantly due to the different characteristics of their housing. It was therefore desirable to compare the situations of the persons by their accommodation type and it was thus necessary to make sure a large enough sample was drawn from each of the three accommodation types to be able to compile representative results for each group.

The target population was grouped into the following strata:

(1) Refugees and asylum seekers in the urban accommodation scheme who had been provided with apartments through the ESTIA (1,280 households).

(2) Refugees and asylum seekers self-accommodated in Thessaloniki, i.e. who were either renting an apartment by themselves, or being hosted by friends, relatives or volunteers (482 households).

(3) Refugees and asylum seekers who were fully registered residents of the Open Reception Facilities (ORF) of Diavata (148 households).

Step 2 : Deciding on the sample size

In strata (3), it was decided to do a full count of all the households, due to the relatively small number of households residing in the camp. In strata (1) and (2), the sample sizes were calculated setting a level of confidence to 95 % and a desired margin of error to 5 %. Furthermore, with an expected non-response-rate of 20 % the sample size was adjusted for each strata to reflect the small size of the target population. The resulting sample size for strata (1) was 230 households, and the sample size for strata (2) was 190 households.

---

14 Unless the point of saturation would be reached, i.e. a point where all persons in the target population are identified and no new referrals are made.

Step 3 : Selecting the sample

As noted above, all persons in strata **(3)** were selected to be included in the sample. Questionnaires were distributed to all households in the facility and no one refused to answer the survey, giving a response-rate of 100 % in this stratum. It should be noted, however, that the population in the strata was relatively small in comparison to the other strata and for some variables, the number of observations was very low, which made the analysis of these results less certain.

Following the determination of the sample sizes of strata **(1)** and **(2)**, a simple random sample was drawn in each of these two strata. The probability of being included in the survey differed across the strata: in strata **(1)** 18 % of the target population was included in the sample, in strata **(2)** 39 % was included, and in strata **(3)** 100 % was included. When estimating a population mean of the whole sample this uneven representation would need to be factored in by the use of weights. Such weights, also called the inverse selection probability, are equal to the number of households each sample household represents. The weights would only be used when calculating figures across the three strata (e.g. estimating the average income of the overall target population) while no weights would be used when estimating a mean for only one of the three strata (e.g. average income of population in ORF of Diavata).

When reaching out to the persons in the samples of strata **(1)** and **(2)** by phone, it became quickly clear that the telephone numbers of a majority of the persons listed in the samples were outdated due to frequent changes of sim-cards. Visiting the households unannounced was not possible due to a limited amount of resources, which meant that these persons could not be reached. This implied that the profiling partners would face a too large non-response-rate within the sample for the results to be representative. It was therefore decided to move away from a random sample selection for these two strata and instead, interview all persons in the strata who could be reached, hoping for an acceptable response-rate.

A census survey was thus carried out in the two strata, combined with snowball sampling, which meant that respondents would refer enumerators to other households in the target population. The response-rate for stratum **(1)** was 34,2 % (out of 1,280 households) and for stratum **(2)** the response-rate was 41,2 % (out of 482 households).
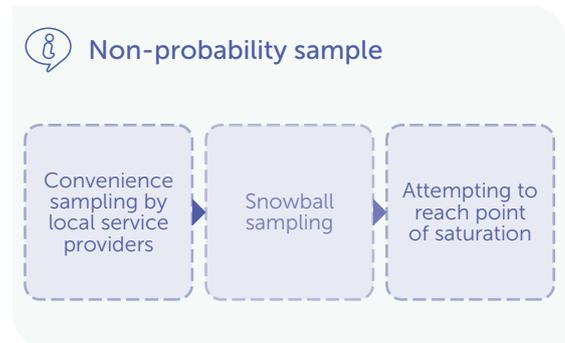
Step 4 : Representativity analysis and post-stratification weights

The relatively high non-response-rates in strata **(1)** and **(2)** introduced a risk of bias in the results, as the persons who replied to the questionnaire might not accurately reflect the characteristics of the whole target population. Hence, an analysis of key demographic characteristics (age, nationality, and sex shares) was carried out among the respondents and compared to those of the target population (available in the ProGres database) to assess the representativity of the respondents. The distribution of characteristics for the respondents resembled the population distributions quite closely. Based on this, it was concluded that the results of the survey could be considered representative.

Doing a census implied that the sampling weights were not necessary to apply. However, to adjust for the different response-rates of the three strata, post-stratification weights were calculated and applied in tables where the respondents of all three strata were considered together, i.e. where the variable of interest was not broken down by strata.

Persons without asylum service documentation

A non-probability sampling approach was applied for this target population, using snowball sampling. Snowball sampling involves selecting a few persons in the target population for interviews, and having these respondents refer enumerators to other members of the target population. Referrals from respondents to new prospective respondents go on in "waves" until enumerators have reached the desirable number of respondents in their sample. The approach taken can be summarized in the following way:

Non-probability sample

Convenience sampling by local service providers ▸ Snowball sampling ▸ Attempting to reach point of saturation

Step 1 : Identifying respondents

Local service providers giving assistance to persons without asylum service documentation were engaged to contact potential respondents and to make appointments for the interviews. In addition, respondents who were approached were also asked to refer to other members of the target population. 200 households had been estimated as an approximate number of the target population and the objective was to reach this number of households, that way hoping to reach the point of saturation or in other words to get a full count of the target population.

Step 2 : Conducting interviews

The number of households interviewed was higher than the number originally foreseen. The enumeration team interviewed 451 persons making up 227 households. The sample covered different accommodation situations: 222 persons were living in a situation of homelessness (living on the street, in unfinished or abandoned buildings, or in a homeless shelter), while 161 persons were living as unregistered residents in the ORF at Diavata, and 27 were living in rented accommodation in Thessaloniki.

A possible explanation for the high number of respondents was an influx of arrivals to Thessaloniki the same month as the interviews were conducted. The survey results supported this theory, as more than half of the survey respondents had been in Thessaloniki for less than a month at the time of the interview. This high number of recent arrivals made the estimate of the total target population more uncertain. In addition, many of the persons who were approached, declined to be interviewed. As a result, it was not possible to assess how representative the interviewees were of the target population.

Challenges and limitations

Profiling partners faced several challenges that limited the possibility of conducting a survey with the preferred sampling approach, both in the initial planning stage as well as during the actual execution of the exercise.

Firstly, challenges when contacting the members of the sample drawn of asylum seekers and refugees resulted in not being able to conduct a survey of a randomly selected sample. The high **non-response-rate** of the target population eventually included led to uncertainty about the results. The representativity analysis of the sample, however, showed that the respondents seemed to be representative of the target populations in key characteristics, and it was thus assumed that the findings were representative and could be extrapolated to the whole target population. Nevertheless, it should be noted that there could still be other characteristics not assessed that differed between the respondents and the target population, which would create more uncertainty about the representativity.

Secondly, **the lack of a sample frame** resulted in a non-probability sample of third country nationals not registered with the Asylum Service. During the data collection stage, the large number of the target **population members that did not want to participate** in the survey further complicated the exercise. Together, this meant that the results were uncertain, and it could not be claimed that the results were representative of the whole target population.

Erbil profiling exercise, Kurdistan Region of Iraq, 2015-2016

Objectives

Erbil Governorate, with a total of 2.01 million people (2014 figures), hosts the capital of the Kurdistan Region of Iraq. The governorate, like the rest of the Kurdistan region, was at the time of the profiling exercise deeply affected by waves of displacement resulting from the conflicts in Syria and the rest of Iraq, as well as by a pervasive financial crisis affecting the public and private sectors of the economy.

While a significant amount of information was available on IDPs and refugees residing in camps, less was known about those residing out of camps. Furthermore, most of the strategies to mitigate the effects of displacement had up until then focused on addressing the needs of IDP and refugee populations, while the needs of the host communities did not receive equal attention.

The specific objectives of the study were the following:

(1) To provide **demographic profiles** disaggregated by gender, age, and displacement status (i.e. refugees, IDPs and host communities) in the targeted areas;

(2) To provide profiles of urban areas with high concentration of out-of-camp displaced populations;

(3) To **analyse the capacities, vulnerabilities and coping mechanisms** of the population in these areas;

(4) To **analyse the relationships** between displaced and host populations.

(5) To analyse the resilience of urban areas in relation to the availability and limitations of services;

(6) To provide a dataset available to the Kurdistan Regional Government and the humanitarian/development community.

Objective **(1)**, **(3)** and **(4)** would be met by using a sample-based household survey.

Target population

Three target populations were defined:

(1) Syrian refugees;

(2) IDP population displaced since December 2013;

(3) Host community (i.e. people that do not fall into the above-mentioned categories including economic migrants).

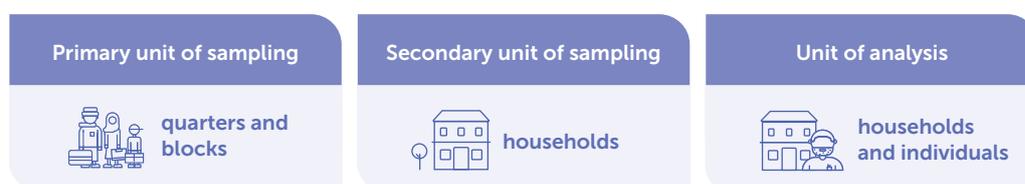All residing in the following three geographic areas:

- Erbil district centre, formed by the city of Hawler (Nawandy Hawler) and Ainkawa;

- Erbil district periphery, formed by the urban nucleus immediately surrounding Erbil district centre and directly connected to it: Baharka, Bnaslawa (Dashty Hawler Centre), Daratu, Kasnazan, Khabat Centre (Nawandy Khabat), and Rizgari;

- Towns, formed by the inner urban centres in the governorate: Harir, Koya Centre, Shaqlawa Centre, and Soran Centre (Nawandy Soran) jointly with Diana.

Sample frame

Three different sample frames were established using different secondary sources, one for each target population:

(1) Syrian refugees: data from UNHCR's ProGres database for registration of refugees;

(2) IDP population displaced since December 2013: data from the first phase of the Comprehensive Registration of Displaced People (CRDP) conducted by Kurdistan Regional Statistics Office (KRSO) in June 2015;

(3) Host community: data from 2009, collected in relation to the preparations of a census survey that was planned but not conducted.

Unit of sampling and analysis

Primary unit of sampling	Secondary unit of sampling	Unit of analysis
quarters and blocks	households	households and individuals

Sampling approach

The objectives of the exercise concerned describing characteristics and situations of the target populations as well as analysing the relationships between the groups. These objectives demanded selecting samples that were representative of the target populations, which meant that a probability sampling approach was preferred. The profiling partners also wanted results that could be compared between the geographic areas as well as between the different target populations within each area. This implied that profiling partners would need to design an approach that allowed for large enough samples of the three target populations within each of the geographic areas. A different sample design was tailored to each of the three target populations. The Kurdistan Regional Statistical Office designed the sampling methodology (partly in collaboration with JIPS), did the selection of the samples as well as the data collection. A short description of the approaches taken for each target population is described below.

Syrian refugees

A simple random sample was drawn from the sample frame. The sample frame was to some extent outdated and/or flawed with no possibility to be updated. Interviews could nevertheless be carried out with households in the sample. Information about response-rate is not available. If the non-response-rate was high, this could introduce a bias in the findings and thus reduce the representativity of the sample.

IDPs

A stratified two-stage approach was used drawing the sample of the IDP population.

Stage 1: The districts with the highest concentration of IDPs were selected from all the three strata. Information about which districts had the highest concentrations of IDPs was available to the Kurdistan Regional Statistics Office through the first phase of the Comprehensive Registration of Displaced. The distribution of the sample between the districts were to be proportionate to the size of the IDP population in each district. Within each of the districts, a number of quarters were picked and enumerated. JIPS does not have information about whether the quarters were randomly selected, or if they were selected based on which quarters had the highest concentration of IDPs.

Stage 2: A sample of IDP households was drawn within each quarter, based on the information from the enumeration.

Host community

The same stratified two-stage approach used for IDPs was also used to draw the sample of the host communities, except that a different primary unit of sampling was used in the 1st stage, i.e. blocks and not districts. JIPS does not have information about why a different sampling unit was chosen and how the blocks were selected (random or not).

Challenges and limitations

There is a general lack of documentation about the sampling methodology in the profiling report. This limits the usefulness of the sample survey results to others than the profiling partners themselves.

The sample of Syrian refugees was picked from an outdated sampling frame and it is thus difficult to determine to what extent the information obtained from the sampled households is representative of the whole population group in each stratum.

Furthermore, because the districts were chosen based on concentration of IDP population within the district, one would not be able to capture IDPs or host communities living in districts that hosted a smaller concentration of IDPs. These households might face different situations compared to the households in high concentration districts. In addition, if IDP households in less IDP-concentrated districts represented a non-negligible share of the total IDP population in the strata, the sample selected would not be representative of the IDP and host community populations living in these districts, but only of the households in high-concentration districts.

Kosovo profiling exercise, 2015-2016

📍 **Kosovo**

Objectives

The population that was forcibly displaced due to the 1998/99 conflict and the 2004 riots in Kosovo has been facing a protracted displacement situation. Estimates at the time of the exercise indicated that the vast majority of the persons displaced within Kosovo resided in private accommodation, while a small proportion resided in collective centers. The profiling exercise aimed to inform a policy on durable solutions, including Durable Solutions strategy of Kosovo's Ministry for Communities and Return, by providing an evidence-based analysis of the displacement situation, according to the IASC Framework for Durable Solutions for IDPs.

The specific objectives were the following:

(1) To produce a **demographic profile** of the displaced population within Kosovo disaggregated by age, sex, location and diversity;

(2) To conduct a **comprehensive analysis of the displacement situation** with regards to the social and economic integration of displaced persons; displacement-related challenges and vulnerabilities; resources and capacities; enjoyment of rights; as well as future intentions and plans;

(3) To enhance institutions' **ability to advocate and design joint programming to support durable solutions** for IDPs through the identification of these groups' specific priorities and enhanced coordination of humanitarian and development analysis.

Target population

The study included four main target populations in below locations, who had all been forced to flee their places of permanent residence in Kosovo between January 1998 and the end of March 2004, in particular as a result of or in order to avoid the effects of armed conflict, the situation of general violence, and violations of human rights, but remained within the territory of Kosovo. The geographical scope of the target populations covered the municipalities with the highest concentration of displaced persons. Persons of other ethnicities were not included in the study.

(1) **Albanian IDPs** in Prishtinë/Pristina, Mitrovica South and Vushtrri/Vucitrn,

(2) **Serb IDPs in private accommodation** in Zveqan/Zvecan, Leposaviq/Leposavić, Mitrovica North, Zubin Potok, Gracanicë/Gracanica, Fushë Kosovë/Kosovo Polje, Lipjan/Lipljane, Shtërpcë/Strpce, Gjilan/Gnjilane, Kamenicë/Kamenica, Viti/Vitina,

(3) **Serb IDPs in collective centres** (all Collective Centers in Kosovo) and

(4) Roma/Ashkali/ Egyptians (RAE) IDPs in Fush K./K. Polje, Gracanicë/Gracanica, Ferizaj/Urosevac, Pejë/Peć, Gjakovë/Dakova and other municipalities where RAE IDPs would be identified.

A comparison group was also defined with the aim to compare the findings for the IDP populations with the situation of non-IDP populations in Kosovo. It was decided that the comparison group would be the general population in Kosovo, and that secondary data would be used for reporting on this target population.

Sample frame

The availability of a single and updated list or register varied between the four strata. The sample frames were, in some cases, compiled based on various sources. Below is a presentation of the different sources used for each of the target populations.

Albanian IDPs: a list of 1,167 households and 5,879 individuals were established. The source was Kosovo Agency of Statistics' (KAS) 2011 census. A set of questions in the census were identified that could be used to determine whether a person was an Albanian IDP and based on this an extraction of the Albanian IDPs was made. To verify this list, additional sources were cross checked and matched with the census extract: a list of Albanian IDPs who had filed claims for lost property with Kosovo Property Agency (mostly in relation to conflict-induced displacement), a list of Albanian IDPs who had directly approached DRC and Mitrovica North Administrative Office (MNAO) in Mitrovicë, as well as through Albanian IDP community leaders who had established contacts with UNHCR and DRC.

Serb IDPs in private accommodation: a complete list of 3,872 households and 16,383 individuals was available. The source was the Serbian Commissariat for Refugees and Migration's (SCRM) database, which had updated data about IDP household names, household sizes, addresses and contact details. The sample frame provided by SCRM was crosschecked with information from UNHCR.

Serb IDPs residing in collective centers: a complete list of 140 households was available. The sources were the Serbian Commissariat for Refugees and Migration's (SCRM) database and a UNHCR database consolidated.

Roma/Ashkali/ Egyptians (RAE) IDPs: a list of 137 households and 638 individuals was established by combining a UNHCR database and lists kept by other organisations that provided assistance to this group. The UNHCR database was, however, several years old and had not been updated. Mobility of these households was expected due to their economic vulnerability, which was expected to further decrease the accuracy of the sample frame.

Unit of sampling and analysis

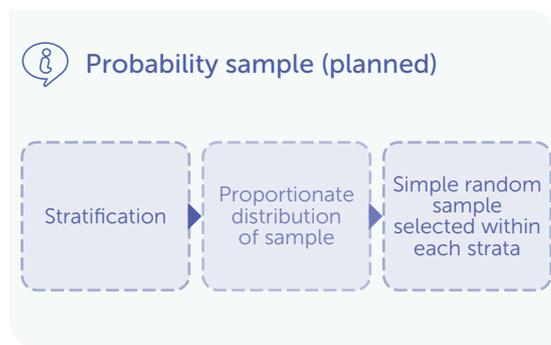| Sampling unit	Analysis unit
household	household and individual

J E T | JIPS
Essential
Toolkit

Sampling approach

The objectives of the exercise were to generate a better understanding of the characteristics and situations of the target populations, which demanded that the findings from the sample survey should be representative of each of the target populations. For this purpose, a probability sampling approach would be the preferred option. However, limitations in the availability of up-to-date sample frames made it necessary to use non-probability approaches for some of the target populations.

As a result, sample selection approaches were tailored to each of the four target populations and designed for the findings to be comparable across the population groups, not to provide a picture of the displaced population as a whole. Below is a description of the different approaches taken for each of the target populations.

Albanian IDPs

Stratified random sampling was initially planned for this target population. However, later in the process the profiling partners had to move away from this approach due to challenges described further below. As a last resort profiling partners made use of snowball sampling. The initially planned sampling approach can be summed up in the following way:

Probability sample (planned)

Stratification → Proportionate distribution of sample → Simple random sample selected within each strata

Step 1 : Stratification

The target population was first divided into three strata based on administrative areas:

(1) Prishtinë/Pristina municipality

(2) Mitrovica South municipality

(3) Vushtrri/Vucitrn municipality

Within each of the three strata the target population was further stratified into:

(1) Urban areas

(2) Rural areas

Step 2 : Deciding on the sample size

A sample size of 510 households was calculated based on a desired confidence level of 95%, a desired margin of error of 5 % and an expected non-response-rate of 15%.

Step 3 : Selecting the sample

The sample size was proportionately distributed among all the strata according to the size of the Albanian IDP population within each of the strata, building on the Kosovo Agency of Statistics' (KAS) 2011 census. A random sample was then drawn for each of the strata based on the census extraction.

After having initiated the fieldwork, and when approaching the households in the sample to conduct the interviews, it was found that in almost none of the addresses listed in the sample, Albanian IDP families were located. It was discovered that in some of the cases the households living at the addresses at the time of the census survey had moved to a different location. However, in most of the cases, the reason why IDP households were not to be found at the listed addresses remained simply unknown.

This resulted in an abandonment of the original sample plan, and profiling partners decided that they would have to go with a snowball sampling approach, but that they would still aim for the same sample size and distribution between strata as initially planned. The Albanian IDP households in the original sample that had actually been found at the listed addresses served as reference points and were asked to identify other Albanian IDP households.

Profiling partners were more or less able to reach the desired overall sample size, as 502 households in total were interviewed. The final distribution of the interviewed households across the three municipality strata, however, did not correspond completely with the planned distribution (as shown in the below table). Due to the non-probability approach, the representativity of the sample for the whole target population was uncertain.

Municipality	Planned distribution of HH	Final distribution of HH
Prishtinë/Pristina	114	45
Mitrovica South	315	435
Vushtrri/Vucitrn	81	12
Other	0	10
Total	510	502

Serb IDPs in private accommodation

Stratified random sampling was chosen for the sample selection of this target population. The approach taken can be summarized in the following way:

Step 1 : Stratification

The target population was stratified into two strata based on geographical location:

Probability sample

Stratification → Disproportionate distribution in two strata, proportionate distribution in the two others → Simple random sample selected in all strata

(1) Northern municipalities (which hosted the largest share of the Serb IDP population)

(2) Southern municipalities (which hosted the smallest share of the Serb IDP population)

Within the two strata listed above, the target population was further divided into two more strata:

(1) Urban areas

(2) Rural areas
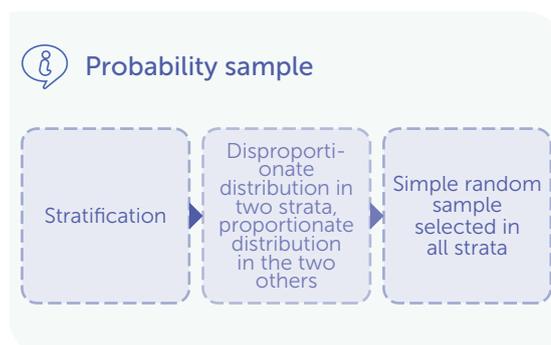
Step 2 : Deciding on the sample size

A sample size of 617 households was estimated based on a desired confidence level of 95%, a desired margin of error of 5 %, and an expected non-response-rate of 15%.

Step 3 : Selecting the sample

The sample size was allocated equally between the two strata of the northern and southern municipalities. The distribution made it necessary to use sampling weights in the analysis of the data, to account for the different Serb IDP population sizes within these two strata. These weights equalled the overall number of households that each of the sampled household in the two strata represented. The weights would be applied whenever results were presented at the overall level, i.e. when aggregating the findings of these two strata. When presenting findings of the strata separately, the weights would not be applied.
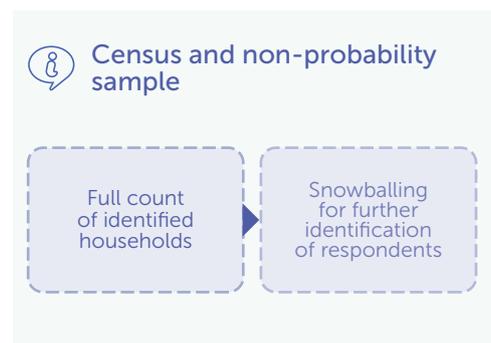
The sample size was distributed proportionately between the urban and rural strata, according to the size of the Serb IDP population living in each stratum within the northern and southern municipalities. The sample was drawn from the sample frame in all the strata, and interviews were carried out for the households in the sample. A total of 567 households responded to the survey, implying a high response-rate of 92 %.

Serbian IDPs in the Collective Centers

The small size of this target population (140 households) led to the decision of including all of the households in the survey, meaning conducting a census for this group. The list of persons in the target population was up to date, and 135 of the households completed the survey, corresponding to a high response-rate of 96 %.

Roma, Ashkali and Egyptian (RAE) IDPs

Like for the Serb IDPs living in collective centers, the profiling partners decided to do a full count of the RAE households that had been identified, given their small number. A total of 137 households had been identified, but the compiled list of members of the target population was somewhat outdated and it was thus expected that several of the households would not be possible to locate. Against this background, a snowball sampling approach was used to complement the list of households from the UNHCR database. The approach taken can be described in the following way:

Census and non-probability sample

Full count of identified households ▶ Snowballing for further identification of respondents

The households on the list were approached by enumerators and asked to complete the interview as well as to refer to other members of the target population that they knew. This approach resulted in 123 households or 491 individuals being interviewed. The surveyed households were located primarily in Mitrovicë/Mitrovica South (39 households) and Gračanica/Graçanicë (26 households) and, to a lesser extent in Obiliq/ Obilić (11 households) and Fushë Kosovë / Kosovo Polje (13 households). Through the snowball approach, enumerators had indications that they were close to reaching a saturation point, meaning that they had identified and interviewed all of the RAE IDPs in the specified municipalities. This led to the conclusion that the findings could be presented as representative of the whole target population.

Challenges and limitations

The original sapling approach had to be re-designed due to the lacking accuracy of the sampling frame provided by the census, making it necessary to apply a different approach. The non-probability sampling approach then chosen to identify Albanian IDPs for the sample did not reach the point of saturation (502 out of an estimated 1,167 households were interviewed), which meant that it was not possible to assess with certainty the representativity of the sample. Furthermore, the majority of the households in the sample were identified in Mitrovice/Mitrovica municipality, with a smaller number in Prishtine/Pristina. Because one expected that there had been some mobility of IDP households moving to Prishtine/Pristina over the last years, it was concluded that the sample identified through the snowballing only could be assumed to be representative for the Mitrovice/Mitrovica municipality.

There were limitations to the availability of secondary data on the general population in Kosovo. Three different sources were used: the 2011 Population and Housing Census, the UNICEF 2013-2014 Multiple Indicator Cluster Surveys, and the UNDP 2015 Mosaic. Comparing the findings from the IDP survey data conducted in 2016 with census data from 2011 was not ideal. The data being representative for two different points in time could make it problematic to compare findings from the two sources, because the relevant characteristics of the overall population in Kosovo might have changed significantly between 2011 and 2016.

Mogadishu profiling exercise, Somalia, 2015-2016

Objectives

Two decades of armed conflicts and severe recurring droughts and floods have forced a remarkable part of the Somali population to leave their homes. Mogadishu hosts the largest estimated protracted internally displaced population in the country, mainly living in informal settlements across the city and its outskirts.

The overall objective of the profiling exercise was to provide an evidence base to inform the planning for durable solutions for IDPs through joint humanitarian and development advocacy and response. Specifically, the exercise aimed to:

(1) Provide a **disaggregated estimate of the IDP population figures** living in settlements;

(2) Identify internal displacement through **analysing household migration history**;

(3) **Analyse the current situation of different target populations**, including their socio-economic situation, living conditions, access to basic services and protection concerns;

(4) **Understand the resilience of families**, defined as the skills, capacities, special needs and coping mechanisms of target populations that inform their decision making for the future.

Target population

Three target populations were included in the sample survey of the profiling exercise:

(1) **IDPs:** persons of Somali origin who had been internally displaced from their place of origin to or within Mogadishu primarily as a result of conflict, disaster, insecurity or rights violation(s).

(2) **Economic migrants:** persons of Somali origin who had left their place of origin and came to Mogadishu primarily to look for livelihoods and other economic reasons.

(3) **Host communities:** persons originating from Mogadishu who have never left their place of origin and live in informal settlements in Mogadishu.

The following target populations were also included in the profiling exercise, but not in the sample survey. Information on these groups were obtained through qualitative data collection approaches:

(4) Returnees: persons of Somali origin who fled Somalia to other countries primarily as a result of conflict or disaster and have returned to Somalia.

(5) Refugees: persons originating from other countries who fled to Somalia and Mogadishu to seek international protection.

Sample frame

No complete list identifying IDPs, economic migrants or host communities living in informal settlements in Mogadishu was available to profiling partners, which was also partly the background for this exercise. It was decided that sample frames would need to be established by conducting a mapping of the informal settlements and doing an enumeration of households living within those settlements.

Unit of sampling and unit of analysis

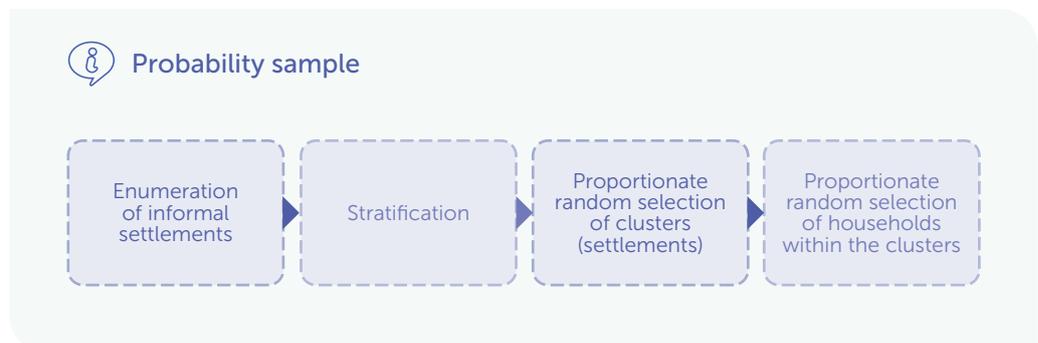| Primary unit of sampling	Secondary unit of sampling	Unit of analysis
Settlements	households	households and individuals
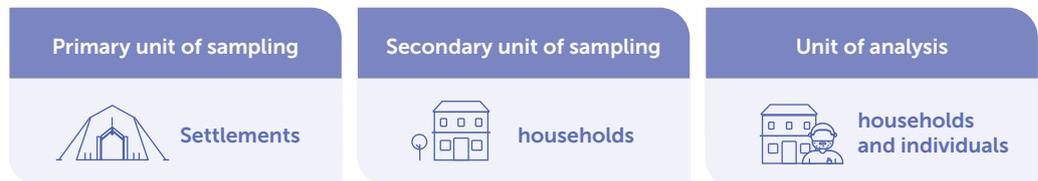
Sampling approach

According to the objectives of the exercise, profiling partners needed to tailor their sampling approach so that they would be able to provide an estimate of the overall number of IDPs living in settlements in Mogadishu, as well as information on the situations and characteristics of IDPs, economic migrants and host communities, that could be extrapolated to the whole of the three target populations. Profiling partners also wanted to make comparisons by between the three target populations – the IDPs, economic migrants and host communities – as well as by geographical location for the IDP target population. Below is a description of the approaches for each of the target populations.

IDPs

A two-stage stratified cluster sampling approach was chosen for this target population, where settlements would serve as clusters (i.e. primary sampling units), and households would serve as secondary sampling units. The approach can be summed up in the following way:

Probability sample

Enumeration of informal settlements ▶ Stratification ▶ Proportionate random selection of clusters (settlements) ▶ Proportionate random selection of households within the clusters

Step 1 : Mapping and enumeration

The profiling partners started out by doing a mapping of all the informal settlements in Mogadishu where IDPs where expected to be living. This desk review focused on identification of settlements in Mogadishu through the review of existing information on settlement names, GPS coordinates and population estimates. The mapping revealed many information gaps, especially in two of the districts (Daynille and Kaxda) that were expected/known to host the largest IDP population and had the highest number of settlements. As a consequence, it was decided that an enumeration exercise would be conducted to identify and count all persons living in the settlements. This would generate population estimates of IDPs living in the settlements (and fulfill objective **(1)**), as well as serve as sample frames for all three target populations for a sample survey.

The enumeration aimed towards conducting a full census of the households in all settlements identified in the mapping phase. In addition, profiling partners worked with district commissioners to identify additional existing settlements in each district to be enumerated. In the Daynille and Kaxda districts, Google Earth spatial imagery was also used to grid the area and to break it down to manageable enumeration areas.

Through the enumeration process, empty settlements were identified, so that they could be disregarded later on in the sampling procedure. For each enumerated household, the shelter structure was spray-painted using a unique code, to avoid households being enumerated twice and to facilitate the identification of targeted households for the survey sample. The questions asked in the enumeration aimed towards identifying members of the different target populations. The enumeration identified 68,796 IDP households, comprising 399,292 individuals.

Step 2 : Stratification

To meet the objectives of analysing the situations and different characteristics of the IDP households, a representative sample of this target population would have to be selected, so that findings could be assumed to hold for the whole target population. The stakeholders also wanted to be able to compare the findings for IDPs in the Daynille and Kaxda districts with the rest of the districts, as the first two were hosting the largest share of IDPs.

The target population was thus divided into three strata, based on geographical location of the IDP population:

(1) Daynille district (hosting 35 % of the IDP households in Mogadishu)

(2) Kaxda district (hosting 20% of the IDP households of Mogadishu)

(3) The rest of the districts of Mogadishu

Step 3 : Deciding on the sample size

The sample size was to be representative with a 95 % confidence level and with a margin of error of 5 % for the overall sample, while when broken down by strata, the margin of error would be 10 %. The sample size was set to 612 households.

Step 4 : Selecting a sample of settlements

Following the mapping and enumeration of settlements, as well as the stratification of geographical areas, a random selection of settlements was picked with probability proportionate to the size of the IDP population in each stratum.

Step ⑤ : Selecting a sample of households

Following the selection of settlements, households within the selected settlements were randomly drawn to take part in the final sample.

Economic migrants

The enumeration of settlements resulted in a database that could be used as a sample frame for this target population as well. Among the households that were identified as economic migrants during the enumeration, a simple random sample was drawn.

The sample was representative with a confidence level of 95 % and a margin of error of 5 %. The sample size was set to 638 households.

⑧ **Probability sample**

Enumeration of settlements ▶ Simple random sample of households
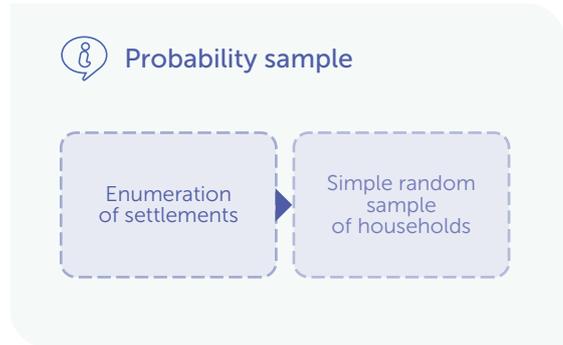
Host community

Like for the economic migrants, the results from the enumeration process could be used as a sample frame for the host community target population. Among the households that were identified as part of the host community, a simple random sample was selected.

The sample was representative with a confidence level of 95 % and a margin of error of 5 %. The sample size was set to 638 households.

⑧ **Probability sample**

Enumeration of settlements ▶ Simple random sample of households
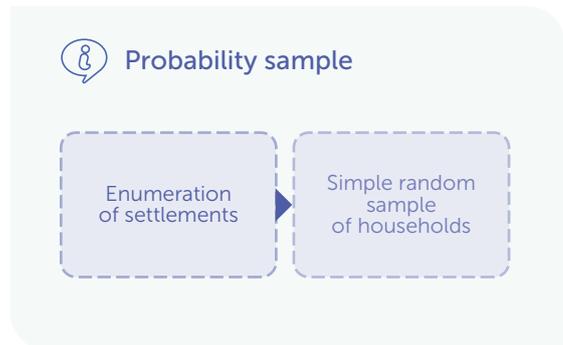
Challenges and limitations

The fluid situation in Mogadishu, notably on account of evictions of households from settlements, posed a significant challenge to the profiling partners. An acceleration of evictions in March and April 2015 required halting the exercise for that period. **To mitigate the effects of the rapidly changing situation, the time lapse between the enumeration and the household survey was minimised.** However, evictions continued to pose a challenge to the exercise throughout the data collection phase, although on a lesser scale. Nevertheless, given that evictions mainly resulted in intra-city displacement, i.e. from one IDP settlement to another, they had limited impact on the overall numbers and key challenges faced by the population in Mogadishu. As a result, this profiling could still provide a comprehensive analysis of the volume and trends of the displacement situation, as well as a thematic analysis to support planning for durable solutions.

Two districts – Heliwa and Abdulaziz – were excluded during the household survey for security reasons, and the results of the profiling cannot be generalised to these two districts. Both were included in the facility mapping and the enumeration area, however.

Sudan profiling exercise, 2017-2019

Objectives

Fifteen years after a violent conflict broke out in Darfur in 2003 the majority of IDPs in the region were still living in camps that resembled permanent settlements and continued to rely on humanitarian assistance. With the protracted displacement unsolved, finding durable solutions to displacement for IDPs and returning refugees had become ever more important.

The profiling exercise was carried out to address the need for comprehensive data giving an accurate picture of the situation faced by displaced people residing in camps on the outskirts of El Fasher town in North Darfur. The objectives of the exercise were the following:

(1) **Provide a comprehensive profile of IDPs** residing in Abu Shouk and El Salaam IDP camps as well as areas of neighbouring El Fasher city. The profile includes core demographic data, livelihoods and socio-economic information, insight into the needs of IDPs, their vulnerabilities, coping mechanisms and capacities. The profiling also takes into account IDPs' settlement intentions and plans as well as the push and pull factors that could trigger onward movement, return, or local integration.

(2) **Offer a comparative analysis** focusing on both displaced and non-displaced populations to identify IDPs' vulnerabilities as a consequence of their displacement. The analysis aimed to foster a deeper understanding of the obstacles and opportunities for reaching durable solutions.

(3) **Help inform programmatic responses** by the Government of Sudan and humanitarian and development actors and strengthen the resilience of their host communities within the IASC Framework.

(4) **Pilot a profiling exercise of displacement and joint durable solutions planning** with a view to replicate the model in other displacement settings in Sudan. As such, the pilot was intended to facilitate the development of agreed data collection tools and methodologies that can eventually be repeated with the necessary adaptations.

Target population

The target populations of the exercise were the following:

(1) The IDP population residing in the two camps of Abu Shouk and El Salam;

(2) The non-displaced population residing in peri-urban El Fasher (an urban area situated on the outskirts of El Fasher forming a 3 km radius immediately surrounding the two camps) and in urban El Fasher (the urban neighbourhoods of El Fasher city).

Sample frame

No register or other lists of the target population were available to the profiling partners. It was decided that a sample frame would need to be created in the process of selecting the sample.

Unit of sampling and unit of analysis

Primary unit of sampling	Secondary unit of sampling	Unit of analysis
Clusters	households	households and individuals
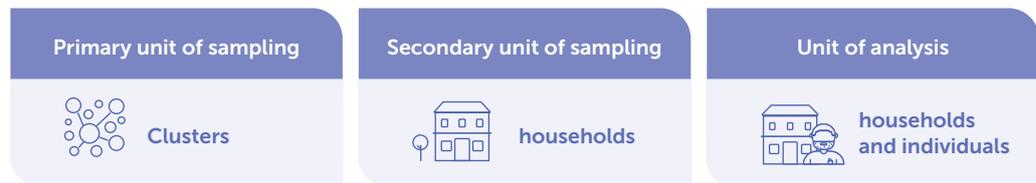
Sampling approach

The objectives of the exercise focused on providing analysis of IDPs' situations and characteristics compared to that of their non-displaced neighbors. It was also desirable to compare the IDPs of the two different camps and the non-displaced living in peri-urban and urban surroundings, because it was expected that these groups would face different situations related to their location. The sampling approach taken would thus need to produce a sample that was representative of all these sub-groups.

In the absence of an already existing sampling frame, profiling partners decided to conduct a two-stage stratified cluster sampling approach. The approach can be described in the following way:

Probability sample

Stratification ▶ Disproportionate random selection of clusters ▶ Enumeration of households in selected clusters ▶ Simple random selection of households in enumerated clusters

Step ① : Stratification

The target population was divided into four strata, according to the different groups that were to be compared:

- IDPs in Abu Shouk camp
- IDPs in El Salam camp
- Non-displaced population in peri-urban El Fasher
- Non-displaced population in urban El Fasher

The four strata were in turn all divided into smaller clusters that were assumed to have similar population sizes, based on a grid developed on a map of the relevant areas.

Step ② : Deciding on the sample size

Within each of the strata, a random sample of clusters would be selected and enumerated, followed by a selection of households for the final sample from the enumerated clusters. The sample size was set to 3,000 households which was expected to equal 252 clusters. The planned sample size of households and clusters were distributed across the strata in the following way:

Population type	Strata	Number of households	Number of clusters
IDP	Abu Shouk camp	996	84
	El Salam camp	996	84
Host Community	Neighboring Al Fashir	504	42
	Non-neighboring Al Fashir	504	42
Total		**3,000**	**525**

Step ③ : Selecting a sample of clusters

A simple random sample of clusters was drawn in each of the strata. There were no reliable population figures available for the clusters, so the clusters could not be selected with probability proportional to size. This meant that sampling weights would need to be applied when aggregating the findings of the different strata to an overall population mean.

Step ④ : Enumeration of selected clusters

As a next step, all the households in each of the selected clusters were enumerated. The listing exercise resulted in a **significantly lower number of enumerated clusters than what was initially planned. One reason for this is that some of the selected clusters were discovered to be inaccessible to the enumerators, mainly in the El Salam camp.** Additionally, clusters that hosted less than 12 households were dropped from the sample. Some of the enumerated households were also listed twice or were living outside of the cluster boundaries, and these were also dropped from the sample.

Step 5 : Selecting a sample of households

The enumeration of households in the selected clusters provided a sample frame for the final sample of households to be selected for interviews by simple random sampling. To select the sample, the enumeration data was restricted to IDP households in the two camp strata and to non-displaced households in the two El Fasher strata. The number of households to be selected was initially planned to be equally distributed across the selected clusters, with 12 households selected from each. However, due to the lower number of enumerated clusters, selecting 12 households per cluster for interviews would have led to a sample size falling significantly short of the planned 3,000 households. Thus, in each stratum some of the listed clusters were randomly selected for oversampling, to compensate for the lower number of clusters. The clusters selected for oversampling were to have 24 interviews in them. Only the clusters with more than 23 listed households were eligible to be selected for this oversampling.

Sampling weights were calculated and applied to clusters depending on their size and the number of households selected in each cluster (12 or 24). The resulting weights could not be tested against population data due to the lack of up-to-date and reliable information. However, for the camps, the population sizes suggested by the weights were close to the latest population estimates by the International Organization of Migration (IOM).

The below table shows the final and actual distribution of listed clusters, interviewed households and individuals included in the interviewed sample from each of the four strata:

Population type	Strata	Number of clusters listed	Number of households interviewed	Number of individuals in interviewed households
IDP	Abu Shouk camp	82	996	5,849
	El Salam camp	50	986	5,960
Host Community	Neighboring Al Fashir	40	509	3,303
	Non-neighboring Al Fashir	41	511	3,376
Total		**213**	**3,002**	**18,533**

Challenges and limitations

The sample in the two El Fasher strata (urban and peri-urban) only included non-displaced households. Therefore, it should be kept in mind that the results for these strata do not represent the situation for all residents living in the peri-urban and urban neighborhoods, given the presence of displaced households in these strata that were not included in the sample. Specifically, the enumeration indicated that approximately 28% of peri-urban and 21% of urban households were IDPs. The comparative analysis thus focused on the differences between the population groups by displacement status and cannot be used to compare the areas as such.

The objectives of the exercise, namely to inform area based and local integration programming would have benefited from an additional comparison by geographic area (city centre vs peri-urban vs camps). Such an approach would have required either a larger sample (to include two target groups within the peri-urban and the urban strata – namely both displaced and non-displaced), or the decrease of each sample drawn (in order not to add resources overall) – which would have led to less statistically powerful samples but could have added programmatic relevance to the analysis. These are considerations that partners are required to have at the outset and agree on what is more important.