Studying Millions of Rescued Documents: Sample Plan at the Guatemalan National Police Archive

Daniel Guzmán, Tamy Guberek, Gary Shapiro and Paul Zador*

Abstract

This paper describes the sample design used at the Guatemalan National Police Archive (GNPA). The Archive contains millions of documents, which were initially found mixed together and in poor physical condition. Given the Archive size and the lack of a traditional sample frame, we opted for a multi-stage random PPS sample using the Archive's topography for stages 1 and 2. For stages 3 and 4, frames were created on location. The sampling faced several challenges, including movement of the documents as they were being restored and organized, and uncertain resource availability. To manage these difficulties we drew iterative sample waves. After rounds of evaluation, we modified the sampling to reduce one stage, making the sampling more efficient. Over 2 years of sampling, we have selected 20,000 documents. Next, we may use adaptive sampling to search more deliberately and probabilistically for documents of interest.

Key Words: probability sampling, complex sample design, sampling frame, cluster sampling, human rights violations, Guatemalan National Police Archive

1. Introduction

In July 2005, an explosion at a military munitions dump near Guatemala City raised concerns about the storage of explosives in nearby residential areas. People who lived in the neighborhood asked investigators to inspect a building at the Guatemala City compound of the National Civil Police. A team from the state-backed Guatemalan Human Rights Ombudsman (PDH) entered the decaying structure and discovered an enormous cache of documents.

The records were stored in a series of dark rooms overrun by rats, bats and cockroaches. Many of the papers were soaked by rainwater from leaks and broken windows. The documents, which numbered at least in the tens of millions, were revealed to be the historic Archive of Guatemala's National Police. The National Police were disbanded after the country's 1996 Peace Accords and were replaced by the National Civil Police. The

*Daniel Guzmán and Tamy Guberek are consultants for the Human Rights Data Analysis Group at The Benetech Initiative www.hrdag.org, www.benetech.org. Paul Zador, PhD., is a Senior Statistician at Westat. Gary Shapiro, recently retired, was a Senior Statistician at Westat during the majority of his contribution to this study. Both Shapiro and Zador are members of the Volunteerism Special Interest Group of the American Statistical Association.

Guatemalan National Police Archive (GNPA) may well be the largest single cache of documents that has ever been made available to human rights investigators.

Since 2005, three simultaneous processes have been underway at the GNPA: 1) archival work, including the restoration, preservation and cataloging of the documents; 2) judicial case research; and 3) sampling and analysis. This paper describes the sample design for the third process, the quantitative study, for sample iterations, or waves, 1 through 9.

This paper is the first in a series of three papers. Here, we describe the sample scheme. The second paper presents the calculation of the sample weights. The third paper presents initial estimates regarding the contents of the National Police Archive.

This paper is structured as follows: Sections 2 - 5 explain the sample design and implementation. Sections 6 and 7 discuss various challenges and the sampling choices that were made to address them. Section 8 provides unweighted descriptive statistics on the data collected. Section 9 discusses the next steps we plan to take as the project evolves. Finally, Section 10 summarizes the paper and draws some general conclusions.

2. The Decision to Sample

Given the Archive's magnitude and poor physical condition, an in-depth study of its contents could take years, maybe decades. Yet, too many important questions needed to be answered sooner rather than later. What kind of information does this body of documents contain? Do these documents hold information about political violence during Guatemala's internal armed conflict? Will they reveal civilian repression, killings and forced disappearances that were carried out by the police to be institutional goals or just the acts of a few officers in the ranks? Probability sampling was an efficient and defensible way to start exploring these questions.

The initial sample design took into consideration three objectives:

1. To understand the scope and heterogeneity of this unexplored, massive Archive
2. To gather data about broad, macro patterns of police operations, such as command structures and communication flow.
3. To estimate the proportion of documents that recorded certain acts and policies of interest to the project, including disappearances, detentions and deaths.

Although the Archive contains documents over a century old, the time period of the study is limited to the years of historical interest due to the internal armed conflict -

between 1960 and 1996 inclusive. The design was intended to be broad rather than deep for an initial phase - waves 1 through 9. This meant we initially drew fewer documents from each sampled location in the Archive, but drew documents from more locations in the Archive. It also meant we focused our coding on the structural information of the sampled documents rather than the details about the content (ie: authors, recipients, dates, presence/absence of acts of violence, etc.) With the data and lessons learned during the process of sampling waves 1-9, incremental changes were made to the design to record more data on the documents' content. These changes and others will be described in Section 6.

3. Basic Sample Design

A multi-stage iterative probability design was used to obtain a sample of documents. Iterative samples were drawn to control for document movement. Probability proportional to size (PPS) sampling was done at each stage. PPS sampling was important because of the variation in the size of units at each stage. Size was measured either in linear meters or cubic meters (volume) of paper for all the stages.

We considered the following challenges as we determined our sample design.

- No direct sample frame of the documents An underlying enumeration of the population contents was not available at the Archive and creating one would be too costly in time and resources.
- Unclear time and resources When the project started, it was unclear how long the jurisdiction and political will of the PDH to organize and research the Archive would last. The project could end soon. Furthermore, the project had to raise funds to maintain itself, so resource availability was uncertain. Given that available time and resources were unclear, it was not realistic to base the sample on a predetermined sample size.
- Movement The sampling was taking place amidst other processes mentioned earlier: archival work and case research. These caused some documents to be moved as they were restored, organized and studied.
- Changing size The initial discovery of the Archive motivated other regional police archives to be revealed. As a result, new documents were brought in to the main Archive on several occasions, altering the total population size from which the sample was drawn.

All stages are presented in detail in the following section, see Table 3 for a summary of sample stages, unit of measurement within each stage, and source of sample frame for

each stage.

Table 1: Sample Stages

Stage	Measure	Sample Frame
Environment	Linear meters	Master List (RMU)
Container	Linear meters	Master List (RMU)
Last unit of aggregation (LUA)	Volume (cm^3)	3D coordinates
Information Unit (IU)	Linear millimeters	LUA's Height

4. Construction of the Sample Frame

While there was no enumerated sample frame that directly accounted for all the Archive documents, one of the first things the GNPA leadership did when they gained jurisdiction over the Archive was to identify and label all the spaces inside the premises that contained documents (The sample excluded anything that was not on paper. Side projects focusing on electronic media are ongoing.). They used a system of nested topography to ensure control and coverage of all the spaces.

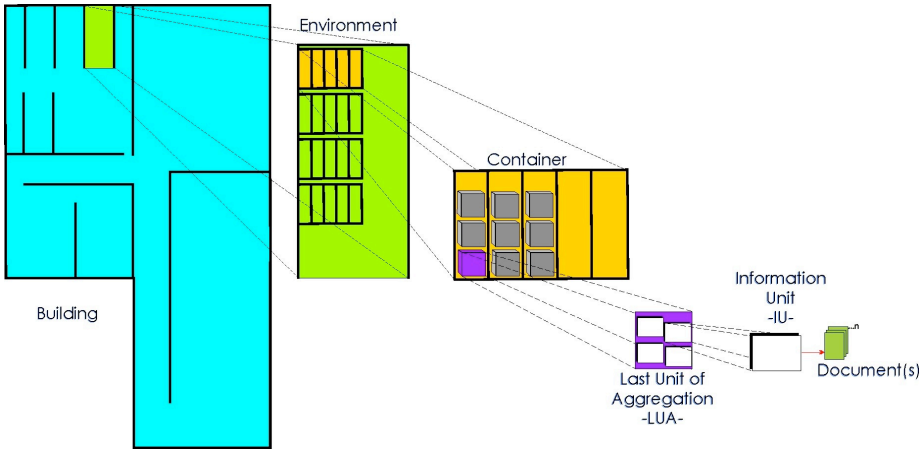
Each space in the Archive was assigned a unique identification. At the highest level, the topography started with the seven buildings that made up the Archive. Inside buildings, they identified "environments", which are often semi-closed off rooms. Once inside an environment, any container in which documents were stored was identified and uniquely labeled. Containers included bookshelves, filing cabinets, tabletops, sacks, wooden platforms and floor space. Then they measured the contents of each container using the archivist system of measurement, usually linear meters.

Figure 1 is a diagram of the nested topography in the Archive, which was used for sampling. It is useful to note that the final sampling stage is Information Unit (IU) which then contains one or multiple documents to be coded.

The topographic inventory was called the Location Master Register or RMU based on the Spanish acronym. RMU was considered the master list of linear meters of paper in containers and also recorded any movement of paper that occurred within the Archive area. There is a continuing need to keep the RMU up to date because the locations and quantities of the documents change frequently as the documents are cleaned, scanned, and archived.

The multi-stage design was able to capitalize on this metric, topographic inventory. It

Figure 1: Diagram of Archive: From Building to Document(s)



served as the primary sample frame for the first two stages of sampling - environments and containers. The most updated version of the RMU was used for each wave. Using the Archive's topography, the field team systematically arrived at a container inside the Archive from which to sample documents. They then created sample frames on location to select Last Units of Aggregation (LUAs) from inside containers.

The measured dimensions of containers were used as the sample frame for stage three, from where a grouping of pages was selected. These groupings, called LUAs, are defined as the smallest recognizable grouping of documents within a container. Some examples of LUAs are bundles of paper tied together with string, folders, cabinet drawers, bags and boxes.

The dimensions of a LUA were used as the sample frame for the fourth and last stage of sampling - called information units (IUs). IUs are defined as one document or a set of documents, which have been filed together as a single unit by the police unit of provenance and relate to a common theme, case or phenomenon. An information unit can be a document consisting of one page or a case file made up of many documents. Typically, an IU would be a group of papers stapled or paper-clipped together. From the documents that made up the sampled IU, we coded information of interest for the analytic objectives of the study. If a sampled IU did not contain any documents written between 1960 through 1996, we noted the IU as ineligible for the study and moved on to the next indicated sampling point.

Figure 2: Documents grouped in bundles within a variety of container types in one environment at the GNPA

5. Sampling Procedure

The four stages used to carry out the sampling are described in detail in this section.

5.1 Stage 1: Environments

In order to get a better sense of the heterogeneity of the contents of the Archive at a given location, and due to the unknown time restrictions of the survey, we decided early on to concentrate on few environments per wave. To calculate the number of environments to select, we used the information provided on the RMU inventory about the total number of linear meters per environment. We calculated an approximate sample size taking the variance of linear meters per environment into account. We chose to select 20 and 23 environments for waves 1 and 2, respectively.

We then noted the linear meters of paper per environment and created a cumulative proportion variable (p_i). All of the environments were assigned mutually exclusive intervals between 0 and 1, as follows:

$$(0, p_1],\ ...,\ (p_{i-1}, p_i],\ ..., (p_{k-1}, p_k],\ with\ p_k = 1, \tag{1}$$

Selection was then carried out without replacement with probability of selection based on these intervals until the specified number of environments had been selected.

Early on, we considered the trade off of using this method with the difficulty of later calculating the selection probability because the intervals were not of the same width. Therefore, as of wave 3, we changed the environment selection procedure.

As of waves 3 through 9, instead of drawing random numbers until a fixed number of environments were selected, we chose a fixed quantity of random numbers (33) drawn from a standard uniform distribution. As above, the selection of environments was determined by the interval (as defined in 1) within which each random number fell. Thirty-three random numbers were chosen with the expectation that this method would yield a similar number of unique environments in sample as those in waves 1 and 2 (where we selected 20 and 23 respectively). As expected, we selected either 22 or 23 unique environments for waves 3 through 9.

5.2 Stage 2: Containers

For each wave three hundred points were distributed among selected environments. Within the environment, containers were selected with known probability. Three hundred points per iteration were chosen based on tests conducted during the pilot study. Sampling documents from 300 points took approximately 2 to 3 weeks given the size of the team working. Based on the knowledge of our partners, a three-week time period was considered short enough to avoid major movement of paper between the sampling of stages.

These 300 points were distributed among the environments (E) selected in the previous stage, as follows:

Let *Ppe* indicate "Points per environment", where:

$$Ppe_i = 300 \times \frac{\text{linear meters } E_i}{\sum \text{linear meters } E_i} \tag{2}$$

Each point represents one container to be selected. At this stage, containers are selected with replacement, proportional to size, based on a combination of points per environment and size of container measured in linear meters. The number of times the same container was sampled determined the number of LUAs to be sampled from inside that container in the subsequent stage.

More equal weights would have been obtained if there had been the same number of sampled containers for each sampled environment, however we chose to use probability proportional to size due to the potentially drastically different sizes of containers. Many environments consisted of very few containers, each of which held large amounts of paper.

A list of selected containers and the number of LUAs to be drawn from each was given to the field team at the Archive. In order to select the sample at the last two stages, the field team had to locate the environments and containers selected. Once they located the sampled containers, the last two stages of selection were completed.

Table 2 is an illustrative snapshot of a list given to the field team after stages one and two were selected. This kind of list was the field team's 'map' to guide them to the location where they should sample LUAs from containers.

Table 2: Snapshot of some containers and the number of LUAs to be drawn from each by the field team

Building No.	Environment No.	Container Type	Container No.	Number of Points to Select
3	10	Platform	11	4
2	2	Filing cabinet	49	1
2	1	Table	2	2
4	1	Bookcase	7	1
4	1	Bookcase	22	2
1	2	Floor	1	2
2	2	Bookcase	63	2
6	1	Floor	1	23

As this point, it is useful to highlight that the sampling of stages one and two occurred at a different moment in time than stages three and four. The movement of papers that occurred during the time lapse was a major challenge and will be the subject of discussion in Section 6.

5.3 Stage 3: Last Units of Aggregation (LUAs)

Containers often hold documents in boxes, drawers, folders or tied bundles, which are called "last units of aggregation." With the list of containers and number of LUAs to draw from each, the field team arrived at their sampling location and began the selection procedure for stage three.

Given that there was no enumerated sample frame available for the contents of each container, the field team measured the extreme dimensions of paper inside a container - its height, width and depth - and used these measurements as the sample frame. In order to draw the sample, the field team members multiplied each dimension by a uniform random number to determine a random coordinate inside the container. This random coordinate would either be located inside a last unit of aggregation or in empty space. If the field team found empty space at the coordinate, they noted this as an "empty hit", and repeated the process until a LUA was found. The empty hits were later used to approximate the total amount of empty space inside a container, and will be mentioned in more detail in Shapiro et al. [1], the second paper in the series on weight calculations.

In cases where the container held only boxes, a different sampling method was used. Here, a box was selected randomly based on the total number of boxes in a container. Each box had an equal probability of selection.

The size of a LUA is its volume, measured in cubic centimeters, not linear meters of paper as used in the first two stages of selection. The field team drew the number of LUAs required by the point distribution in the previous stage. For each wave, 300 LUAs were selected from three-dimensional space inside containers, as described above in Section 3.

An advantage of sampling according to three-dimensional space was that we avoided handling the documents more than absolutely necessary. The only time the fragile documents were touched by members of the sampling team was when they arrived at the point from which to draw out a LUA and when drawing the information units from inside the LUA - stage four.

5.4 Stage 4: Information Units (IU)

An "information unit" (IU) is defined as a set of documents which have been filed together by the owners of the original filing system and relate to a common theme, case or phenomenon. As an example, an information unit may be a single document or a set of documents making up a case file. At the most aggregate level, the information unit is the unit ultimately sampled and studied at the GNPA. The reason we used information unit as the last stage of selection rather than documents is to respect the contents of the Archive as they were filed.

Information units are grouped together inside LUAs. The selection of IUs was based on the height of a LUA, measured in linear millimeters. The field team multiplied the height of the sampled LUA by a random number to determine a millimeter point from

which to select an IU. They drew the IU at the selected millimeter point plus the next two consecutive IUs and coded all three of them. By drawing three, we believed it was possible to gauge the similarity or difference of IU content within a grouping of documents. We did not want to select more than three, since a large cluster size would increase the sampling error. The total sample size of IUs per wave is 300 $points \times 3\ UIs = 900\ IUs$. Note that with this procedure it is impossible to know the precise probability of selection for each IU, since the probability is a function of the size of preceding IUs for which we usually have no information. Shapiro et al. [1] describes the challenges of estimating the selection probability for all three IUs. This was corrected in wave 11.

6. Challenge: Movement

By selecting small samples every 2 to 3 weeks, document movement during a wave was initially minimal. As a result, the probability of selection during this time was sufficiently stable.

Five months after the start of the iterative samples, the document movement was exacerbated by a few other events. This made the transition from stage two to stage three increasingly difficult.

Documents were not moved in a consistent way: Sometimes all the documents were moved from one container to another; other times documents in one container were split into several other containers; less often, several containers were consolidated into one or mixed with other documents not in the originally-sampled container. The staff at the Archive made an effort to track the papers, but movement was often hard to control. Table 2 shows that approximately 80% of the containers either had no movement or movement from one container to one other container, and therefore did not present difficulties for calculating selection probabilities. The remaining 20% were more challenging.

Table 3: Container movement table

From	To	Freq	Pct
No movement		772	65
1	1	189	16
Many	1	102	9
1	Many	61	5
Many	Many	9	1
Unknown		48	4

Shapiro et al. [1] describes how the increased movement of paper complicated calculating weights and the decisions we took to minimize the negative impact.

At the point in time when the movement was no longer considered manageable, we intervened and modified the sample design to eliminate dependency on the original inventory of environments and containers at the beginning of each wave. These changes are described in more detail in Section 7.

7. Evaluation and Changes

After the ninth wave, the changes in location, contents and size of containers that occurred during the time lapse between selecting containers and selecting LUAs threatened to make the calculation of selection probabilities too hard. We paused the process for approximately one month to make several improvements in the sample design.

For wave ten, the sample design was modified to only three sample stages. To start, we calculated the entire volume of the physical Archive (the volume of the buildings, not the paper). From the total volume of the Archive buildings, we approximated a proportion of 'occupied space', where:

$$\text{occupied space} = \frac{m^3 \text{ paper}}{m^3 \text{ total archive}} \tag{3}$$

We assumed that 3D coordinate in occupied space contained LUAs. We knew our target number of LUAs to sample was 300 (since we had sampled 300 LUAs in each wave 1 through 9). In order to sample this number of LUAs, we had to over-sample 3D coordinates (n), since only a proportion of the Archive had occupied space.

$$n = \left\lfloor \frac{300}{\text{occupied space}} \right\rfloor \tag{4}$$

As in waves 1 through 9, it was difficult to get at LUAs directly. First we had to allocate the n 3D coordinates in the Archive's environments. We used the same mechanism described in Section 5.1, with the only difference being that instead of using 33 random numbers, we used n random numbers.

n_j is the fraction of all 3D coordinates n that will be drawn from any environment j. So:

$$n = \sum_j n_j \tag{5}$$

Many 3D coordinates will point to 'empty' space, but the number of empty hits will be part of the data used to calculate the selection probability of the LUAs.

This modification in design effectively overcame the challenges posed by movement because we no longer depended on the Archive's contents remaining in the same place between the stages of sampling that occurred with a time lapse. By eliminating a stage, the design was consequently more efficient. Furthermore, the probabilities of selection should be much more uniform than in the original design.

In addition to changing the sampling design for selecting a LUA, the sampling procedure for IUs was also changed. We chose to select 10 IUs from each LUA, rather than 3 as in the previous waves. Furthermore, the selected IUs were not all consecutive. The field team selected a point inside a LUA as before, and from that point they drew 6 consecutive IUs. Then, they divided the remaining height of the LUA into four equal parts, drawing an IU from each fourth. This provides the opportunity to look more rigorously at the question of homogeneity/heterogeneity inside an entire LUA, not just consecutive IUs.

8. Status of Sampling

As of April 2009, 20,000 documents have been sampled. Approximately 8,000 were in sample for waves 1-9 using the sample design described in this paper. The other 12,000 were collected during a much longer wave 10 using the subsequent design described in Section 7.

The tables below present descriptive statistics about the documents sampled in waves 1-9. Table 4 lists the total number of information units (IUs) and documents sampled that were eligible for the study. An IU was considered eligible if at least one document within it was written between 1960-1996. Documents *related* to the period of study includes all the documents in the eligible IUs, although some may not have been written in the time period of the study. All these data will be used to calculate weighted estimates in the third paper of this series.

Table 4: General Summary of the Sampled Data

Category	Frequency
Information Units	5371
Documents *related* to 1960-1996	8162
Documents written between 1960-1996	7241
Documents with acts	1806
Documents with policies	90
Number of acts mentioned	3699
Number of policies mentioned	310

Figure 3 plots the year of creation of all the documents in sample. There is a gradual increase in document production after the mid 70's. Document year will be used to analyze when documents of certain subjects, such as acts of interest (listed below) were created and whether they have a correlation with periods of violence in Guatemala.

Figure 3: Sampled documents by year written

Table 5 lists the "acts of interest" for the quantitative research about human rights violations at the Archive. As Table 5 shows, *Deprivation of Liberty* and *Deaths* are by the far most commonly listed types of acts of interest and their frequencies surpass the the frequencies of *Interrogation* and *Surveillance* by several orders of magnitude. To what extent frequency differences in Table 5 reflect differences between (1) the documentation process or (2) the actual frequencies of the acts themselves is unclear. In the future, we

will try using adaptive sampling to target documents for infrequent act types of interest.

Table 5: Acts of interest to the GNPA research team

Acts	Frequency	Percent of all acts
Deprivation of liberty	2375	64
Deaths	596	16
Denunciations	203	5
Disappearances	179	5
Intimidation	138	4
Physical abuse	93	3
Proof of person	39	1
Entering private property	28	1
Sexual abuse	27	1
Interrogation	13	< 1
Surveillance	8	< 1

9. Next Steps

Four years have passed since the discovery of the Archive. Large portions of it have been cleaned and similar types of documents have been organized and stored together. In the future, we plan to develop and use adaptive sampling or other sample techniques to search more deliberately and probabilistically for documents of interest. One option might be to sample by document type. For example, we may be able to sample books with correspondence registers, which can provide a baseline of communication flow that we can then compare with specific document types such as confidential documents. Another option could be to sample identity cards that contain information that the police collected on individuals, often an indication of surveillance and social control. At the conclusion of wave 11, we will pause and evaluate the road ahead.

10. Conclusion

Although the sampling was challenging and resource-intensive, the GNPA has been committed to conducting scientific and rigorous research of the Archive documents since they began their work. As shown above, the documents sampled contain important information related to our research questions about subjects of interest such as acts of violence and administrative policies. After presenting the weight calculations based on our sample plan

in the second paper of this series of three, the third paper will present estimates directly related to the research questions motivating the study.

Acknowledgements

The bulk of the work on the ground, including data collection, measurement and data entry has been carried out under the leadership of Jorge Villagran and his team at the GNPA. Professor David Banks, PhD., also offered his invaluable advice during the initial pilot phase. The sample design was reviewed by Fritz Scheuren, PhD., VP Statistics NORC, University of Chicago during the early stages of the project.

References

[1] Shapiro, G., Guzmán, D. and Zador, P. and Guberek, T. (2009), "Weighting for the Guatemalan National Police Archive Sample: Unusual Challenges and Problems," in *JSM Proceedings*, Section on Survey Research Methods. Alexandria, VA: American Statistical Association.