Weighting for the Guatemalan National Police Archive Sample: Unusual Challenges and Problems

Gary Shapiro, Daniel Guzmán, Paul Zador, Tamy Guberek, Megan Price and Kristian Lum*

Abstract

This paper describes the weighting procedures used for the sample of records from the Guatemalan National Police Archive. The decisions made about sampling procedures to address the structure of the Archive resulted in a number of special problems when calculating the weights. First and foremost, the universe from which samples were selected is highly unusual and fluid, with sparse knowledge of measures of size. At more than one stage of selection, probabilities of selection were not known, requiring weights to be based on estimates of the probabilities of selection. There were difficulties stemming from operational issues, such as the movement of documents while the survey was conducted and empty space within containers.

Key Words: probability sampling, weights, post-stratification, weight trimming, Guatemalan National Police Archive, human rights

1. Background and Introduction

This paper, the second in a series of three papers, describes the weighting for the sample of documents in the first nine waves of sampling at the Guatemalan National Police Archive (GNPA). The documents were haphazardly stored in warehouses, presenting challenges in both selecting the sample and in subsequent weight calculations to appropriately account for sample design decisions. As described in the first paper [1], the documents are potentially useful to learn about the National Police's role in the violence during the internal armed conflict in Guatemala from 1960-1996.

The following section of this paper will present the weight calculations for each of the four stages of sampling described in Guzmán et al [1]: environment, container, last unit of aggregation (LUA), and information unit (IU). Analyses presented in the following paper in this series [2] are conducted at the document level (IUs may contain multiple documents). However, since documents were selected with certainty once an IU was selected,

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

IUs are the last stage at which weights are calculated. Section 2 discusses special problems in determining environment weights in waves 1 and 2, and the LUA and IU base weights. Section 3 discusses the additional special problem created by movement of documents during the sampling process. Section 4 discusses an additional weighting step that could be implemented in the future. Section 5 discusses data quality and imputation. Finally, in Section 6, we briefly summarize what has happened since the end of wave 9. This paper particularly focuses on the special weighting problems as discussed in Sections 2 and 3.

A traditional Horvitz-Thompson weighting approach was used for the survey, with base weights equal to the inverse of the probability of selection or, in some cases, to estimates of the probability of selection. The measure of size is height, in linear meters, for all stages of selection except the third stage (LUA) where volume was used.

There were an initial nine waves of sample selection. For simplification, we will discuss the weighting for a single wave, with comments about combining across waves in Section 2.5

2. Basic Weights

The weighting approach for the survey is the traditional approach (described in all survey sampling textbooks), with base weights being equal to the inverse of the probability of selection, and appropriate post-stratification adjustments applied to the base weights. As described by Guzman et al. [1], the measure of size for the first 2 stages was height, the measure of size for the third stage of selection was volume and the measure of size for the last stage of selection was height.

2.1 Environment Weights

For waves 3-9, the weighting was straightforward. Using 33 random numbers (as explained in [1]) we selected either 22 or 23 environments per wave, with probability proportional to the linear meters of paper within the environment, with replacement. Since the probability of not selecting E_i, the ith environment was $(1-p_i)$, this resulted in the following probability of selection for E_i:

$$Pr(E_i) = 1 - (1 - p_i)^{33} \tag{1}$$

where:

E_i is the environment i, and $p_i = \frac{\text{Linear meters } E_i}{\sum \text{Linear meters } E_i}$.

Linear meters E_i is the measure of size for E_i. Therefore, the environment base weight for E_i is $W(E_i) = 1/Pr(E_i)$.

For waves 1 and 2, we also sampled with replacement, but the number of unique environments in sample was fixed rather than the number of random numbers used to select the sampled environments. For wave 1, we used 25 random numbers to obtain 20 unique selected environments. For wave 2, we set the number of unique environments at 23, but had to select a sample of 58 random numbers to obtain the 23 unique environments. This sampling methodology made it difficult to calculate weights directly. A Monte Carlo simulation was used, repeating sample selection 10,000 times. Based on how many times the environments in our samples were in the simulated samples, we computed the estimated environment selection probabilities. The environment base weight for these waves is also the inverse of the probability of selection.

2.2 Container Weights

A container may consist of a bookshelf, filing cabinet, tabletop, sack, wooden platform, floor space, or other object located in an environment. For a given environment in a given wave, a fixed number of points were sampled as explained in Guzmán et al., [1]. Each point represents one container to be selected, and the same container can be selected more than once. The number of containers selected per environment was determined according to the size of the environment. The optimal sample design would have been equal sample sizes for each selected environment rather than sample size proportional to the measure of size of the environment. That is, we would have obtained less variable weights if there had been the same number of sampled containers for each sampled environment. Using PPS to select environments should have solved the problem of drastically different sizes of containers, however, in the early stages of sampling we remained concerned that PPS at the first stage was not adequately accounting for the differences in sizes of containers and therefore also selected containers according to PPS.

To calculate the conditional probability of selecting container j in environment i, we used:

$$Pr(C_{ij}) = \text{Points}_i \frac{\text{Linear meters } C_{ij}}{\sum_j \text{Linear meters } C_{ij}} \tag{2}$$

where:

C_{ij} = Container j in Environment i,

$Points_i$ = Number of selected points in Environment i, and

Linear mts C_{ij} is the number of linear meters in container C_{ij}.

The conditional container weight is then $W(C_{ij}) = \frac{1}{Pr(C_{ij})}$.

2.3 Last Unit of Aggregation (LUA) Weights

There were special problems in determining weights for the Last Unit of Aggregation (LUA), due to our inability to know or reliably estimate the probability of selection. A LUA is a group of documents that have been stored together in the archive. Examples of LUAs are folders, drawers, bundles of documents tied together, boxes of documents filed together, plus many others. A single LUA was drawn per selected container. However, since containers were selected with replacement in the previous step, it is possible for multiple LUAs to be drawn independently from the same container.

Sampling of LUAs was done based on random coordinates in three-dimensional space inside containers, as defined in Guzmán et al [1]. Thus, this sampling was based on volume, unlike the sampling for the first two stages, which were based on linear meters. As mentioned in Guzmán et al, these random coordinates could correspond to a hit in occupied or empty space inside the container. Thus, for a container that was selected once, a LUA may be selected after one or many hits (if the first few hits were in empty space). The true conditional probability of selection of a LUA within a container is a direct function of the proportion of space that it occupies. However, we were not able to directly estimate this proportion. The number of hits required to select occupied space provides an estimate of this proportion, but when there is only a single or small number of LUAs to be sampled, the estimate is quite poor.

To overcome this lack of information about the proportion of occupied space in each container, we used a Bayesian hierarchical model to borrow information across the containers. We assume that the number of empty hits for the ith container, empty_i, comes from a Negative Binomial distribution, NB(empty_i; required_i, p_i), where required_i is the number of LUAs required from the ith container before we can stop drawing, and p_i is the unknown probability of selecting a LUA from container i. We also assume that all of the p_i come from a common Beta distribution, with parameters α and β. Conditional on α, and β, each of the p_i is distributed Beta(α + required_i, β + $empty_i$) with mean $\frac{required_i+\alpha}{required_i+\alpha+empty_i+\beta}$. Notice that the mean of each of the p_is contains common parameters α, and β, thus sharing information across containers. To finish the model specification, we place vague Gamma priors on α, and β and run a simple Gibbs Sampler to sample from the distribution of the p_i, α, and β. We use the posterior mean of each p_i as the probability of drawing a LUA from the ith container, from which we can easily calculate the expected

number of empty hits that will occur before we draw the required number of LUAs.

The procedure results in weights that are only indirectly related to the actual probability of selection of a LUA.

The conditional probability of selection for a LUA is:

$$Pr(L_{ijk}) = \frac{V_{ijk}H_{ij}}{V_{ij}S_{ij}} \tag{3}$$

where:

V_{ij} = Volume of container C_{ij}, including empty space,

V_{ijk} = Volume of LUA k in container C_{ij},

S_{ij} = Empirical Bayes estimate of the proportion of occupied space for container C_{ij}, and

H_{ij} = number of times C_{ij} is selected in the second sample stage.

The conditional LUA weight is then $W(P_{ijk}) = \frac{1}{Pr(L_{ijk})}$.

Unfortunately, examination of initial final weights revealed that this procedure still inadequately estimated the probability of selecting a LUA. This was determined by comparing the estimated total linear meters of paper to the known total linear meters of paper in the entire Archive, based on the initial inventory. The estimated linear meters for the entire Archive was calculated from cumulative linear meters based on LUA weights. This comparison showed that we were consistently overestimating linear meters, which implied that we had not yet adequately accounted for and excluded empty space. Additional weight adjustments based on these findings will be discussed in Section 2.6.

Note that since the weights for LUAs are in terms of volume whereas the weights for environments and containers are in terms of linear meters, the desired cancellation of terms when conditional weights are multiplied together does not happen. There is more variation in weights when the cancellation does not occur.

There are special cases where LUA weights were calculated more directly. Given the necessity of the GNPA project to clean and organize the documents, a subset of the Archive has been gradually moved to equally-sized archival boxes. If a selected container only

contained these boxes, the probability of selecting a LUA (in this case a box) was calculated as one over the total number of LUAs (boxes) in that container.

2.4 Information Unit (IU) Weights

Determining the probability of selection for Information Units (IUs) was also difficult. An IU is 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. An IU may be a single document or a set of documents making up a case file. Sampling was done based on the height of a sampled LUA. A random point in this single dimension was identified by multiplying a random number by the height (in linear millimeters) of the LUA. The IU located at that point was sampled. The next 2 consecutive IUs were then also included in the sample. Although the probability of selection of the specific group of 3 IUs is clear, the probability of selection of the individual IUs is more challenging to determine.

Consider the following sorting of IUs in a LUA: Y, Z, A, B, and C, where A is the initially selected IU in the particular sample, and thus A, B, and C are brought into the sample. Y and Z are the IUs preceding A in the LUA.

Pr (C in sample) = Pr (A selected) + Pr (B selected) + Pr (C selected), where Pr(A selected) is the probability of an initial selection of IU A, Pr(B selected) is the probability of an initial selection of IU B, and Pr(C selected) is the probability of an initial selection of IU C. This is easily calculated, since we know the heights of A, B, and C.

However, the probability of selection for that first IU is a function of not only its height but also the height of the preceding 2 IUs, since they could have also brought A into the sample.

Pr(A in sample) = Pr(Y selected) + Pr(Z selected) + Pr(A selected)

Pr(B in sample) = Pr(Z selected) + Pr(A selected) + Pr (B selected)

Although the probability of selecting A, B and C is clear, there was unfortunately no information obtained on the height of the preceding Y and Z. We estimated the height of Y and Z as the average height of the 3 IUs (A, B, and C) that were in the sample. The probability of selection of Y and Z can then be easily calculated, so that the probabilities of being in the sample for A, B and C can be calculated (as above).

To justify this substitution, the average number of pages was calculated for each IU selected first, second and third across the sample. Since an analysis of variance concluded that these three means were not statistically significantly different from each other, we felt

comfortable substituting the average height of the selected IUs for the unknown heights of Y and Z.

The conditional probabilities for IU_{ijkr} were calculated as:

$$Pr(\text{IU}_{ijk1}) = \frac{\#\text{ pages}(\text{IU}_{ijk1})}{\#\ \widehat{\text{pages}}(\text{LUA}_{ijk})} + 2\frac{\sum_{r=1}^{3}\#\text{ pages}(\text{IU}_{ijkr})/3}{\#\ \widehat{\text{pages}}(\text{LUA}_{ijk})} \quad (4)$$

$$Pr(\text{IU}_{ijk2}) = \frac{\#\text{ pages}(\text{IU}_{ijk1})}{\#\ \widehat{\text{pages}}(\text{LUA}_{ijk})} + \frac{\#\text{ pages}(\text{IU}_{ijk2})}{\#\ \widehat{\text{pages}}(\text{LUA}_{ijk})} + \frac{\sum_{r=1}^{3}\#\text{ pages}(\text{IU}_{ijkr})/3}{\#\ \widehat{\text{pages}}(\text{LUA}_{ijk})} \quad (5)$$

$$Pr(\text{IU}_{ijk3}) = \frac{\#\text{ pages}(\text{IU}_{ijk1})}{\#\ \widehat{\text{pages}}(\text{LUA}_{ijk})} + \frac{\#\text{ pages}(\text{IU}_{ijk2})}{\#\ \widehat{\text{pages}}(\text{LUA}_{ijk})} + \frac{\#\text{ pages}(\text{IU}_{ijk3})}{\#\ \widehat{\text{pages}}(\text{LUA}_{ijk})} \quad (6)$$

Where i indexes environments, j indexes containers, k indexes LUAs and r indexes IUs.

The expected number of pages per LUA, $\#\ \widehat{\text{pages}}(\text{LUA}_{ijk}) = 11.17 \times \text{Height}(\text{LUA}_{ijk})$ was based on archivist knowledge and empirical data from our sample. Height was measured in millimeters.

The conditional IU weight is then $W(\text{IU}_{ijkr}) = \frac{1}{Pr(\text{IU}_{ijkr})}$.

2.5 Combining Weights Across Waves and the Final Weight

The final weight, except for trimming and post-stratification adjustment, for a single wave is the product of each of the weights in the preceding sections divided by the number of waves:

$$W_{ijkr} = \frac{W(E_i) \times W(C_{ij}) \times W(\text{LUA}_{ijk}) \times W(\text{IU}_{ijkr})}{9} \quad (7)$$

Since each wave was designed to represent the entire population of the Archive, combining weights across waves required division by nine.

2.6 Post-Stratification

As mentioned in Section 2.3, additional weight adjustment was considered necessary. To determine this adjustment, we used known population values to compare with the estimated values. The most reliable known population value was the total linear meters of paper in the Archive per wave. We applied a one-cell post-stratification adjustment using

the ratio of the estimated linear meters (as described in Section 2.3) to the known linear meters across all environments as the adjustment factor. This is a somewhat unconventional adjustment factor - traditional post-stratification adjustment involves sampled elements themselves (e.g., number of LUAs) rather than attributes of sampled elements (e.g., linear meters of paper). However, in the absence of population-level information on the number of LUAs, linear meters were considered a reasonable proxy. This adjustment was carried out for weights at the LUA level (3rd stage) because we did not measure IUs or documents in terms of linear meters and therefore could not make this weight adjustment at a later stage.

Table 1 shows an example of how the adjustment was done. The first column in the table is a unique id for each LUA. The linear meters of a LUA are equivalent to the height of the LUA, as presented in the second column of the table (LM). The third column has the basic weight (Wt) for each LUA. The fourth column (LM*Wt) shows the weighted linear meters for each LUA. Therefore, the sum of this column would be the estimate of the total linear meters of paper in the Archive (if there were only ten LUAs in the entire archive). Then, we take the known population value of the total linear meters in the Archive (Pop LM), which is 1100 linear meters in this example. With this known value we created a factor to correct the weights as noted in the fifth column (Pop LM/Estimated Total Linear Meters). This value represents the adjusted weights (basic weight * the factor). Finally, in the last column, we calculate the weighted linear meters by LUA using the adjusted weights. The sum of the last column is the estimate of the total linear meters using the adjusted weights. This estimate should equal the known total linear meters of paper in the Archive.

Table 1: Post-Stratification Adjustment

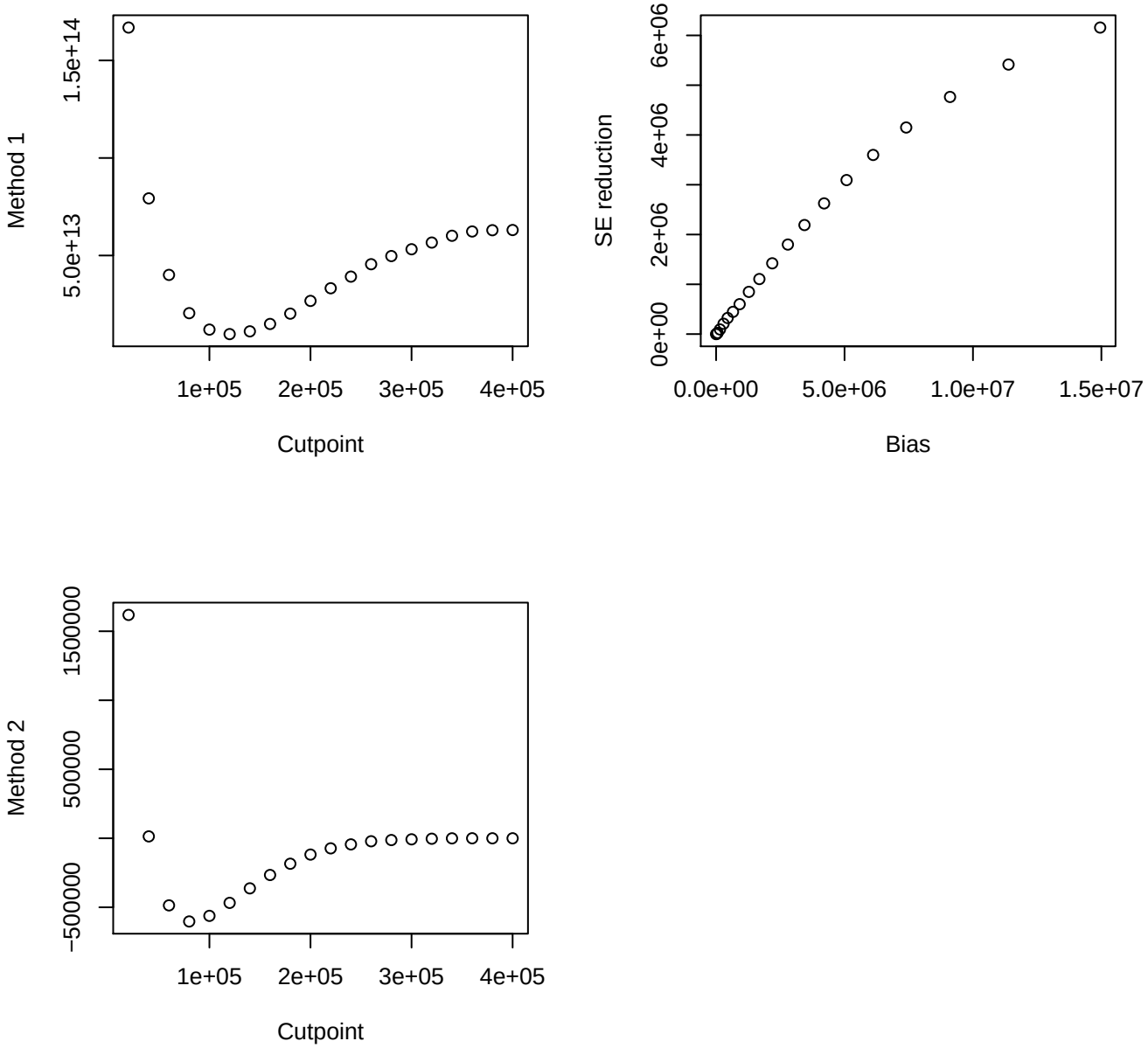
LUA	Linear Meters (LM)	Weight (Wt)	Weighted LM (LM*Wt)	Adjusted Wt (Wt* $\frac{\text{Pop LM}}{\text{Est. LM}}$)	Adjusted LM Estimates (Adjusted Wt*LM)
1	0.52	423	219.96	266.80	138.74
2	0.30	500	150.0	315.36	94.61
3	0.40	315	126.0	198.67	79.47
4	0.10	817	81.7	515.29	51.53
5	0.32	511	163.5	322.29	103.13
6	0.48	482	231.4	304.00	145.92
7	0.79	230	181.7	145.06	114.60
8	0.61	270	164.7	170.29	103.88
9	0.50	352	176.0	222.01	111.01
10	0.57	437	249.1	275.62	157.10
Estimated Total LM			1744.06		
Total Adjusted LM					1100

2.7 Trimming of Large Weights

It is known that in complex surveys unplanned large weights may occur. Given our sampling design we expected some unequal sampling weights. Nevertheless large variation in sampling weights can point to other problems, such as errors in the sample frame or coding process, or from adjustment procedures. As Potter warns, "[t]his unplanned or extreme variation in sampling weights can result in inflated sampling variances and a few extreme weights can offset the precision gains from an otherwise well- designed and executed survey design."[3]

Unfortunately, our initial weight calculations revealed extreme variation - the maximum estimated weight was more than 460 times larger than the median weight. We are continuing to review the weighting output to understand why there was so much variation. The weight calculations depended on several measurements that the field team had to take at stage three, during their work on the ground. Taking these measurements was often difficult, given the physical conditions of the Archive. The fourth stage also had large variation but less than the third. We tried to correct for bad measurements using knowledge from the field team. One option would have been to trim large weights at each stage, but in theory large weights at one stage could have been compensated by very small weights in the next stage. We decided to use visual exploration to choose the cut point for large cumulative weights, taking into account the trade off between bias and variance reduction (see Figure 1). A cut point was chosen (according to the method described below) that resulted in 56 records (0.7%) being trimmed. Post-trimming, the ratio of the maximum estimated weight to median weight was 115. The estimate of the total number of documents remained constant pre- and post-trimming.

Figure 1: Visual Exploration of Weights to Determine Cut Point for Trimming



We began exploring the appropriate cut point by directly plotting bias against standard error reduction, as shown in the plot in the upper right of Figure 1 (each point represents one potential cut point value). While we can see the behavior of bias against standard error reduction, it is hard to know the value of the cut point from the graph. So next, we used two methods to take into account the trade off between bias and the standard error reduction for specific cut points. We plotted the values of equations 8 and 9 below for cut points, as seen in the two plots on the left side of Figure 1. For both equations, the best cut point is determined at the minimum value (i.e., the trough in each of the plots above). Combining information from the three plots we chose the best cut point for the trimming process.

$$\text{Method 1} = \text{bias}(\widehat{LM_c})^2 - SE(\widehat{LM_o})^2 + 2 * SE(\widehat{LM_c})^2 \tag{8}$$

$$\text{Method 2} = \frac{\text{bias}(\widehat{LM_c})^2}{\widehat{LM_o}} - \frac{(SE(\widehat{LM_o}) - SE(\widehat{LM_c}))^2}{SE(\widehat{LM_o})} \tag{9}$$

where:

$\widehat{LM_o}$ = the estimate of total linear meters using weights before trimming, and

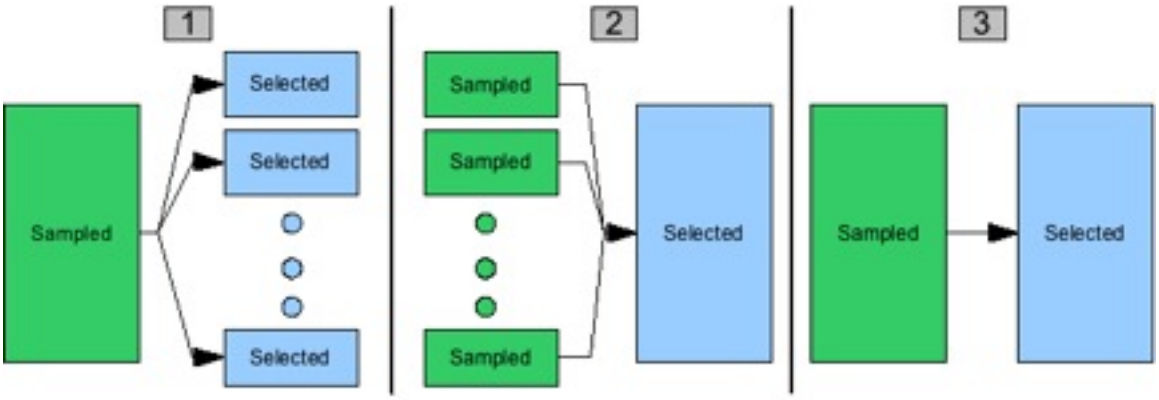
$\widehat{LM_c}$ = the estimate of total linear meters using trimmed weights at cut point c.

3. Additional Special Problem

Sections 2.1, 2.3 and 2.4 discussed problems resulting from our inability to accurately estimate the probability of selection for Environments in waves one and two, and LUAs and IUs in all waves. One additional problem in the weighting was due to the continuous movement of contents of containers around the warehouses. Thus, the makeup of an environment may have changed from the time of initial measures of size and the selection of containers, so that the contents of a sampled container were no longer in the same location. As Figure 2 shows, there were three main possibilities for the movement: One container was split into many containers; several containers were merged into one container; or all the documents were moved from a sampled container into another new container. A rare fourth possibility is not depicted in Figure 2. This possibility is referred to as movement from many containers to many other containers (see Table 2), and occurs when the contents of several

sampled containers are combined into several new containers.

Figure 2: Possible Container Movement



However, as Table 2 shows, the majority of the containers (65%) did not experience any movement. The second row of the table shows those situations where a container was moved from one location to another, and where we can assume there were no documents removed from the container nor documents added to the container from elsewhere, as is the case for 16% of the sampled containers. For 9% of the sampled containers, documents from many sampled containers were collapsed into one single container that was selected (but not originally sampled). Five percent of the containers were split from one container to many and one percent were moved from many containers to many containers. Lastly, we do not have information regarding movement for four percent of containers.

Table 2: 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

Weights can only be calculated in the usual manner for cases in which old and new containers were correctly matched, the new, receiving container(s) was (were) found, and there are estimates of the linear meters for all merged containers. If a sampled container is merged with another unsampled container and this is unknown, the weight will be based on a probability of selection that is lower than the true probability of selection, i.e., the weight ignores the fact that there was a second container that could have brought us to the same sampled container. In most cases, however, correct tracing of containers was possible. Therefore we do not believe that the movement seriously impacted the weights (see Table 2 above), with only the 48 containers in the last row of table 2 likely to have incorrect weights.

4. Future Weighting Step

We are planning to add an additional step in the weighting process. Since we independently select 33 environments within each wave (for waves 3-9) and make independent selections across waves, some environments are over-sampled and some are under sampled. We plan to rake the environment weights $W(E_i)$ to two sets of controls. This raking has not yet been done, and the estimates presented in the third paper in this session [2] use weights that do not include this raking adjustment. Note that the linear meters of paper in containers in an environment can change over time, and thus vary by wave. Thus, one set of controls for raking are the linear meters of paper in containers in each wave (across all environments). The second set of controls for raking are the linear meters of containers in each environment across waves. In doing the raking, we will probably do some combining of small environments, thus doing the raking by groups of environments rather than individual environments. A raking factor R_j will be applied immediately after the environment base weight.

5. Imputation

As mentioned in Section 2.7, many things can cause large weights. It is important to investigate these causes as they may be an indication of bad data. During our data cleaning step, we found two main sources of bad data in our study - implausible and missing data.

In order to detect implausible data we conducted the following face validity check: in terms of linear meters, $\mathrm{IU}_{ijkr} < \mathrm{LUA}_{ijk} < C_{ij} < E_i$. We found some measurements for containers and LUAs that were too large to be plausibly correct (i.e., LUAs that were larger than the containers from which they were sampled). This occurred in approximately 8% of containers and 17% of LUAs. Some of these errors occurred during data entry. Others may have been due to confusion in the unit of measurement, the place of a decimal point or missing data. These kinds of errors were likely exacerbated by high turnover in the field

team personnel over the course of the nine waves. Where possible we went back to the hard-coded forms to recover data that was incorrectly digitalized.

For the remaining erroneous data, we used the standard imputation method k nearest neighbors in the R software package called "impute."[4] The imputation was done for erroneous container and LUA measurements as well as for missing number of pages in an IU.

6. Conclusion

We have described the weighting for the sample of the Guatemalan National Police Archives and the primary challenges in calculating these weights. Despite the highly non-traditional nature of the sampling procedure dictated by the structure of the Archive, we are confident that we have achieved representative measures. Lessons learned from the challenges presented in this paper are already being implemented in ongoing sampling at the Archive today. Most importantly, samples as of wave 10 are no longer sensitive to paper movement inside the Archive. Another approximately 12,000 documents have been sampled in this wave and we expect the resulting weighting procedure to be simplified and thus less time-consuming. We will no longer have problems of unknown probabilities of selection, and expect that the weights will be much less variable. We believe these documents will enrich our understanding of the National Police's role in the violence in Guatemala's internal armed conflict between 1960 to 1996.

References

[1] Guzmán, D., Guberek, T., Shapiro, G., Zador, P. (2009), "Studying Millions of Rescued Documents: Sample PLan at the Guatemalan National Police Archive," in *JSM Proceedings*, Section on Survey Research Methods. Alexandria, VA: American Statistical Association.

[2] Price, M., Guberek, T., Guzmán, D., Zador, P., Shapiro, G. (2009), "A Statistical Analysis of the Guatemalan National Police Archive: Searching for Documentation of Human Rights Abuses," in *JSM Proceedings*, Section on Survey Research Methods. Alexandria, VA: American Statistical Association.

[3] Potter, F. (1988), "Survey Procedures to Control Extreme Sampling Weights," in *Proceedings of the Section on Survey Research Methods.* Alexandria, VA: American Statistical Association. 453–458.

[4] Hastie, T., Tibshirani, R., Narasimhan, B., Chureference, G. (2009), "Imputation for Microarray Data - An R Package," http://cran.r-project.org/web/packages/impute.index.