

## A Statistical Analysis of the Guatemalan National Police Archive: Searching for Documentation of Human Rights Abuses

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

### Abstract

This paper offers a first statistical look at the Guatemalan National Police Archive (GNPA). Preliminary estimates are provided of the degree to which specific groups within the National Police were aware of violent acts reported in the institution's documents. Weighted estimates were calculated based on a multi-stage sample of documents from the GNPA. As outlined in the previous papers, the physical structure of the Archive presented many challenges, resulting in complicated selection probabilities. The statistical analysis assumed containers of documents (stage 2) as the primary sampling unit. The proportion of all documents in the Archive of which any specific group had knowledge was compared to the proportion of documents *of interest* of which a group had knowledge. Approximately 20% of documents in the Archive are believed to contain information about acts of interest.

**Key Words:** multi-stage sampling, complex sample design, human rights violations, Guatemalan National Police Archive

### 1. Introduction

When the Guatemalan National Police Archive (GNPA) was discovered in 2005, families of victims, human rights groups and historians felt renewed hope at the prospect that these millions of documents could hold valuable information about crimes committed during the internal armed conflict. There was initial skepticism about whether or not documents of interest for the study of human rights violations would still exist in the Archive, or whether these kinds of documents would have been “cleaned” out of the cache of paper. Whether or not the Archive was purged is still an open question. Nonetheless, this paper offers a first statistical look at the Archive and confirms that its content are rich in information to learn about the country's violent past, even if future analyses reveal strategic gaps in documentation.

---

\*Megan Price, PhD., is a Statistician for the Human Rights Data Analysis Group at The Benetech Initiative. Tamy Guberek and Daniel Guzmán are consultants for the Human Rights Data Analysis Group at The Benetech Initiative [www.hrdag.org](http://www.hrdag.org), [www.benetech.org](http://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.

This study includes documents written or filed between 1960 and 1996, the time period of the internal armed conflict in Guatemala. Of the estimated 31,227,378 documents in the Archive within the period of interest, 20% contain information about acts of interest, such as deaths, disappearances and detentions. This paper describes the initial findings regarding the contents of the Archive and presents a comparison of knowledge of documents of interest versus knowledge of all documents for specific subsets of the National Police. “Knowledge” of a document and documents “of interest” are defined in Section 2.1.

One of the goals of this paper is to raise questions for the investigators and historians at the Archive who are local subject-matter experts. We hope these results will assist them in their ongoing research, which will add to the work of previous institutions to uncover Guatemala’s violent past such as the Historical Clarification Commission, the Catholic church’s REHMI project and numerous NGO reports from which information from the National Police is largely absent.[1]

The discovery of the historic Archive of Guatemala’s National Police, the choice to sample from the Archive and the methodology involved in the sampling project, as well as the calculation of weights resulting from the sampling method are all covered in our previous two papers.[2, 3] As a reminder, the general sampling structure was multi-stage, consisting of topographic levels inside the Archive: Environments, Containers, Last Units of Aggregation (LUAs), and Information Units (IUs). All documents within a selected IU were brought in to the sample.

Section 2 of this paper describes the variables we coded from the sampled documents and their relevance to the objectives of this study. Section 3 presents the statistical results. Section 4 offers some discussion about the relevance of the results and some of the statistical challenges we encountered in the calculations. Section 5 concludes the paper. Lastly, Section 6 foreshadows some analyses for future papers on the GNPA.

## 2. Methods

### 2.1 Coding the Documents

The sampling design aimed to respect the conditions and filing structures found in the GNPA as much as possible. Since the National Police often linked and filed together several individual documents concerning the same subject as one unit, we sampled information units (IU) rather than individual documents. Information was then coded from each of the documents within a sampled IU. The data used for this statistical analysis comes from the disaggregated documents within IUs (i.e., each record in the analysis dataset corresponds to one document).

---

Our coding scheme was designed to collect data from the documents in order to answer two specific research questions: What could we learn from these documents about the command and control structures of the National Police? Would analyses of patterns and magnitudes of possible human rights violations mentioned in the Archive's documents add to the understanding of Guatemala's violent past, in particular about forced disappearances? We discuss the initial results pertaining to these questions in Section 4.

The variables we coded include general information about each document such as document type, date, and condition. Other variables critical to the analyses contain information on what we call "document-level actors", such as authors, direct and indirect recipients and attributes about these actors such as police unit or other affiliation and rank of any individuals associated with the creation or transmission of the document. The last level of information we recorded was about the contents of documents, specifically if the documents included information on acts of interest to the study. Documents were considered to be "of interest" if they contained information on any of a list of 11 acts potentially related to human rights violations (see Table 2).

Because a primary question of interest in this study is who within the National Police had knowledge of various acts, additional indicator variables were created from the document-level variables described above. For example, document actors were grouped according to ranks within the police hierarchy to define positions of "leadership" within the National Police. Police leadership was then considered to have knowledge of a document if a person with leadership rank authored, received (directly or indirectly) or typed any aspect of the document.

The National Police was also divided into units and subunits. As suggested by our partners in Guatemala, we combined the approximately 50 subunits according to the area of focus of the unit into 13 new categories. Each of these categories was considered to have knowledge of a document if anyone affiliated with that unit authored or received it (directly or indirectly).

Since the information contained in the Archive documents is primarily qualitative in nature, we created a controlled vocabulary for all the variables to be coded. The controlled vocabulary was instrumental in helping maintain high levels of agreement between coders whose job it was to transform the qualitative information to quantitative units.[4] The inter-rater reliability was calculated to be consistently over 75% throughout the nine waves of sampling.

## 2.2 Calculating Estimates

The R package ‘survey’ was used to calculate all of the estimates described in the following sections.[5] The package includes functions for creating survey design ‘objects’ based on cluster and stratification information, calculated weights, and sampling method. By the end of the ninth wave we had conducted essentially a census of environments within the Archive; therefore, for variance estimation purposes, the second stage, containers, was considered the primary sampling unit in these analyses. The survey package also provides a function for calculating replicate weights; in our case using the JK1 method.[6] These replicate weights were then used to estimate the variance of each point estimate reported in this paper.

## 3. Results

This section presents estimates of the National Police’s knowledge of acts of interest, based on the multi-stage sample. While this report alone is not sufficient to draw definitive conclusions regarding the National Police’s role in Guatemala’s violent past, we believe that these estimates will raise important questions for subject matter experts, will help direct future research into the command structure of the National Police and will aid in maintaining interest in and resources for future research.

### 3.1 Global Estimates

Table 1 presents global estimates about the Archive documents based on the multi-stage sample. Since information units were sampled rather than documents directly, an IU may have included documents from years before or after 1960-1996 as well as at least one document from within that time period. An IU may have also contained a document with no date of creation along with a document with a valid date. Identity cards are one type of document that typically do not have a single document creation date, but rather dates for each entry added to the card. Since both cases mentioned above (out-of-range date or missing a valid document date) may be included in an eligible IU, two totals are presented in Table 1: documents *related* to the study’s time period and documents *within* the time period.

The last row in Table 1 presents an estimate of the total number of acts mentioned in the Archive documents. Note that this estimate does not account for the fact that the same act can be mentioned in multiple documents. In the future, we plan to de-duplicate the acts in the sample and offer an estimate of unique acts in the Archive documents.

**Table 1:** Estimates of Documents in Archive

	<b>Estimate (SE)</b>	<b>95% CI</b>
<b>Total Related to 1960-1996</b>	34,971,341 (5,263,488)	(24,654,904, 45,287,778)
<b>Total Within 1960-1996</b>	31,227,378 (5,109,612)	(21,212,538, 41,242,219)
<b>Total of Interest</b>	6,285,487 (1,058,066)	(4,211,678, 8,359,297)
<b>Proportion of Interest</b>	0.20 (0.02)	(0.16, 0.24)
<b>Total Number of Acts</b>	12,079,401 (2,275,698)	(7,619,033, 16,539,769)

### 3.2 Acts of Interest

Table 2 is a disaggregated list of the 11 acts of interest coded from the documents. It shows in decreasing order the total estimated number of documents containing information on each type of act and what proportion of documents of interest this constitutes. Documents *of interest*, refer to documents that contain information about the acts in this list. Documents may contain information on more than one act, so the summation of the totals in Table 2 is greater than the estimated total number of documents of interest in Table 1. Also, any specific act may be referred to in several documents and the frequency with which acts are mentioned may depend on several factors, including the act's type and the document's date.

**Table 2:** Disaggregation of Acts of Interest

	<b>Total (SE)</b>	<b>Proportion (SE)</b>
<b>Detentions</b>	4,658,367 (904,385)	0.741 (0.04)
<b>Deaths</b>	1,365,218 (345,939)	0.217 (0.04)
<b>Denunciations</b>	572,299 (165,704)	0.091 (0.02)
<b>Physical and psychological abuse</b>	439,666 (189,983)	0.070 (0.03)
<b>Intimidation</b>	302,113 (99,268)	0.048 (0.02)
<b>Disappearances</b>	223,749 (62,035)	0.036 (0.01)
<b>Writ of habeas corpus</b>	65,519 (27,956)	0.010 (0.004)
<b>Interrogations</b>	30,171 (22,163)	0.005 (0.004)
<b>Entering private property</b>	29,484 (12,408)	0.005 (0.002)
<b>Sexual abuse</b>	14,102 (4,984)	0.002 (0.001)
<b>Surveillance</b>	1,307 (1,199)	<0.0001 (<0.0001)

Note that even in the case of relatively elusive acts, such as disappearances, which comprise only 3.6% of documents with acts of interest, the total estimated number of documents (in this case 223,749) may still be non-trivial.

Various police subunits' knowledge of documents over time will be described in the following sections, but it is worth noting here one subunit's knowledge of disaggregated acts. The departmental subunits, those that patrolled areas outside of Guatemala City, had disproportionately high amounts of information about specific acts of interest: they were informed about approximately 75% (SE = 17.1) of the physical and psychological abuse documented in the police records, 54% (SE = 13.2) of the deaths and 45% (SE = 34.6) of the trespassing onto private property. These are higher rates of knowledge about specific acts of interest than any other subunit except for the units who conduct investigations, which had knowledge of 90% of interrogations (it seems intuitive that an investigative group of units would have knowledge of most interrogations, however the interpretation of this statistic remains an open question for our partners at the Archive). This raises important questions for subject-matter experts conducting research in Guatemala - are these higher rates of knowledge due to the departmental subunit's broader coverage of the country? Is it because the regions were more violent than the city? Is it possible that the city headquarters Archive was purged of documents of interest before its discovery in 2005?

### 3.3 Documents over Time

Next, we compare the proportion of documents *of interest* of which a group had knowledge to the proportion of *all* documents of which a group had knowledge over time. This baseline comparison is important since a group may have what appears to be increased knowledge of documents of interest during a specific time period, but this may or may not be due to an increase in the amount of *all* documents of which a group has knowledge. In other words, we suspect that differences in patterns of knowledge of different classes of documents hold the most potential for identifying areas where additional information may be found regarding systematic human rights violations.

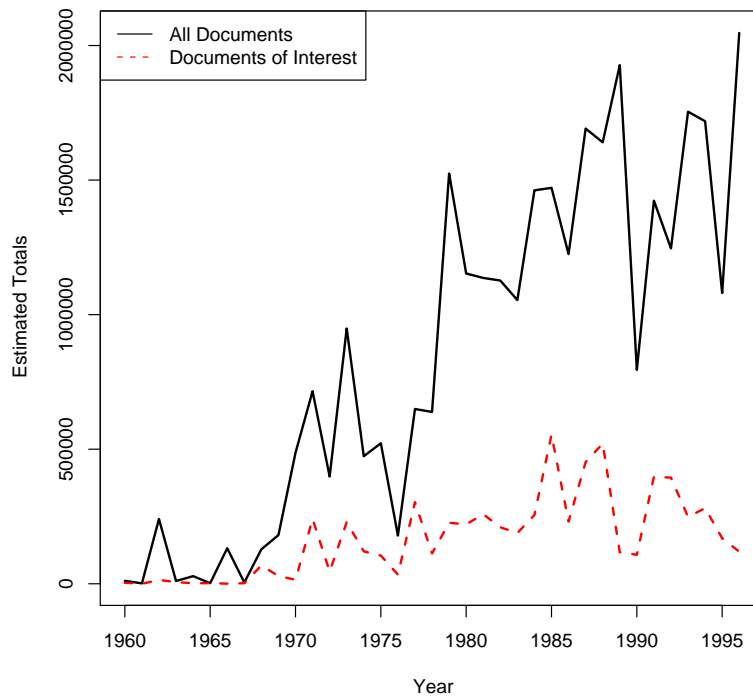
Note that dates used in these analyses are dates of document creation rather than date of occurrence of the act. Future analyses will compare patterns of document dates to act dates. It is also important to note that approximately 11% of documents in the sample (902 out of 8,162) did not contain date information (for reasons described above in section 3.1).

Lastly, it is also important to note that although estimates appear to be very close to zero during the beginning of the period of interest (the early 1960s) this is simply an artifact of the larger estimates in subsequent years - estimates from the early 1960s are one to two orders of magnitude smaller than later estimates. These early estimates are also

based on a much smaller sample size than later estimates, and thus it should be kept in mind that these measures have much larger estimates of variance (excluded from figures to make pattern detection easier).

Figure 1 shows that over time the National Police produced an increasing amount of documents (solid line), however there was a much smaller increase in the number of documents of interest (dotted line) created across the same time period.

**Figure 1:** All Documents 1960-1996 vs. Documents of Interest 1960-1996



This same pattern can be seen among some subunits (headquarters in particular) in Section 3.4.

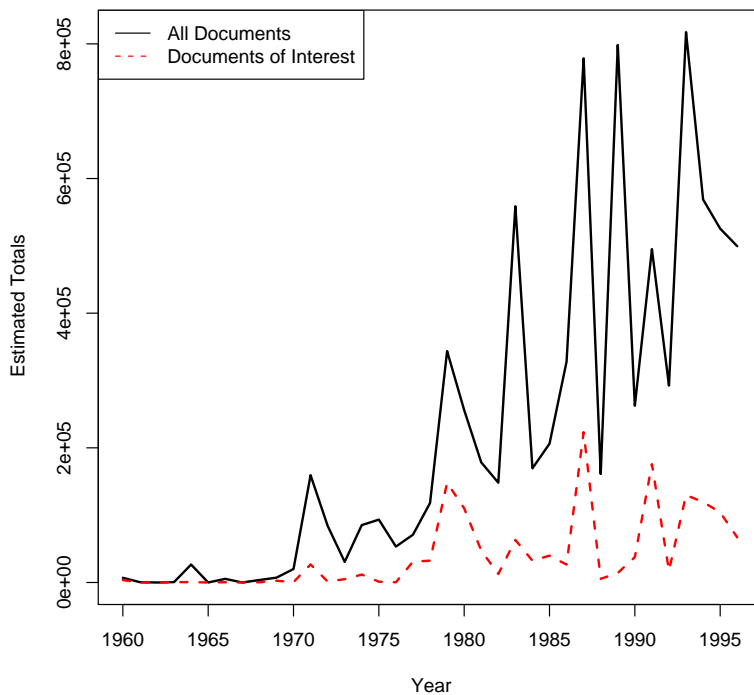
It is also important to note that the pattern of documents *of interest* in Figure 1 differs from generally accepted patterns of violence across Guatemala's history. Potential reasons for, and important questions raised by this difference will be addressed in Section 4.

### 3.4 Police Subunit Knowledge of Documents of Interest

The subunits presented in this paper were chosen because of qualitative knowledge of their role in repression and because of notable patterns observed in the data. Sometimes, as in the case of the police headquarters (see Figure 2), this pattern matches the overall pattern observed in Figure 1. Others, as noted below in Figures 3 and 4, are unique.

Similar to the pattern noted in Figure 1, Figure 2 indicates that the police headquarters (referred to by the Spanish acronym DIGE) had knowledge of an increasing number of all documents, in general, over time (solid line) but a relatively static (with a few exceptions) number of documents of interest (dotted line).

**Figure 2:** All Documents with DIGE Knowledge vs. Documents of Interest with DIGE Knowledge



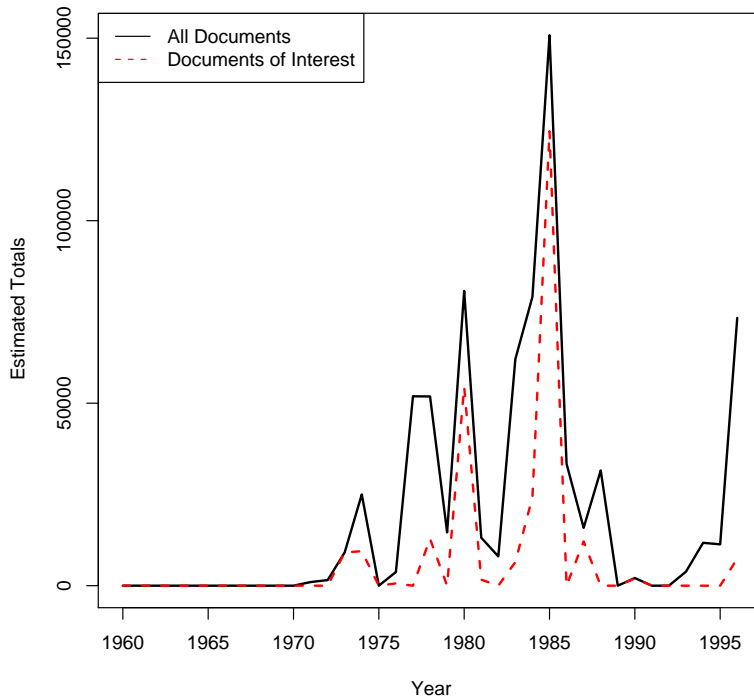
The motor vehicle police subunit was the unit tasked with patrolling Guatemala City's



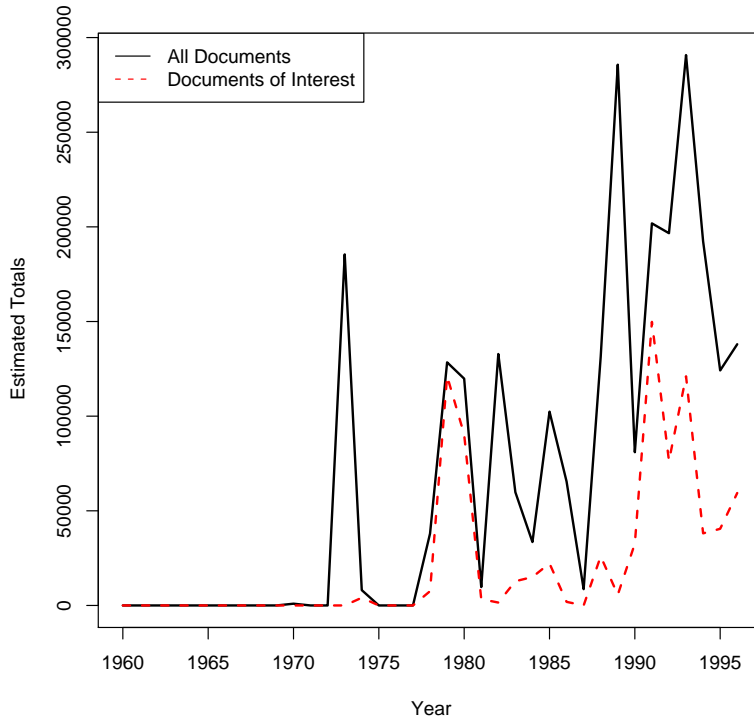
streets, and was suspected of taking part in surveillance. Although they had knowledge of a relatively small proportion of the entire archive, Figure 3 shows that in both 1980 and 1985 this subunit experienced non-trivial increases in knowledge of both all documents (solid line) and specifically documents of interest (dotted line).

Figure 4 illustrates two interesting details of the kinds and amount of documents known to the Joint Operations Unit (*Centro de Operaciones Conjuntas - COC*) – the unit that worked directly with the Guatemalan military. First, in 1973 they experienced a drastic increase in their knowledge of all documents (solid line), but not a corresponding increase in knowledge of documents of interest (dotted line). Then in 1979 they experienced another drastic increase in knowledge, this time also in documents of interest. Subsequent years (1982, 1985, 1989) indicate similarities to 1979 - sudden increases in knowledge of all documents without corresponding increases in documents of interest.

**Figure 3:** All Documents Known by the Motor Vehicle subunit Knowledge vs. Documents of Interest Known by the the subunit



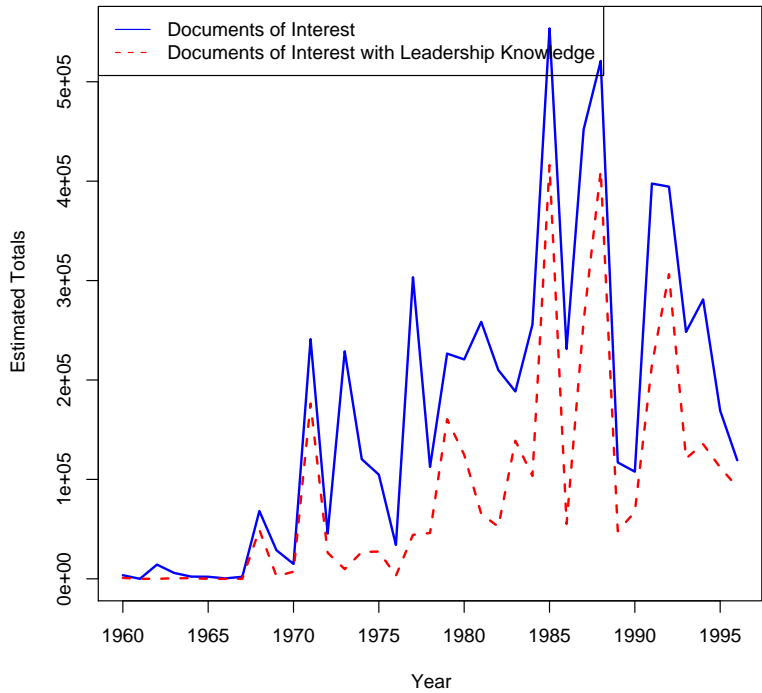
**Figure 4:** All Documents with COC Knowledge vs. Documents of Interest with COC Knowledge



### 3.5 Leadership Knowledge of Documents of Interest

Understanding the information particularly known to the leadership of the National Police was a primary motivation for this first round of estimates. Note that the solid line in Figure 5 represents a different set of data than in previous figures - Figure 5 compares the number of documents of interest to documents of interest with police leadership knowledge. This graph indicates that for most years, leadership had knowledge of the majority of documents including information on acts of interest. Police leadership cannot claim ignorance of the operations of the National Police with respect to acts of interest.

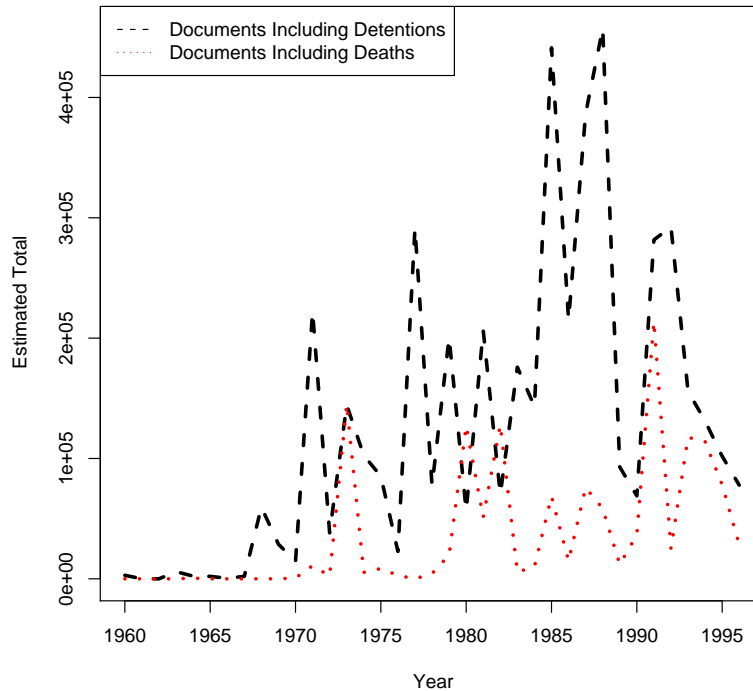
**Figure 5:** All Documents with Leadership Knowledge vs. Documents of Interest with Leadership Knowledge



### 3.6 Disaggregated Acts Over Time

Lastly, Figure 6 indicates a particularly interesting pattern in 1980 and 1982, when there are actually more documents including information about deaths than detentions. If one considers that a primary activity of a police force is to detain individuals, it is quite shocking that in these two years there are a higher number of documents dealing with deaths. These two years will certainly warrant additional investigation.

**Figure 6:** All Documents with Information on Deaths vs. Documents with Information on Disappearance



## 4. Discussion

### 4.1 Magnitude of Possible Human Rights Violations

As a first estimation of the Archive's contents, we believe our study succeeds in indicating that the Archive provides a rich source of information regarding both the pattern and magnitude of possible human rights violations. As shown in Table 2 there are likely hundreds of thousands, and in some cases millions, of documents with information on deaths, detentions, disappearances, intimidation, and physical and psychological abuse. As an example, the 223,749 estimated documents on disappearances have the potential to offer a wealth of information that could lead to knowing the fate of the many people who were disappeared during this chapter of Guatemala's history. Although some of the other acts of interest (listed in Table 2) fall under the purview of standard police knowledge, the sheer magnitude

of information contained within the Archive makes it well worth continued study.

## 4.2 Who Knew and to What Degree?

While the magnitude of documents with acts in the Archive is noteworthy, the pattern of acts the National Police documented over time also provides important insight into this period in Guatemala's history. In particular, it is important to contrast the pattern shown in Figure 1 with historically established patterns of violence in Guatemala during the same time period. Although the acts of interest listed in Table 2 are not all acts of violence, they do include several violent acts such as death, disappearance and physical abuse. Multiple historical sources report a rural counterinsurgency campaign between 1981 and 1983. During these campaigns, the Historical Clarification Commission (CEH) found that the Guatemalan army committed acts of genocide against the Mayan populations.[1, 7, 8] It is troubling that there does not appear to be an increase in the number of documents of interest during this time period in Figure 1 as one would expect if the National Police were documenting the number of deaths that would result from acts of genocide. The same sources report a peak in urban violence in 1979 and 1980, during which time many people were forcibly disappeared. Again, this pattern is not reflected in Figure 1. This second pattern is particularly noteworthy as the National Police's records of violent acts are primarily focused on urban areas.

Although we would not expect police documentation to exactly match patterns of violence, we would expect that during periods of known heightened violence, police documentation of violent acts would also increase. Therefore, it is surprising that the pattern of documents with acts of interest is divorced from the accepted patterns of known violence in Guatemala.[1] The possibility that police units were commanded not to document certain types of systematic violence, or that such documentation, if it ever existed, was systematically purged, has not escaped our attention. As mentioned in Section 1, this possibility, and how to formally test this as a hypothesis, remains an open question.

## 4.3 Statistical Challenge

As described in the previous paper in this series [3], the complicated calculation of weights and resulting source of high levels of variability presented challenges in comparing the estimates included in this paper. However, the primary goals of this paper are to summarize the Archive's contents, present estimates of patterns of knowledge, and help to identify more targeted areas for future research. We are optimistic that the lessons learned throughout this study will help us to improve weight calculations in future waves, thus reducing one main source of variability, and improving our ability to compare key estimates.

Another specific source of additional variance which we hope to remedy in future analy-

ses is incorrect treatment of self-representing units. As mentioned in Shapiro [3] the highly variable weights were in part due to the choice to conduct sampling with probability proportional to size for containers, which was driven by the variety of sizes of containers. In the few cases where containers were self-representing (i.e., only one container in an environment), the appropriate primary sampling unit would have been LUA. Similarly, in cases with only a few containers, either LUAs should have been used as the primary sampling units or a finite population correction factor should have been considered. Although failure to carry out either of these analytical steps likely resulted in slight over-estimation of variance, we do not believe the one-or-few container case to have occurred often.

## 5. Conclusions

Despite the numerous physical challenges presented to the sampling scheme by the nature of the Archive, we believe we have applied a rigorous probability sampling technique to a unique data source for human rights research. We are hopeful that this first step in summarizing the contents of the GNPA will not only draw attention to a potentially rich source of information but also assist other researchers in conducting more targeted research in the future.

## 6. Future Work

The results presented here are but a tiny fraction of the research that is possible with such an abundance of data. Future analyses will not only incorporate additional variables from our existing dataset (act dates in addition to document dates, authorship and recipient information, and policy decisions) but also the 12,000 additional documents that have been coded in wave 10. Subsequent sampling waves will also rely on lessons learned from this study to continue simplifying the sampling and weighting procedures. Lastly, we hope to direct more targeted future sampling projects with a focus on more specific time periods and/or subunits of interest as a result of the patterns identified in this paper.

## References

- [1] Guatemalan Historical Clarification Commission (CEH), (1999). *Memoria del Silencio*, Volume I, page 50 paragraph 95.
- [2] Guzmán, D., Guberek, T., Shapiro, G., Zador, P. (2009), “Studying Millions of Rescued Documents: Sample PPlan at the Guatemalan National Police Archive,” in *JSM Proceedings*, Section on Survey Research Methods. Alexandria, VA: American Statistical Association.
- [3] 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.

- [4] Silva, Romesh (2002) “On the Maintenance and Measurement of Inter-Rater Reliability when Documenting Large-Scale Human Rights Violations,” *Proceedings of the Joint Statistical Meetings of the American Statistical Association, the International Biometric Society ENAR and WNAR, the Institute of Mathematical Statistics, and the Statistical Society of Canada*.
- [5] Lumley, T. (2009) “survey: analysis of complex survey samples”. R package version 3.16.
- [6] Quenouille, M. (1949) “Approximation Tests of Correlation in Time Series,” *Journal of the Royal Statistical Society B*. 11, 18-84.
- [7] Ball, P., Kobrak, P., and Spierer, H.F. ‘*State Violence in Guatemala, 1960-1996: A Quantitative Reflection*. American Association for the Advancement of Science and the International Center for Human Rights Research (CIIDH)
- [8] Proyecto Interdiocesano de Recuperacion de la Memoria Historica (REMHI) Guatemala (1998) *Nunca Mas*. Guatemala, ODHAG.