

# Artículos presentados ante la Joint Statistical Meeting –JSM- en 2009

La Joint Statistical Meeting –JSM- es un congreso que reúne anualmente a estadísticos en Norte América. Son cinco las sociedades estadísticas consideradas como miembros oficiales del congreso: la Asociación Americana de Estadística –ASA- (por sus siglas en inglés), el Instituto de Estadística Matemática –IMS- (por sus siglas en inglés), la Sociedad Biométrica Internacional del este –IBS- (por sus siglas en inglés), la Sociedad Biométrica Internacional del oeste –IBS- (por sus siglas en inglés) y la Sociedad Estadística de Canadá. En agosto del 2009 en Washington D.C se realizó uno de los congresos, en el cuál Benetech presentó tres artículos, utilizando de referencia 9 de las 11 iteraciones del Estudio Cuantitativo desarrollado en el Archivo Histórico de la Policía Nacional, que se tradujeron al español bajo la supervisión de Benetech, siendo los que a continuación se presentan. Las versiones en inglés se encuentran disponibles en: <http://www.amstat.org/meetings/jsm/2009/index.cfm> y en <http://www.hrdag.org/resources/publications>.

Guzman, D.R., Guberek, T., Shapiro, G.M., Zador, P. 2009. "Studying Millions of Rescued Documents: Sampling Plan at the Guatemalan National Police Archive (GNPA)." In JSM Proceedings, Survey Research Methods Section. Alexandria, VA: American Statistical Association. ("Estudiando Millones de Documentos Rescatados: Plan de Muestreo para el Archivo Histórico de la Policía Nacional de Guatemala", traducción propia AHPN).

Shapiro, G.M., Guzman, D.R., Zador, P., Guberek, T., Price, M.E., Lum, K. 2009. "Weighting for the Guatemalan National Police Archive Sample: Unusual Challenges and Problems." In JSM Proceedings, Survey Research Methods Section. Alexandria, VA: American Statistical Association. ("Ponderación de la Muestra del Archivo Histórico de la Policía Nacional de Guatemala: Retos y Problemas Inusuales", traducción propia AHPN).

Megan Price, Tamy Guberek, Daniel Guzmán, P. Zador, and Gary Shapiro. A statistical analysis of the Guatemalan National Police Archive: Searching for documentation of human rights abuses. In JSM Proceedings, Survey Research Methods Section, Alexandria, VA, 2009. American Statistical Association. URL: <http://www.hrdag.org/resources/publications/JSM-GT-estimates.pdf>. (Análisis Estadístico del Archivo Histórico de la Policía Nacional de Guatemala: en Busca de Documentación sobre Violaciones de Derechos Humanos). El universo utilizado para realizar dichas inferencias es menor al utilizado para presentar las inferencias en esta publicación.

## **Estudiando Millones de Documentos Rescatados: Plan de Muestreo para el Archivo Histórico de la Policía Nacional de Guatemala**

Daniel Guzmán, Tamy Guberek, Gary Shapiro y Paul Zador<sup>1</sup>

### **Resumen**

El presente trabajo describe el diseño muestral empleado en el Archivo Histórico de la Policía Nacional de Guatemala (AHPN). El Archivo contiene millones de documentos encontrados inicialmente desordenados y en malas condiciones físicas. Debido a las dimensiones del Archivo y a la falta de un marco muestral tradicional, optamos por utilizar un muestreo PPT multietápico y aleatorio usando la topografía del Archivo para las etapas 1 y 2. Para las etapas 3 y 4 se crearon marcos *in situ*. Enfrentamos una gran cantidad de retos a la hora de llevar a cabo el muestreo, tales como la incierta disponibilidad de recursos y el traslado de los documentos mientras éstos se restauraban y organizaban. Para enfrentar estas dificultades diseñamos iteraciones muestrales iterativas. Después de varias rondas de evaluación, modificamos el muestreo para reducir una etapa, haciendo el muestreo más eficaz. En más de 2 años de muestreo hemos seleccionado 20,000 documentos. Posteriormente, podremos utilizar un muestreo adaptable para buscar documentos de interés con mayor intencionalidad y probabilística.

**Palabras clave:** muestreo probabilístico, diseño muestral complejo, marco muestral, muestreo por conglomerados, violaciones a los derechos humanos, Archivo Histórico de la Policía Nacional de Guatemala.

### **1. Introducción**

En julio de 2005, la explosión en un depósito militar de municiones cerca de Ciudad de Guatemala acrecentó las preocupaciones en relación con el almacenamiento de explosivos en zonas residenciales cercanas. Las personas que vivían en el vecindario pidieron a las autoridades que investigaran un edificio del complejo de la Policía Nacional Civil en Ciudad de Guatemala. Un equipo de la Procuraduría de los Derechos Humanos (PDH), entidad estatal, entraron en el edificio semi-derruido y descubrieron un enorme acervo de documentos.

Los archivos estaban almacenados en una serie de cuartos oscuros, infestados por ratas, murciélagos y cucarachas. Muchos de los papeles se encontraban empapados por el agua que se filtraba por las goteras y las ventanas rotas. Los documentos, cuyo número alcanzaba como mínimo decenas de millones, resultaron ser el Archivo Histórico de la Policía Nacional de Guatemala (AHPN). La Policía Nacional fue disuelta después de la firma de

---

<sup>1</sup>Daniel Guzmán y Tamy Guberek son consultores del Grupo de Análisis de Datos sobre Derechos Humanos de la organización Benetech (*Human Rights Data Analysis Group, HRDAG*, por su nombre y sigla en inglés). Paul Zador, PhD., es Estadístico Principal de Westat. Gary Shapiro, recientemente jubilado, fue estadístico principal de Westat durante la mayor parte de su participación en este estudio. Ambos, Shapiro y Zador, son miembros del Grupo Voluntario de Intereses Especiales de la Asociación Americana de Estadística.

los Acuerdos de Paz en 1996 y fue reemplazada por la Policía Nacional Civil. Muy probablemente, el AHPN puede ser el acervo documental más grande que jamás se haya puesto a disposición de los investigadores en derechos humanos.

A partir de 2005 se han estado realizando tres procesos simultáneos en el AHPN: 1) trabajo archivístico, que incluye la restauración, preservación y catalogación de documentos; 2) una investigación judicial de casos; y 3) muestreo y análisis. Este trabajo describe el diseño muestral para el tercer proceso, el estudio cuantitativo para las iteraciones muestrales 1 a 9. El presente documento es el primero de una serie de tres, y en él se describe el esquema muestral. El segundo documento presenta el cálculo de las ponderaciones del muestreo. El tercer documento comprende las estimaciones iniciales en lo que respecta a los contenidos del Archivo Histórico de la Policía Nacional.

Este informe está estructurado como se describe a continuación: los apartados 2 a 5 explican el diseño e implementación del muestreo. Los epígrafes 6 y 7 analizan los diversos retos del muestreo y las decisiones tomadas para hacerles frente. El capítulo 8 provee estadísticas descriptivas no ponderadas sobre los datos recogidos. El apartado 9 expone las próximas medidas a adoptar conforme evolucione el proyecto. Por último, en el acápite 10 se resume el informe y se aportan algunas conclusiones generales.

## **2. La decisión de muestrear**

Dada la magnitud y las precarias condiciones físicas del Archivo, un estudio en profundidad de su contenido podría llevar años e incluso décadas. Sin embargo, hay muchas interrogantes que deben responderse y mientras más pronto, mejor. ¿Qué tipo de información contiene este conjunto de documentos? ¿Los documentos contienen información acerca de la violencia política ocurrida durante el conflicto armado interno del país? ¿Revelarán que la represión, los asesinatos y las desapariciones forzadas cometidos contra civiles por parte de la policía fueron parte de los objetivos institucionales, o solamente actos de unos cuantos oficiales del cuerpo? El muestreo probabilístico es un método eficiente y se justifica para empezar a investigar estas cuestiones.

El diseño muestral original tuvo en cuenta tres objetivos:

1. Entender el alcance y la heterogeneidad de este Archivo masivo e inexplorado.
2. Recolectar datos acerca de los macropatrones de actuación de los operativos policiales, tales como las estructuras de mando y el flujo de información.
3. Hacer estimaciones acerca de la proporción de documentos que registraron ciertos actos y políticas de interés para el proyecto, lo cual incluye desapariciones, detenciones y muertes.

A pesar de que el Archivo contiene documentos de más de un siglo de antigüedad, el período de tiempo del estudio se limita a los años de interés histórico que comprenden el conflicto armado interno, entre 1966 y 1996, ambos inclusive. El diseño se elaboró con la intención de que, en su fase inicial (iteraciones 1-9), fuera más extensivo que profundo. Esto implicó que, inicialmente, escogiéramos menos documentos por cada ubicación muestreada en el Archivo, pero que seleccionáramos documentos de más ubicaciones. Esto

también implicó que centráramos la codificación en la información estructural de los documentos muestreados más que en los detalles de sus contenidos (por ejemplo: autores, receptores, fechas, presencia/ausencia de hechos de violencia, etc.). Con los datos y las lecciones aprendidas durante el proceso de muestreo de las iteraciones 1-9, se llevaron a cabo cambios incrementales en el diseño para registrar una mayor cantidad de información referida al contenido de los documentos. Estos y otros cambios se describen en el apartado

### **3. El diseño muestral básico**

Para obtener una muestra de los documentos se empleó un diseño probabilístico multietápico iterativo. Se realizaron muestras iterativas para obtener control sobre el traslado de los documentos. Se realizó un muestreo de probabilidad proporcional al tamaño (PPT) en cada una de las etapas. El muestreo PPT fue importante debido a la varianza en el tamaño de las unidades en cada etapa. El tamaño se midió en metros lineales o en metros cúbicos (volumen) de papel en todas las etapas.

Para la determinación de nuestro diseño muestral tomamos en consideración los siguientes retos:

- **INEXISTENCIA DE UN MARCO MUESTRAL DIRECTO DE LOS DOCUMENTOS.** En el Archivo no había disponible una enumeración de los contenidos del universo que sirviera de base para el trabajo, y la creación de tal enumeración habría costado demasiado tiempo y recursos.
- **IDEAS POCO CLARAS CON RESPECTO AL TIEMPO Y A LOS RECURSOS.** Cuando se inició el proyecto no se tenía claro cuánto tiempo duraría la jurisdicción y la voluntad política del PDH para organizar e investigar en el Archivo. El proyecto podía haberse terminado en poco tiempo. Además, el proyecto debía recaudar fondos para mantenerse a sí mismo, por lo que la disponibilidad de recursos era incierta. Debido a ello, no resultaba realista fundamentar el muestreo en un tamaño muestral predeterminado.
- **TRASLADO.** El muestreo se realizó al mismo tiempo que otros procesos mencionados previamente: trabajo archivístico e investigación de casos. Esto causó que se trasladaran algunos documentos para restaurarlos, organizarlos y estudiarlos.
- **TAMAÑO CAMBIANTE.** El descubrimiento inicial del Archivo motivó el descubrimiento de otros archivos policiales regionales. A raíz de estos hallazgos, en varias ocasiones se introdujeron nuevos documentos al Archivo principal, alterando así el tamaño del universo para el cual se había diseñado la muestra.

En el próximo apartado se presentan todas las etapas en detalle (ver la Tabla 3 para un resumen de las etapas muestrales), la unidad de medida dentro de cada etapa y la fuente del marco muestral para cada una de las fases.

**Tabla 1: Etapas de Muestreo**

Etapa	Medida	Marco Muestral
Ambiente	Metros lineales	Registro Maestro de Ubicación (RMU)
Contenedor	Metros lineales	Registro Maestro de Ubicación (RMU)
Última Unidad de Agregación (UUA)	Volumen ( $cm^3$ )	Coordenadas 3D
Unidad de Información (UI)	Milímetros lineales	Altura de UUA

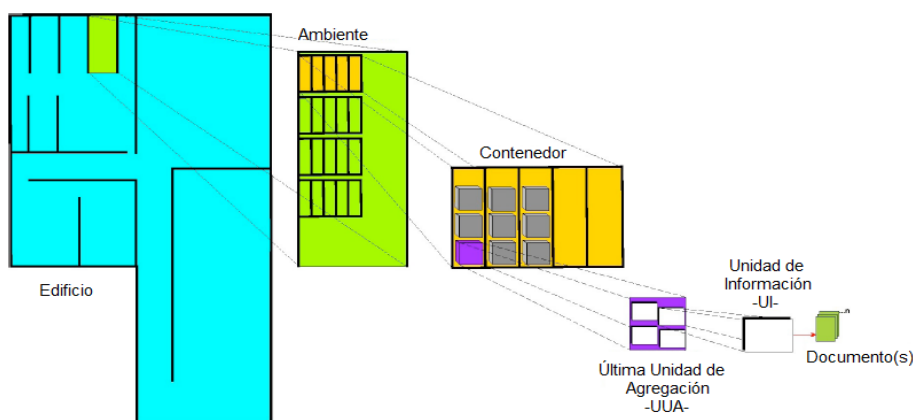
#### 4. Construcción del marco muestral

Aunque no había un inventario detallado que enumerara todos los documentos del Archivo y que sirviera de marco muestral, una de las primeras acciones que el equipo de coordinación del AHPN adoptó cuando tuvo jurisdicción sobre el Archivo fue identificar y etiquetar, en cada local, todos los espacios que contuvieran documentos (el muestreo excluyó todo lo que no estuviera en papel. Hay proyectos paralelos que se centran en los medios electrónicos y que se encuentran actualmente en marcha). Se utilizó un sistema de topografía jerarquizada para asegurar el control y la cobertura de todos los espacios.

A cada uno de los espacios del Archivo se le asignó una identificación única. Al nivel más alto, la topografía comenzó con los siete edificios que componen el Archivo. Dentro de los edificios se identificaron “ambientes”, los cuales, en muchas ocasiones, son cuartos parcialmente cerrados. Una vez dentro del ambiente, se identificó y etiquetó de manera individual cualquier contenedor con documentos. Dentro de la categoría de contenedor se incluyen estanterías, muebles de archivo, la superficie de las mesas, sacos, plataformas de madera y espacios en el piso. A continuación se midió el contenido de cada contenedor utilizando el sistema archivístico de medición, usualmente metros lineales.

La Gráfica 1 es un diagrama de la topografía jerarquizada del Archivo que se utilizó para el muestreo. Vale la pena indicar que la última etapa del muestreo es la unidad de información (UI), que a su vez contiene uno o más documentos a codificar.

**Gráfica 1:** Diagrama del Archivo: del edificio al(a los) documento(s)



El inventario topográfico lleva por nombre Registro Maestro de Ubicación (RMU). El RMU constituye la lista maestra de los metros lineales de papel en los contenedores y también registra todo desplazamiento de papel que se lleva a cabo en el área del Archivo. Existe una necesidad permanente de mantener el RMU actualizado porque la ubicación y las cantidades de documentos cambian frecuentemente mientras los documentos se limpian, escanean y archivan.

El diseño multietápico sacó provecho de este inventario métrico y topográfico. Éste sirvió como marco muestral principal para las primeras dos etapas del muestreo: ambientes y contenedores. Para cada iteración se empleó la versión más actualizada del RMU. Haciendo uso de la topografía del Archivo, el equipo de campo llegaba sistemáticamente a un contenedor del cual podía tomar muestras de los documentos dentro del Archivo. A continuación se crearon marcos muestrales *in situ* para seleccionar las últimas unidades de agregación (UUA) de dentro de los contenedores.

Las dimensiones medidas de los contenedores se utilizaron como el marco muestral para la tercera etapa, a partir de la cual se escogió un grupo de documentos. Estas agrupaciones, llamadas UUA, se definen como las más pequeñas agrupaciones de documentos reconocibles dentro de un contenedor. Entre los ejemplos de UUA están paquetes de documentos atados juntos con cuerdas, carpetas, gavetas de gabinetes, bolsas y cajas.

Las dimensiones de las UUA se utilizaron como marco muestral para la cuarta y última etapa de muestreo, las llamadas unidades de información (UI). Las UI se definen como un documento individual o un conjunto de documentos que fueron archivados juntos, como una sola unidad, por la unidad de policía de la que provienen y que están relacionados con un tema, caso o fenómeno que les es común. Tanto un documento de una sola página como el expediente de un caso compuesto de varios documentos pueden ser una unidad de información. La UI típica sería un conjunto de papeles engrapados juntos o agrupados con un clip. De conformidad con los objetivos analíticos de este estudio, codificamos información de interés a partir de los documentos que conforman las UI muestreadas. Si una UI muestreada no contenía ningún documento escrito entre 1960 y 1996, clasificábamos la UI como ilegible para el estudio y pasábamos al siguiente punto muestral indicado.

**Gráfica 2:** Documentos agrupados en paquetes dentro de una variedad de contenedores en uno de los ambientes del AHPN



5. Procedimiento muestral

En el siguiente apartado se describen las cuatro etapas del muestreo.

### 5.1 Etapa 1: Ambientes

Para obtener una mejor impresión de heterogeneidad de los contenidos del Archivo en una ubicación dada, y debido al desconocimiento de las limitaciones de tiempo para realizar el sondeo, decidimos desde un principio concentrarnos en unos pocos ambientes en cada iteración. Para calcular el número de ambientes que debíamos seleccionar, utilizamos la información que provea el inventario del RMU acerca de la cantidad total de metros lineales por ambiente. Calculamos un tamaño muestral aproximado tomando en cuenta la varianza de metros lineales por ambiente. Decidimos seleccionar 20 y 23 ambientes para las iteraciones 1 y 2, respectivamente.

A continuación, registramos la cantidad de metros lineales por ambiente y creamos una variable de proporción acumulada ( $p_i$ ). Para todos los ambientes asignamos intervalos mutuamente excluyentes entre 0 y 1, de la siguiente manera:

$$(0, p_1], \dots, (p_{i-1}, p_i], \dots, (p_{k-1}, p_k], \text{ con } p_k = 1, \quad (1)$$

La selección se llevó a cabo sin reemplazo y con una probabilidad de selección basada en estos intervalos hasta que un número específico de ambientes hubiesen sido seleccionados. Desde el principio consideramos el beneficio de utilizar este método debido a la posterior dificultad de calcular la probabilidad de selección porque los intervalos no eran de la misma amplitud. Por lo tanto, a la altura de la iteración 3, cambiamos el procedimiento de selección de los ambientes.

En las iteraciones 3 a 9, en lugar de seleccionar números aleatorios hasta que un número previamente determinado de ambientes fuese seleccionado, establecimos una cantidad fija de números aleatorios (33) extraídos de una distribución estándar y uniforme. Tal como se señaló anteriormente, la selección de ambientes se determinó por el intervalo dentro del cual cayó cada número aleatorio (tal como se definió en el numeral 1). Se escogieron treinta y tres números aleatorios con la expectativa de que este método pudiera dar paso a un número similar de ambientes únicos en la muestra, como los de las iteraciones 1 y 2 (en que se seleccionaron 20 y 23 respectivamente). Tal como esperábamos, seleccionamos 22 ó 23 ambientes únicos para las iteraciones 3 a 9.

### 5.2 Etapa 2: Contenedores

Para cada iteración se distribuyeron trescientos puntos entre los ambientes seleccionados. Dentro de cada ambiente se seleccionaron contenedores con una probabilidad conocida. Con base en pruebas llevadas a cabo durante el estudio piloto, se escogieron trescientos puntos por iteración. Dado el tamaño del equipo de trabajo, el muestreo de documentos de 300 puntos nos tomó entre 2 y 3 semanas. Sobre la base del conocimiento de nuestros colegas, un período de tres semanas se consideró lo suficientemente corto para evitar importantes traslados de documentos entre las etapas de muestreo.

Estos 300 puntos se distribuyeron entre los ambientes (A) seleccionados en la etapa anterior, tal como se presenta a continuación:

Si  $Ppa$  indica “Puntos por ambiente”, donde:

$$Ppa_i = \frac{300 \times \text{metros lineales } E_i}{\sum \text{metros lineales } E_i} \quad (2)$$

Cada punto representa un contenedor a ser seleccionado. En esta etapa los contenedores son seleccionados con reemplazo proporcional al tamaño, basado en la combinación de puntos por ambiente y midiendo el tamaño del contenedor en metros lineales. El número de veces que se muestrea el mismo contenedor determina el número de UUA que deberán ser muestreadas dentro de ese contenedor en la etapa subsiguiente.

Se hubieran obtenido ponderaciones más parejas si hubiese habido el mismo número de contenedores muestreados en cada ambiente muestreado; sin embargo, decidimos utilizar la probabilidad proporcional al tamaño debido a la potencialmente drástica diferencia de tamaños de los contenedores. Muchos ambientes abarcan muy pocos contenedores, cada uno de los cuales contiene grandes cantidades de papel.

A cada uno de los equipos de trabajo de campo del Archivo se le proporcionó una lista de contenedores seleccionados y el número de UUA que debía extraer de cada uno. Con el objeto de seleccionar la muestra en las últimas dos etapas, el equipo de trabajo de campo tuvo que ubicar los ambientes y los contenedores seleccionados. Una vez que los contenedores muestreados fueron localizados, se completaron las últimas dos etapas de selección.

La Tabla 2 es un ejemplo extraído de una de las listas proporcionadas al equipo de trabajo de campo después de que las etapas uno y dos fueran seleccionadas. Esta clase de lista era el 'mapa' que servía de guía al equipo de trabajo para encontrar el lugar donde debían muestrear las UUA de los contenedores.

**Tabla 2:** Ejemplo de algunos contenedores y del número de UUA que debía extraer cada uno de los equipos de trabajo de campo

Edificio No.	Ambiente No.	Tipo de Contenedor	Contenedor No.	Número de Puntos a Seleccionar
3	10	Plataforma	11	4
2	2	Mueble de archivo	49	1
2	1	Mesa	2	2
4	1	Librera	7	1
4	1	Librera	22	2
1	2	Suelo	1	2
2	2	Librera	63	2
6	1	Piso	1	23

Llegado este punto, es útil resaltar que el muestreo de las etapas uno y dos se produce en un momento diferente en el tiempo al de las etapas tres y cuatro. El traslado de documentos



que se produjo durante este período de tiempo fue uno de los principales retos y será el tema de debate del apartado 6.

### **5.3 Etapa 3: Últimas unidades de agregación (UUA)**

Frecuentemente los contenedores tienen documentos en cajas, gavetas, carpetas o legajos atados; todos estos se denominan “últimas unidades de agregación”. Con la lista de contenedores y el número de UUA a extraer de cada uno, el equipo de trabajo de campo llega a la ubicación de muestreo e inicia el procedimiento de selección para la etapa tres.

Dado que no había un marco muestral enumerativo disponible para los contenidos de cada contenedor, el equipo de trabajo midió las dimensiones extremas del papel dentro del contenedor -su altura, anchura y profundidad- y utilizó estas medidas como su marco muestral. Con el objeto de extraer la muestra, los miembros del equipo de trabajo de campo multiplican cada dimensión por un número aleatorio uniforme para determinar una coordenada aleatoria al interior del contenedor. Esta coordenada aleatoria podría estar localizada dentro de una última unidad de agregación o en un espacio vacío. Si el equipo de trabajo de campo encuentra un espacio vacío en esa coordenada, lo anota como un “intento vacío” y repite el proceso hasta encontrar una UUA. Más adelante se utilizaron los intentos vacíos para aproximarse a la cantidad total de espacio vacío dentro de un contenedor, tema que se abordará con mayor detalle en Shapiro et al. [1], en el segundo documento de esta serie, que trata acerca de la estimación de las ponderaciones.

En los casos en que los contenedores sólo contienen cajas, se utilizó un método de muestreo diferente. En estos casos se seleccionó una caja aleatoriamente basándose en el número total de cajas en un contenedor. Cada caja tiene la misma probabilidad de selección.

El tamaño de una UUA es su volumen medido en centímetros cúbicos y no en metros lineales, como se hizo en las dos primeras etapas de selección. El equipo de trabajo de campo extrajo el número de UUA requeridas por el punto de distribución durante la etapa previa. Se seleccionaron 300 UUA para cada iteración a partir de los espacios de tres dimensiones al interior de los contenedores, tal como se describió anteriormente en el epígrafe 3.

Una ventaja del muestreo realizado de conformidad con el espacio de tres dimensiones es que se evita manipular los documentos más allá de lo estrictamente necesario. Los únicos momentos en que los miembros del equipo de muestreo tocan los frágiles documentos es cuando llegan al lugar donde deben extraer una UUA y cuando extraen las unidades de información del interior de la UUA (etapa cuatro).

### **5.4 Etapa 4: Unidades de información (UI)**

Una “unidad de información” se define como un conjunto de documentos que han sido archivados juntos por los propietarios del sistema original de archivo y tienen un tema, caso o fenómeno en común. Como ejemplo, una unidad de información puede ser un solo documento o un conjunto de documentos que conforman un expediente. En el nivel más agregado, la unidad de información es aquella que al final es muestreada y estudiada por el AHPN. La razón por la cual utilizamos unidades de información como la última etapa de

selección, en lugar de usar documentos, responde a la idea de respetar los contenidos del Archivo tal y como fueron archivados.

Las unidades de información están agrupadas juntas dentro de las UUA. La selección de UI se basó en la altura de una UUA medida en milímetros lineales. El equipo de trabajo de campo multiplicó la altura de una UUA muestreada por un número aleatorio, con el objeto de determinar un punto milimétrico a partir del cual seleccionar una UI. El equipo luego extrae la UI en el punto milimétrico seleccionado junto a las siguientes dos UI consecutivas y codifica las tres. Creemos que al extraer tres fue posible calcular las similitudes o diferencias del contenido de la UI dentro de un grupo de documentos. No deseábamos seleccionar más de tres, debido a que un grupo de gran tamaño hubiera podido incrementar el error muestral. El tamaño total de la muestra de UI por iteración es de  $300 \text{ puntos} \times 3 \text{ UI} = 900 \text{ UI}$ . Nótese que con este procedimiento es imposible saber con exactitud la probabilidad de selección de cada UI, pues la probabilidad es una función del tamaño de las UI precedentes, por lo que usualmente no disponemos de esa información. Shapiro, et al. [1] describe los retos que implica calcular la probabilidad de selección para las tres UI. Esto se corrigió en la iteración 11.

## 6. El reto: el movimiento

Al seleccionar pequeñas muestras cada 2 ó 3 semanas, el movimiento de documentos durante una iteración fue mínimo inicialmente. Como consecuencia de ello, la probabilidad de selección durante este tiempo fue lo suficientemente estable.

Cinco meses después de iniciar las muestras iterativas, el movimiento de documentos aumentó debido a una serie de acontecimientos. Esto significó que la transición de la etapa dos a la tres fue cada vez más difícil.

No se movían los documentos de manera consistente: algunas veces todos los documentos eran trasladados de un contenedor a otro; en otras ocasiones los documentos de un contenedor se separaban entre varios contenedores; con menor frecuencia se unían varios contenedores en uno o se mezclaban con otros documentos que no se encontraban en el contenedor originalmente muestreado. El equipo del Archivo se esmeró en dar seguimiento a los papeles, pero muy frecuentemente resultaba difícil mantener el control sobre los traslados. La Tabla 3\* muestra que aproximadamente el 80% de los contenedores no sufrieron traslado alguno o el movimiento se dio de un contenedor a otro y, por lo tanto, no causó dificultades al calcular las probabilidades de selección. El restante 20% fue un reto mayor.

---

\* **Nota de la Traductora:** el documento original dice “Tabla 2”, sin embargo el contexto indica que se está haciendo referencia a la Tabla 3 por lo que se ha introducido el cambio correspondiente.

**Tabla 3:** Movimiento de Contenedores

Desde	Hacia	Frecuencia	Porcentaje
Sin movimiento		772	65
1	1	189	16
Muchos	1	102	9
1	Muchos	61	5
Muchos	Muchos	9	1
Desconocido		48	4

### 7. Evaluación y cambios

Después de la novena iteración, los cambios de ubicación, contenidos y tamaño de los contenedores ocurridos en el lapso transcurrido entre la selección de los contenedores y la selección de las UUA amenazaron con dificultar demasiado el cálculo de selección de probabilidades. Durante aproximadamente un mes, hicimos una pausa en el proceso con el fin de mejorar el diseño muestral.

Para la décima iteración, modificamos el diseño muestral hacia sólo tres etapas muestrales. Para empezar, calculamos el volumen completo del Archivo físico (el volumen de los edificios, no del papel). Del volumen total de los edificios del Archivo nos aproximamos a una proporción de 'espacios ocupados' donde:

$$\text{espacio ocupado} = \frac{\text{m}^3 \text{ papel}}{\text{m}^3 \text{ archivo total}} \quad (3)$$

Asumimos que las coordenadas en 3D en espacios ocupados contienen UUA. Sabíamos que el objetivo era muestrear un número de 300 UUA (en tanto habíamos muestreado 300 UUA en cada una de las iteraciones de la 1 a la 9). Con el fin de muestrear este número de UUA, debíamos sobre-muestrear las coordenadas en 3D ( $n$ ), en tanto sólo una proporción del Archivo tiene espacios ocupados.

$$n = \left\lceil \frac{300}{\text{espacio ocupado}} \right\rceil \quad (4)$$

Al igual que en las iteraciones 1 a 9, fue difícil conseguir las UUA directamente. Primero tuvimos que asignar las coordenadas  $n$  3D en los ambientes del Archivo. Utilizamos el mismo mecanismo descrito en el apartado 5.1, con la única diferencia de que, en vez de usar 33 números aleatorios, utilizamos  $n$  números aleatorios.

$n_j$  es la fracción de todas las coordenadas  $n$  3D que se obtendrán de cualquier ambiente  $j$ . De modo que:

$$n = \sum_j n_j \quad (5)$$

Muchas coordenadas 3D conducirán a espacios 'vacíos', pero el número de intentos vacíos formará parte de los datos utilizados para calcular la probabilidad de selección de las UUA. Esta modificación al diseño redujo eficazmente los retos planteados por los traslados, ya que no dependeremos nunca más de que los contenidos del Archivo permanezcan en el mismo lugar entre una etapa de muestreo y otra que ocurre tras un lapso de tiempo.

Eliminar una etapa trajo como consecuencia un diseño más eficaz. Además, las probabilidades de selección deberían ser mucho más uniformes que las del diseño original. No sólo se modificó el diseño de la muestra para seleccionar las UUA, sino también el procedimiento de muestreo para las UI. Decidimos seleccionar 10 UI de cada UUA, en lugar de 3, que es como se llevó a cabo en las iteraciones previas. Además, las UI seleccionadas no fueron todas consecutivas. El equipo de campo seleccionó un punto dentro de la UUA, como antes, y desde ese punto escogieron 6 UI consecutivas. Luego, dividieron la altura restante de la UUA en cuatro partes iguales, seleccionando una UI de cada uno de esos cuartos. Esto proporciona la oportunidad de observar más rigurosamente el tema de la homogeneidad/heterogeneidad dentro de una UUA completa y no sólo de UI consecutivas.

### 8. El estado del muestreo

Hasta abril de 2009 se habían muestreado 20,000 documentos. Aproximadamente 8,000 formaban parte de la muestra de las iteraciones 1 a 9 en las que se utilizó el diseño muestral descrito en este documento. Las otras 12,000 se recogieron durante una décima iteración mucho más larga, utilizando para ello el subsecuente diseño descrito en el epígrafe 7.

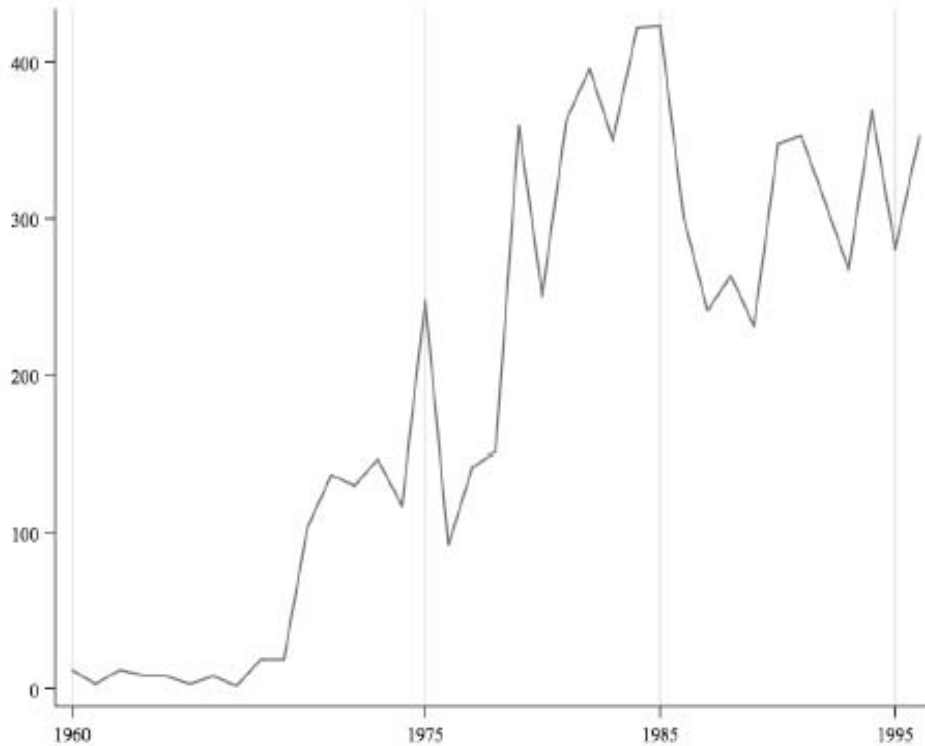
Las tablas que se presentan a continuación muestran estadísticas descriptivas acerca de los documentos muestreados en las iteraciones 1 a 9. La Tabla 4 presenta el número total de unidades de información (UI) y documentos muestreados que fueron elegibles para el estudio. Una UI se considera elegible si al menos uno de los documentos que contiene fue escrito entre 1960 y 1996. Los documentos *relacionados* con el período de estudio incluyen todos los documentos en una UI elegible; sin embargo, algunos pudieron no haber sido escritos en el período de tiempo comprendido en esta investigación. Todos estos datos se utilizarán para calcular las estimaciones ponderadas, en el tercer documento de esta serie.

**Tabla 4:** Resumen General de los Datos Muestreados

Categoría	Frecuencia
Unidades de información	5371
Documentos <i>relacionados</i> con 1960-1996	8162
Documentos escritos entre 1960 y 1996	7241
Documentos con hechos	1806
Documentos con políticas	90
Número de hechos mencionados	3699
Número de políticas mencionadas	310

La Gráfica 3 representa el año de creación de todos los documentos de la muestra. Se observa un incremento gradual en la producción de documentos luego de mediados de los 70. El año del documento será utilizado para analizar cuándo se crearon los documentos relacionados con ciertos temas, tales como hechos de interés (listados a continuación), y para saber si tienen correlación con períodos de violencia en Guatemala.

**Gráfica 3:** Documentos muestreados, por año en que fueron escritos



La Tabla 5 enumera, para la investigación cuantitativa, los “hechos de interés” relacionados con violaciones a los derechos humanos recogidas en el Archivo. Tal como muestra la Tabla 5, la *Privación de Libertad* y las *Muertes* son, con diferencia, los tipos de hechos de interés más comúnmente registrados y la magnitud de sus frecuencias sobrepasa un gran número de veces las frecuencias de *Interrogatorios* y *Vigilancia*. No está claro si la amplitud de las diferencias de frecuencias en la Tabla 5 refleja diferencias entre (1) el proceso de documentación o (2) las frecuencias reales de los hechos mismos.

En el futuro, trataremos de utilizar un muestreo adaptativo para la selección de documentos para tipos de hechos de interés poco frecuentes.

**Tabla 5:** Hechos de Interés para el Equipo de Investigación del AHPN

Hechos	Frecuencia	Porcentaje de todos los hechos
Detenciones	2375	64
Muertes	596	16
Denuncias	203	5
Desapariciones	179	5
Intimidación	138	4
Abuso físico	93	3
Exhibición personal	39	1
Ingreso ilegal a propiedad privada	28	1
Abuso sexual	27	1
Interrogatorio	13	< 1
Vigilancia	8	< 1

## 9. El trabajo futuro

Han pasado cuatro años desde el descubrimiento del Archivo. Se han limpiado amplias secciones y se han organizado y reunido documentos que comparten las mismas características. En el futuro planeamos desarrollar y utilizar muestreos adaptativos u otras técnicas de muestreo para investigar documentos de interés de forma más deliberada y probabilística. Una opción puede ser muestrear por tipo de documento. Por ejemplo, podríamos estar en capacidad de muestrear libros de registro de correspondencia que pueden proveer una línea base acerca del flujo de comunicación, lo cual luego podríamos comparar con tipos específicos de documentos, tales como los documentos confidenciales. Otra opción podría ser muestrear las fichas de identificación que contienen información que la policía recogió acerca de individuos; lo cual, con mucha frecuencia, resulta ser un indicador de vigilancia y control social. Al concluir la iteración 11, tomaremos una pausa y evaluaremos el camino que queda por delante.

## 10. Conclusión

A pesar de que el muestreo fue desafiante y exigió el uso de muchos recursos, desde el momento en que se empezó a trabajar, el AHPN se ha comprometido a llevar a cabo una investigación científica y rigurosa de los documentos del Archivo. Tal como se expuso anteriormente, los documentos muestreados contienen información importante relacionada con las preguntas de investigación que tenemos acerca de temas de interés, tales como hechos de violencia y políticas administrativas. Después de presentar, en el segundo documento de esta serie de tres, los cálculos de la ponderación basados en nuestro diseño muestral, el tercer informe planteará las estimaciones directamente relacionadas con las preguntas de investigación que motivaron el estudio.

## Agradecimientos

El peso del trabajo de campo, que incluye la recolección de datos, medición y tabulación de información, se llevó a cabo bajo la dirección de Jorge Villagrán y su equipo en el AHPN. El profesor Dr. David Banks puso a nuestra disposición su invaluable consejo durante la fase inicial piloto. El doctor Fritz Scheuren, Estadístico Vicepresidente del Centro Nacional de Estudios de Opinión de la Universidad de Chicago (*NORC*, por sus siglas en inglés), revisó el diseño de la muestra durante las primeras etapas del proyecto.

## Referencias

- [1] Shapiro, G., Guzmán, D. y Zador, P. y Guberek, T. (2009), “Ponderación de la Muestra del Archivo Histórico de la Policía Nacional de Guatemala: Retos y Problemas Inusuales”, en *JSM Proceedings*, Sección de Métodos de Investigación Pericial (Survey Research Methods). Alexandria, VA: Asociación Americana de Estadística.

Traducción del inglés al español: Carla Villagrán, traductora jurada  
Traducción revisada por Beatriz Bejarano, Benetech.

---

---

## La Ponderación de la Muestra del Archivo Histórico de la Policía Nacional de Guatemala: Problemas y Desafíos Inusuales

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

### Resumen

El presente informe describe los procesos de ponderación utilizados para el muestreo de los registros del Archivo Histórico de la Policía Nacional de Guatemala (AHPN). Las decisiones tomadas en cuanto a los procedimientos de muestreo que permitieran lidiar con la estructura del Archivo desembocaron en una serie de problemas a la hora de calcular las ponderaciones. Lo primero y principal fue que el universo del cual se extrajeron las muestras es extremadamente inusual y fluido y se posee escaso conocimiento acerca de sus dimensiones. En más de una de las etapas de selección, las probabilidades de selección eran desconocidas, lo que requirió que las ponderaciones se basaran en estimaciones de las probabilidades de selección. Además, surgieron problemas derivados de cuestiones operativas, tales como el movimiento de los documentos mientras se llevaba a cabo el muestreo y la existencia de espacios vacíos en los contenedores.

**Palabras clave:** muestreo probabilístico, ponderaciones, post-estratificación, recorte de ponderación, Archivo Histórico de la Policía Nacional de Guatemala, derechos humanos.

### 1. Antecedentes e introducción

Este documento, el segundo de una serie de tres, describe la ponderación realizada en las primeras nueve iteraciones de muestreo de los documentos del Archivo Histórico de la Policía Nacional de Guatemala (AHPN). Los documentos del Archivo fueron almacenados desordenadamente, lo cual significó grandes desafíos, tanto para seleccionar las muestras como para realizar los posteriores cálculos de las ponderaciones y justificar apropiadamente las decisiones tomadas en cuanto al diseño muestral. Como se describe en el primer informe [1], los documentos son potencialmente útiles para estudiar el papel de la Policía Nacional en los actos de violencia cometidos durante el conflicto armado interno en Guatemala entre 1960 y 1996.

El siguiente acápite del documento presenta los cálculos de las ponderaciones para cada una de las cuatro etapas de muestreo descritas en Guzmán, et al. [1]: ambiente, contenedor, última unidad de agregación (UUA) y unidad de información (UI). Los análisis que se presentarán en el siguiente documento de esta serie [2] se llevaron a cabo a nivel de documentos (las UI pueden contener más de un documento). No obstante, ya que los documentos fueron seleccionados con certidumbre, una vez se hubo extraído una UI, las unidades de información constituyen la última etapa en la que se calculan las ponderaciones. En el apartado 2 se presentan los inusuales problemas que hubo que

---

\* Gary Shapiro, recientemente jubilado, fue estadístico principal de Westat durante la mayor parte del tiempo en que contribuyó con este estudio. Paul Zador, PhD., es estadístico principal de Westat. Ambos, Shapiro y Zador, son miembros del Grupo Voluntario de Intereses Especiales de la Asociación Americana de Estadística. Daniel Guzmán y Tamy Guberek son consultores del Grupo de Análisis de Datos sobre Derechos Humanos de la organización Benetech (*Human Rights Data Analysis Group, HRDAG*, por su nombre y sigla en inglés). Kristian Lum es candidata al doctorado del Departamento de Ciencias Estadísticas de la Universidad de Duke. [www.hrdag.org](http://www.hrdag.org), [www.benetech.org](http://www.benetech.org).



enfrentar para determinar las ponderaciones de los ambientes en las iteraciones 1 y 2, así como las ponderaciones base de la UUA y las UI. El capítulo 3 trata el problema especial y adicional ocasionado por el traslado de los documentos durante el proceso de muestreo. El acápite 4 presenta una medida de ponderación adicional que podría implementarse en un futuro. La sección 5 trata de la calidad de los datos y de los datos de imputación. Por último, en el apartado 6 resumiremos brevemente lo que ha ocurrido desde el final de la iteración 9. Este documento se centra particularmente en los problemas especiales de la ponderación como se plantea en los acápites 2 y 3.

Para el estudio muestral se utilizó el enfoque de muestreo tradicional de Horvitz-Thompson, con ponderaciones base iguales al inverso de la probabilidad de selección o, en algunos casos, a los cálculos aproximados de la probabilidad de selección. Para todas las etapas de la selección la medida del tamaño es la altura en metros lineales, exceptuando la tercera etapa (UUA), en la que se empleó el volumen.

Se realizó una selección inicial de nueve iteraciones de muestreo. Con el objetivo de simplificar, en la Sección 2.5 plantearemos la ponderación de una sola iteración, con comentarios acerca de las otras.

## 2. Ponderaciones base

El enfoque de ponderación para el muestreo es el tradicional (descrito en todos los libros de texto sobre encuestas muestrales), las ponderaciones base son iguales al inverso de la probabilidad de selección y se aplican ajustes de post-estratificación apropiados a las ponderaciones base. Como se explica en Guzmán, et al. [1], la medida de tamaño para las dos primeras etapas es la altura, la medida de tamaño para la tercera etapa de selección es el volumen, y para la última etapa de selección, la altura.

### 2.1 Ponderaciones de los ambientes

Para las iteraciones 3-9, el cálculo de la ponderación fue directo. Utilizando 33 números aleatorios (como se explica en [1]), seleccionamos 22 ó 23 ambientes por iteración, con una probabilidad proporcional a los metros lineales de papel contenidos en el ambiente, con reemplazo. Dada la probabilidad de no escoger  $A_i$ , el ambiente “ $i$ ” sería  $(1 - p_i)$ , lo cual resultaba en la siguiente probabilidad de selección para  $A_i$ :

$$Pr(A_i) = 1 - (1 - p_i)^{33} \quad (1)$$

en donde:

$$A_i \text{ es el ambiente } i, \text{ y } p_i \text{ es } = \frac{\text{Metros lineales } A_i}{\sum \text{Metros lineales } A_i}$$

Los metros lineales de  $A_i$  es la medida del tamaño para  $A_i$ . Por lo tanto, la ponderación base del ambiente para  $A_i$  es  $P(A_i) = 1/Pr(A_i)$ .

Para las iteraciones 1 y 2, también muestreamos con reemplazo, pero el número de ambientes únicos en la muestra se ajustó, en vez de hacerlo con la cantidad de números aleatorios utilizados para seleccionar los ambientes muestreados. Para la iteración 1 usamos 25 números aleatorios para obtener 20 ambientes únicos seleccionados. Para la iteración 2

fijamos en 23 el número de ambientes únicos, pero tuvimos que seleccionar una muestra de 58 números aleatorios para obtener 23 ambientes únicos. Esta metodología de muestreo dificultó el cálculo directo de las ponderaciones. Se utilizó una simulación Monte Carlo y la selección de la muestra se repitió 10,000 veces. Con base en el número de veces que los ambientes de nuestras muestras aparecían en las muestras simuladas, calculamos las probabilidades estimadas de selección de un ambiente. La ponderación base del ambiente para estas iteraciones es también el inverso de la probabilidad de selección.

## 2.2 Ponderación de los contenedores

Un contenedor puede ser una estantería, un mueble de archivo, la superficie de una mesa, un saco, una plataforma de madera, un espacio sobre el suelo u otro objeto localizado dentro de un ambiente. Para un ambiente dado en una iteración determinada, se muestreó un número ya fijado de puntos, tal como se explica en Guzmán, et al. [1]. Cada punto representa un contenedor que debe seleccionarse y el mismo contenedor puede ser seleccionado más de una vez. El número de contenedores seleccionados en cada ambiente se determina de acuerdo con el tamaño del ambiente. El diseño muestral óptimo hubiese consistido en muestras de igual tamaño para cada ambiente seleccionado en vez de muestras con un tamaño proporcional a la medida del tamaño del ambiente. De este modo hubiésemos obtenido ponderaciones menos variables si hubiese habido el mismo número de contenedores muestreados en cada ambiente muestreado. Utilizando PPT (probabilidad proporcional al tamaño) para seleccionar los ambientes debimos haber resuelto el problema de las drásticas diferencias de tamaño de los contenedores; sin embargo, en las primeras etapas del muestreo estábamos preocupados de que el uso de PPT en la primera etapa no estimara adecuadamente las diferencias en los tamaños de los contenedores, y por lo tanto también escogimos los contenedores de acuerdo con PPT.

Para calcular la probabilidad condicional de seleccionar un contenedor  $j$  en un ambiente  $i$ , utilizamos:

$$Pr(C_{ij}) = \text{Puntos}_i \frac{\text{Metros lineales } C_{ij}}{\sum_j \text{Metros lineales } C_{ij}} \quad (2)$$

donde:

$C_{ij}$  = Contenedor  $j$  en el Ambiente  $i$ ,

$\text{Puntos}_i$  = Número de puntos seleccionados en el ambiente  $i$ , y

Los metros lineales  $C_{ij}$  es el número de metros lineales en el contenedor  $C_{ij}$ .

La ponderación condicional del contenedor será entonces  $P(C_{ij}) = \frac{1}{Pr(C_{ij})}$ .

## 2.3 Ponderación de la Última Unidad de Agregación (UUA)

Debido a nuestra incapacidad de saber o estimar una probabilidad de selección confiable, se presentaron problemas particulares al determinar las ponderaciones para las Últimas Unidades de Agregación (UUA). Una UUA es un grupo de documentos archivados juntos en el Archivo. Ejemplos de UUA son carpetas, gavetas, paquetes de documentos atados juntos, cajas con documentos archivados juntos y muchos otros. Se extrajo una única UUA

por cada contenedor seleccionado. Sin embargo, debido a que en los pasos previos los contenedores fueron seleccionados con reemplazo, es posible que se hayan extraído múltiples UUA independientemente de un mismo contenedor.

El muestreo de las UUA se hizo basándose en coordenadas aleatorias en un espacio de tres dimensiones dentro de los contenedores, tal como se define en Guzmán, et al. [1]. Por lo tanto, a diferencia de las dos primeras etapas, que se fundamentaron en metros lineales, este muestreo se basó en el volumen. Como lo indican Guzmán et al., estas coordenadas aleatorias pueden corresponderse con un espacio ocupado o vacío dentro del contenedor. Por consiguiente, para un contenedor que fue seleccionado una vez, una UUA puede ser seleccionada después de uno o varios intentos (si los primeros intentos correspondieron a espacios vacíos). La verdadera probabilidad condicional de selección de una UUA dentro de un contenedor es una función directa de la proporción de espacio que está ocupado. No obstante, no estábamos en condiciones de estimar directamente esta proporción. El número de intentos requeridos para seleccionar un espacio ocupado aporta una estimación de esta proporción; pero cuando sólo hay uno o un número pequeño de UUA para muestrear, la estimación resulta un tanto pobre.

Para superar esta falta de información acerca de la proporción de espacio ocupado en cada contenedor, utilizamos un modelo jerárquico bayesiano para tomar prestada información de varios contenedores. Suponemos que el número de intentos en espacios vacíos para el contenedor “ $i$ ”,  $\text{vacío}_i$ , proviene de una distribución binomial negativa,  $\text{BN}(\text{vacío}_i; \text{requerido}_i, p_i)$ , en donde  $\text{requerido}_i$  es el número de UUA requeridas de cada contenedor “ $i$ ”, antes de que podamos dejar de hacer extracciones, y  $p_i$  es la probabilidad desconocida de seleccionar una UUA de un contenedor  $i$ . También suponemos que todos los  $p_i$  provienen de una distribución común Beta, con parámetros  $\alpha$  y  $\beta$ . Con la condición de que en  $\alpha$  y  $\beta$ , cada uno de los  $p_i$  se distribuyan  $\text{Beta}(\alpha + \text{requerido}_i, \beta + \text{vacío}_i)$  lo cual significa:

$$\frac{\text{requerido}_i + \alpha}{\text{requerido}_i + \alpha + \text{vacío}_i + \beta}$$

Nótese que el significado de cada uno de los  $p_i$  contiene parámetros comunes  $\alpha$  y  $\beta$  y, por lo tanto, comparten información a través de los contenedores. Para terminar la especificación del modelo, ubicamos priores Gamma imprecisos en  $\alpha$  y  $\beta$ , y corremos un Muestreo de Gibbs simple para muestrear desde la distribución de los  $p_i$ ,  $\alpha$  y  $\beta$ . Utilizamos el significado posterior de cada  $p_i$  como la probabilidad de extraer una UUA del contenedor “ $i$ ”, a partir de lo cual podemos fácilmente calcular el número esperado de intentos vacíos que ocurrirán antes de que extraigamos el número necesario de UUA.

El procedimiento da como resultado ponderaciones que sólo están indirectamente relacionadas con la verdadera probabilidad de selección de una UUA.

La probabilidad condicional de selección de una UUA es:

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

donde:

$V_{ij}$  = Volumen del contenedor  $C_{ij}$ , espacios vacíos incluidos

$V_{ijk}$  = Volumen de la UUA  $k$  en el contenedor  $C_{ij}$ ,

$S_{ij}$  = Estimación empírica de Bayes acerca de la proporción de espacio ocupado en el contenedor  $C_{ij}$ ,

y,

$H_{ij}$  = número de veces que se selecciona  $C_{ij}$  en la segunda etapa de muestreo.

La ponderación condicional de UUA es entonces,  $W(P_{ijk}) = \frac{1}{Pr(L_{ijk})}$ .

Desafortunadamente, el examen de las ponderaciones finales inicialmente formuladas reveló que este procedimiento aún estima inadecuadamente la probabilidad de selección de la UUA. Esto se determinó comparando la estimación total de metros lineales de papel con el dato conocido del total de metros lineales de papel en todo el Archivo, a partir del inventario inicial. Los metros lineales estimados para el Archivo en su conjunto se calcularon sobre la acumulación de metros lineales basada en las ponderaciones de las UUA. Esta comparación demostró que estábamos sobreestimando los metros lineales sistemáticamente, lo cual implicaba que todavía no teníamos una manera apropiada de dar cuenta de los espacios vacíos excluidos. Los ajustes adicionales a estas ponderaciones, basados en estos hallazgos, se presentarán en el apartado 2.6.

Cabe destacar que, dado que las ponderaciones de las UUA se realizan en volumen, en tanto las ponderaciones para los ambientes y contenedores se dan en metros lineales, la cancelación de términos deseada que se presenta cuando las ponderaciones condicionales se multiplican juntas no sucede en este caso. Cuando no se da la cancelación, hay más varianza en la ponderación.

Hay casos especiales en que las ponderaciones de las UUA se calcularon más directamente. Dada la necesidad del Proyecto del AHPN de limpiar y organizar los documentos, un subconjunto del Archivo se fue trasladando gradualmente hacia cajas de archivo de igual tamaño. Calculamos que si un contenedor seleccionado contuviera sólo este tipo de cajas, la probabilidad de seleccionar una UUA (en este caso una caja) sería de una más sobre el número total de UUA (cajas) en el contenedor.

## **2.4 Ponderación de la Unidad de Información (UI)**

También fue difícil determinar la probabilidad de selección de las Unidades de Información (UI). Una UI es un conjunto de documentos archivados juntos por los dueños del sistema original de archivo y relacionados con un tema, caso o fenómeno que les es común. Una UI puede ser un documento individual o un conjunto de documentos que conforman un expediente de archivo. Realizamos el muestreo basándonos en la altura de una UUA muestreada. Multiplicando un número aleatorio por la altura de la UUA (en milímetros lineales), se identificó un punto aleatorio en esta sola dimensión. Se muestreó la UI localizada en ese punto. Las dos unidades de información siguientes y consecutivas se incluyeron también en la muestra. Aunque sea clara la probabilidad de selección de un grupo específico de 3 UI, la probabilidad de selección de una sola UI es más difícil de determinar.

Considerando la siguiente selección de UI en una UUA: Y, Z, A, B y C, donde A es la UI inicialmente seleccionada en una muestra específica y por consiguiente, A, B y C se incluirían en la muestra. Y y Z son las UI que preceden a A en la UUA.

$\Pr(C \text{ en la muestra}) = \Pr(A \text{ seleccionada}) + \Pr(B \text{ seleccionada}) + \Pr(C \text{ seleccionada})$ , donde  $\Pr(A \text{ seleccionada})$  es la probabilidad de selección inicial de UI A,  $\Pr(B \text{ seleccionada})$  es la probabilidad de selección inicial de UI B y,  $\Pr(C \text{ seleccionada})$  es la probabilidad de selección inicial de UI C. Esto se calcula fácilmente a partir del conocimiento de las alturas de A, B y C.

Sin embargo, la probabilidad de selección de esa primera UI es una función no sólo de su altura sino también de la altura de las 2 UI precedentes, dado que éstas también pudieron atraer a A a la muestra.

$$\begin{aligned} \Pr(A \text{ en la muestra}) &= \Pr(Y \text{ seleccionada}) + \Pr(Z \text{ seleccionada}) + \Pr(A \text{ seleccionada}) \\ \Pr(B \text{ en la muestra}) &= \Pr(Z \text{ seleccionada}) + \Pr(A \text{ seleccionada}) + \Pr(B \text{ seleccionada}) \end{aligned}$$

Aunque la probabilidad de seleccionar A, B y C esté clara, desafortunadamente no se obtuvo información acerca de la altura de las precedentes Y y Z. Calculamos la altura de Y y Z como un promedio de la altura de las 3 UI (A, B y C) que estaban en la muestra. La probabilidad de seleccionar Y y Z puede entonces calcularse fácilmente, de modo que pueda estimarse la probabilidad de que A, B y C estuvieran en la muestra (como se planteó anteriormente).

Como justificación de esta sustitución, se calculó el número promedio de páginas para cada UI que se seleccionó como primera, segunda y tercera en la muestra. Dado que el análisis de varianza determinó que en términos estadísticos, estos tres promedios no eran sustancialmente diferentes el uno del otro, nos sentimos cómodos al sustituir la altura promedio de las UI seleccionadas por las alturas desconocidas de Y y Z.

Las probabilidades condicionales para  $UI_{ijk}$  se calcularon así:

$$\Pr(UI_{ijk1}) = \frac{\# \text{ páginas}(UI_{ijk1})}{\# \text{ páginas}(UUA_{ijk})} + 2 \frac{\sum_{r=1}^3 \# \text{ páginas}(UI_{ijkr})/3}{\# \text{ páginas}(UUA_{ijk})} \quad (4)$$

$$\Pr(UI_{ijk2}) = \frac{\# \text{ páginas}(UI_{ijk1})}{\# \text{ páginas}(UUA_{ijk})} + \frac{\# \text{ páginas}(UI_{ijk2})}{\# \text{ páginas}(UUA_{ijk})} + \frac{\sum_{r=1}^3 \# \text{ páginas}(UI_{ijkr})/3}{\# \text{ páginas}(UUA_{ijk})} \quad (5)$$

$$\Pr(UI_{ijk3}) = \frac{\# \text{ páginas}(UI_{ijk1})}{\# \text{ páginas}(UUA_{ijk})} + \frac{\# \text{ páginas}(UI_{ijk2})}{\# \text{ páginas}(UUA_{ijk})} + \frac{\# \text{ páginas}(UI_{ijk3})}{\# \text{ páginas}(UUA_{ijk})} \quad (6)$$

donde  $i$  indexa los ambientes,  $j$  indexa los contenedores,  $k$  indexa las UUA y  $r$  indexa las UI.

El número esperado de páginas por UUA,  $\# \text{páginas}(UUA_{ijk}) = 11.17 \times \text{Altura}(UUA_{ijk})$  se basó en el conocimiento y la información empírica del Archivo, obtenida a partir de nuestra muestra. La altura se midió en milímetros.

Entonces, la ponderación condicional de la UI es  $P(UI_{ijkr}) = \frac{1}{Pr(UI_{ijkr})}$ .

## 2.5 La Combinación de las Ponderaciones a través de las Iteraciones y las Ponderaciones Finales

La ponderación final para una sola iteración, con excepción de los ajustes por recortes y post-estratificación, es el producto de cada una de las ponderaciones expuestas en los apartados previos, dividida por el número de iteraciones:

$$P_{ijkr} = \frac{P(A_i) \times P(C_{ij}) \times P(UUA_{ijk}) \times P(UI_{ijkr})}{9} \quad (7)$$

Dado que cada iteración se diseñó para representar el universo Archivo, la combinación de ponderaciones a través de las iteraciones requirió la división entre nueve.

## 2.6 Post-estratificación

Tal como se mencionó en el acápite 2.3, se consideró necesario hacer ajustes adicionales a las ponderaciones. Para determinar estos ajustes utilizamos los valores de la población para compararlos con los valores estimados. El valor más confiable de la población conocida era el total de metros lineales de papel, en el Archivo, para cada iteración. Aplicamos el ajuste de post-estratificación de una celda; como factor de ajuste utilizamos la proporción de metros lineales estimados (tal como se describió en el apartado 2.3) con respecto a los metros lineales conocidos a través de todos los ambientes. Este es un factor de ajuste poco convencional –el ajuste tradicional post-estratificación implica el uso, por sí mismos, de los elementos muestreados (por ejemplo, el número de UUA) en vez de los atributos de los elementos muestreados (por ejemplo, metros lineales de papel). No obstante, ante la ausencia de información a nivel de población acerca del número de UUA, consideramos que los metros lineales eran una representación razonable. Realizamos este ajuste solamente para las ponderaciones a nivel de las UUA (tercera etapa), ya que ni las UI ni los documentos se midieron en metros lineales y por lo tanto el ajuste no se podía aplicar a mayor escala.

La Tabla 1 presenta un ejemplo de cómo se aplicó el ajuste. La primera columna de la tabla es una identificación única para cada UUA. Los metros lineales de una UUA son equivalentes a la altura de la UUA, tal como se muestra en la segunda columna de la tabla (ML). La tercera columna representa la ponderación base (Pt) para cada una de las UUA. La cuarta columna sería el cálculo del total de metros lineales para cada UUA y por lo tanto su suma sería la estimación del total de metros lineales de papel que hay en el Archivo (si hubiera sólo diez UUA en el archivo entero). A continuación, tomamos el valor conocido de la población del total de metros lineales en el Archivo (Pob ML) que, en este ejemplo, es equivalente a 1100 metros lineales. Con este valor conocido creamos un factor para corregir las ponderaciones, tal como se aprecia en la quinta columna (Pob ML/Total Estimado de Metros Lineales). Este valor representa las ponderaciones ajustadas (ponderación base \* el

factor). Finalmente, en la última columna calculamos los metros lineales ponderados por UUA, utilizando las ponderaciones ajustadas. La suma de la última columna es el cálculo del total de metros lineales utilizando las ponderaciones ajustadas. Esta estimación debería igualar el total conocido de metros lineales de papel en el Archivo.

**Tabla 1:** Ajuste Post-estratificación

UUA	Metros lineales (MT)	Ponderación n (Pt)	Metros lineales ponderados	Pt Ajustada (Pt*Pob. MT)	Estimación de ML Ajustados
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
Total ML Estimados			1744.06		
Total ML Ajustados					1100

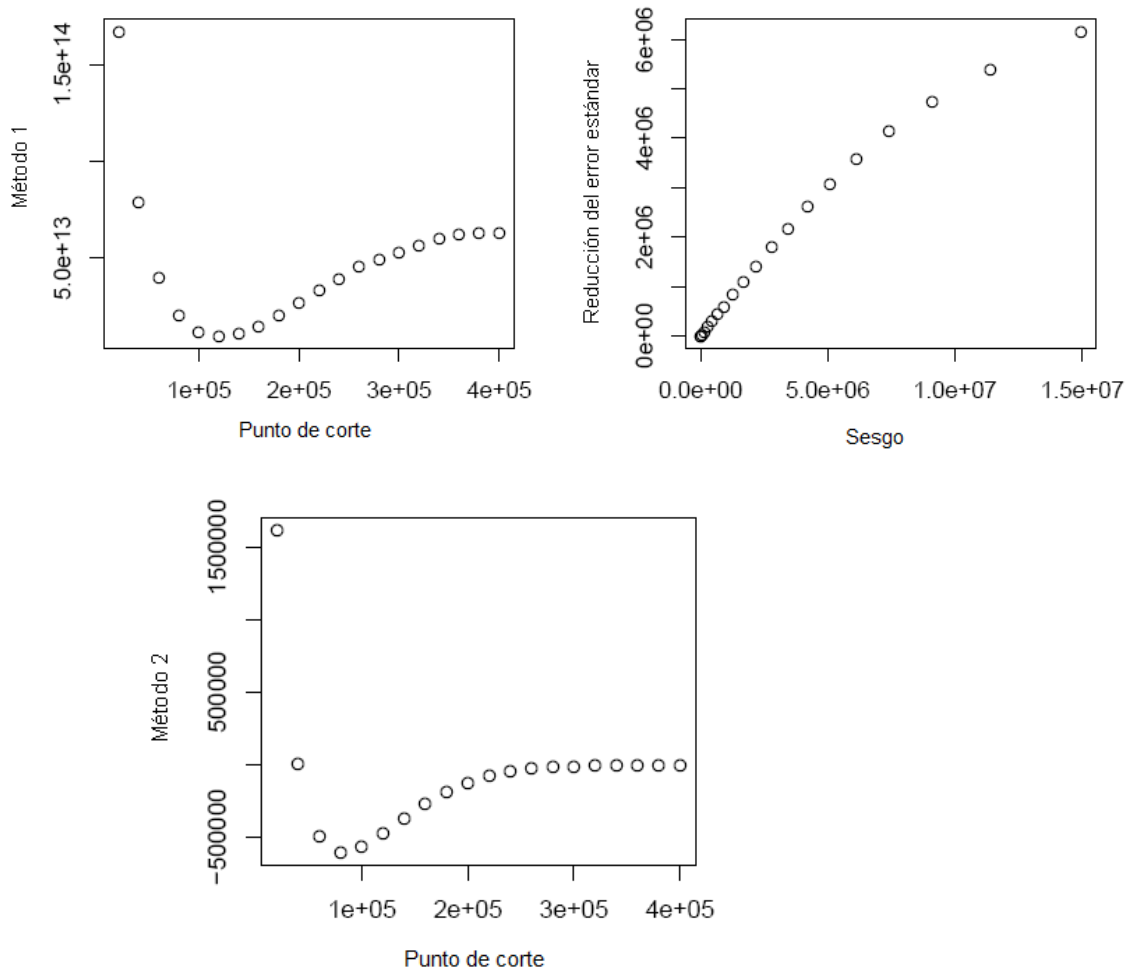
## 2.7 El recorte de grandes ponderaciones

Es bien conocido que en las encuestas muestrales complejas pueden presentarse grandes ponderaciones no planificadas. Dado nuestro diseño muestral, esperábamos algunas ponderaciones muestrales desiguales. Sin embargo, la presencia de grandes variaciones en las ponderaciones muestrales puede reflejar otros problemas, como errores en el marco muestral, en el proceso de codificación o en el procedimiento de ajuste. Tal como advierte Potter, “esta variación extrema o no planificada en las ponderaciones muestrales puede provocar varianzas muestrales exageradas, y unas cuantas ponderaciones extremas pueden contrarrestar los aumentos en la precisión de un diseño muestral que, por otra parte, está bien diseñado y ejecutado” [3].

Desafortunadamente, nuestros cálculos iniciales de la ponderación revelaron una variación extrema - la ponderación máxima calculada fue más de 460 veces mayor que la ponderación media. Continuamos revisando el producto ponderado para entender por qué hubo una variación tan grande. Los cálculos de la ponderación dependían de muchas medidas que el equipo de trabajo de campo tuvo que tomar durante la tercera etapa, mientras trabajaban sobre el terreno. Dadas las condiciones físicas del Archivo, tomar estas

medidas a menudo fue muy difícil. La cuarta etapa también presentó una gran variación, pero en menor grado que la tercera. Tratamos de corregir las malas mediciones utilizando para ello el conocimiento del equipo de trabajo de campo. Una opción hubiera sido recortar las grandes ponderaciones en cada etapa pero, teóricamente, las grandes ponderaciones en una etapa podían compensarse por ponderaciones muy pequeñas en la siguiente etapa. Decidimos recurrir a la exploración visual para escoger el punto de corte para las grandes ponderaciones acumulativas, tomando en consideración la compensación entre el sesgo y la reducción de la varianza (ver Gráfica 1). Se escogió un punto de corte (de conformidad con el método descrito más adelante) que dio como resultado el recorte de 56 registros (0.7%). Posterior al recorte, la razón entre la estimación máxima y la ponderación media era de 115. Antes y después del recorte estadístico, la estimación del número total de documentos se mantuvo constante.

**Gráfica 1:** Exploración Visual de las Ponderaciones para Determinar el Punto del Recorte Estadístico





Empezamos a explorar el punto de corte adecuado trazando directamente el sesgo y contraponiéndolo a la reducción del error estándar, tal como se muestra en la Gráfica 1, arriba a la derecha (cada punto representa el valor de un punto potencial de corte). Aunque podemos ver el comportamiento del sesgo en relación con la reducción del error estándar, resulta difícil determinar en la gráfica el valor del punto de corte. Posteriormente utilizamos dos métodos para tomar en cuenta la compensación entre el sesgo y la reducción del error estándar para puntos de corte específicos. Trazamos los valores de las ecuaciones 8 y 9 (infra) para obtener puntos de corte, tal como se puede observar en las dos representaciones ubicadas a la izquierda de la Gráfica 1. Para ambas ecuaciones, el mejor punto de corte se estableció en el valor mínimo (por ejemplo, en la depresión de cada una de los trazados de arriba). Combinando información de los tres, escogimos el mejor punto de corte para el proceso de recorte estadístico.

$$\text{Método 1} = \widehat{\text{sesgo}}(ML_c)^2 - \widehat{EE}(ML_o)^2 + 2 * \widehat{EE}(ML_c)^2 \quad (8)$$

$$\text{Método 2} = \frac{\widehat{\text{sesgo}}(ML_c)^2}{ML_o} - \frac{(\widehat{EE}(ML_o) - \widehat{EE}(ML_c))^2}{ES(ML_o)} \quad (9)$$

donde:

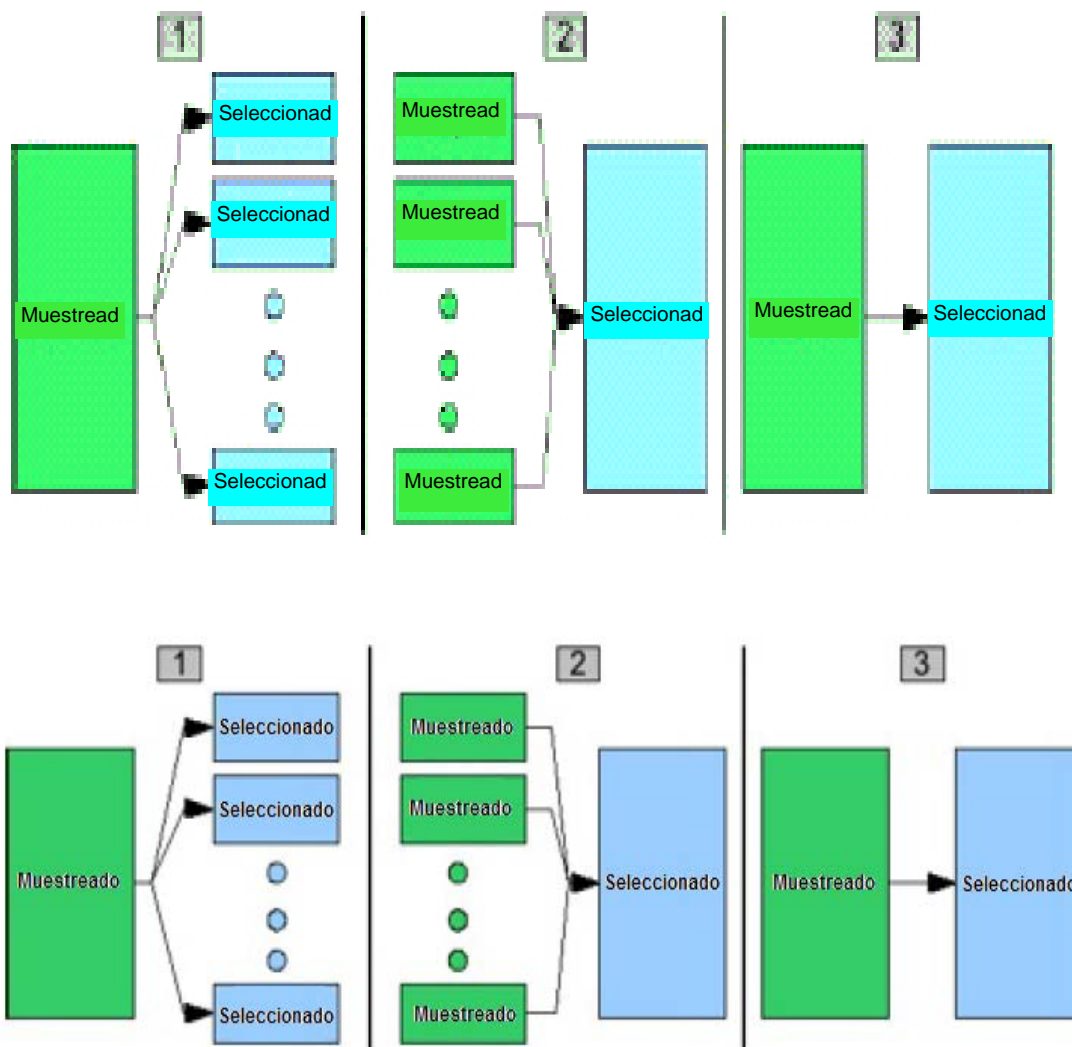
$ML_o$  = estimación del total de metros lineales utilizando las ponderaciones antes del recorte estadístico, y

$ML_c$  = estimación del total de metros lineales utilizando las ponderaciones recortadas en el punto de corte  $c$ .

### 3. Problemas Especiales Adicionales

En los apartados 2.1, 2.3 y 2.4 se expusieron los problemas generados por nuestra incapacidad para estimar con exactitud la probabilidad de selección de los ambientes en las iteraciones uno y dos, así como las UUA y las UI en todas las iteraciones. El continuo traslado del contenido de los contenedores al interior de los almacenes provocó problemas adicionales para la ponderación. A causa de ello, la composición de un ambiente pudo haber cambiado de tal manera que desde el momento en que se tomaron las medidas iniciales de tamaño hasta que se seleccionaron los contenedores, el contenido de un contenedor muestreado podía no encontrarse ya en la misma ubicación. Como muestra la Gráfica 2, el traslado podía realizarse principalmente de tres maneras: un contenedor se dividía en varios contenedores, varios contenedores se combinaban en uno solo, o todos los documentos procedentes de un contenedor muestreado se reubicaban en un nuevo contenedor. Una cuarta posibilidad, poco común, no se ha representado en la Gráfica 2, y consiste en el traslado desde varios contenedores hacia muchos otros contenedores (ver Tabla 2), y sucede cuando el contenido de múltiples contenedores muestreados se reubica en varios contenedores nuevos.

**Grafica 2:** Posible Traslado de Contenedor



No obstante, como muestra la Gráfica 2, la mayoría de los contenedores (65%) no sufrieron ningún traslado. La segunda fila de la tabla presenta aquellas situaciones en las que un contenedor fue trasladado de un lugar a otro, y en la que asumimos que no se retiró ningún documento del contenedor ni se añadieron al contenedor documentos de otra procedencia, tal y como sucedió con el 16% de los contenedores muestreados. En el 9% de los contenedores muestreados, los documentos procedentes de muchos contenedores muestreados fueron reubicados en uno solo que ya había sido seleccionado (pero no muestreado originalmente). El 5% de los contenedores se dividieron a partir de un contenedor hacia varios, y un uno por ciento de ellos fueron trasladados de varios contenedores hacia otra serie de contenedores. Por último, no tenemos información con respecto al movimiento del cuatro por ciento de los contenedores.

**Tabla 2:** Reubicación de Contenedores

Desde	Hacia	Frecuencia	Porcentaje
Sin movimiento		772	65
1	1	189	16
Muchos	1	102	9
1	Muchos	61	5
Muchos	Muchos	9	1
Desconocido		48	4

Los pesos sólo pueden calcularse de la manera habitual cuando los viejos y los nuevos contenedores se corresponden idénticamente, cuando se han hallado los nuevos - es decir, los contenedores que reciben la documentación - y cuando existen estimaciones, en metros lineales, para todos los contenedores que se combinaron. Si un contenedor muestreado se combina con otro contenedor que no ha sido muestreado y esto se desconoce, la ponderación se basará en la probabilidad de selección, que es menor que la verdadera probabilidad de selección; por ejemplo, la ponderación hace caso omiso del hecho que había un segundo contenedor que hubiera podido conducirnos al mismo contenedor muestreado. No obstante, en la mayoría de los casos fue posible hacer un correcto seguimiento de los contenedores. Por lo tanto, no creemos que el traslado haya afectado seriamente las ponderaciones (ver Tabla 2, supra); sólo los 48 contenedores que aparecen en la última fila de la Tabla 2 serían susceptibles de haber sido ponderados equivocadamente.

#### 4. La Futura Etapa de Ponderación

En la actualidad, estamos planificando agregar un paso adicional en el proceso de ponderación. Ya que en cada iteración seleccionamos 33 ambientes independientemente (para las iteraciones 3 a 9) e hicimos selecciones independientes a través de las iteraciones, resulta que algunos ambientes están sobre-muestreados y otros están sub-muestreados. Planificamos rastrear las ponderaciones de los ambientes  $P(E_i)$  para obtener dos grupos de control. Este rastreo todavía no se ha realizado, y los cálculos que se presentan en el tercer documento de esta serie [2] utilizan ponderaciones que aún no han incluido el ajuste del rastreo. Nótese que los metros lineales de papel en los contenedores de un ambiente pueden cambiar a través del tiempo y, por lo tanto, cambiar para cada iteración. Por consiguiente, un grupo de control para rastreo son los metros lineales de papel en los contenedores en cada iteración (a través de todos los ambientes). El segundo grupo de control que ha de rastrearse son los metros lineales de contenedores en cada ambiente a través de las iteraciones. Para hacer el rastreo, probablemente combinemos algunos ambientes pequeños, de manera que hagamos el rastreo por grupos de ambientes en vez de hacerlo por ambientes individuales. Se aplicará un factor de rastreo  $R_j$  inmediatamente después de la ponderación base del ambiente.

## 5. Imputación

Tal como se mencionó en el apartado 2.7, hay muchos motivos por los que pueden provocarse grandes ponderaciones. Es importante investigar estas causas, en tanto podrían ser un indicador de mala información. Durante la etapa de limpieza de datos, encontramos que en nuestra investigación había dos fuentes principales de mala información: datos poco confiables y datos faltantes.

Con el objeto de detectar los datos poco confiables, llevamos a cabo una comprobación de validez: en términos de metros lineales,  $UI_{ijk} < UUA_{ijk} < C_{ij} < A_i$ . Encontramos que algunas mediciones de los contenedores y de las UUA eran demasiado grandes para ser correctas de manera convincente (por ejemplo, UUA que eran mayores que los contenedores de donde habían sido muestreadas). Esto ocurrió en aproximadamente el 8% de los contenedores y en el 17% de las UUA. Algunos de estos errores ocurrieron durante la tabulación de los datos. Otros de los errores pueden haberse debido a confusiones en la unidad de medida, al lugar en donde se colocó el punto decimal o a datos faltantes. Esta clase de errores pueden haber sido exacerbados por la alta rotación de personal de campo durante el curso de las nueve iteraciones. En aquellos casos en que fue posible, volvimos a los formatos de código duro (originales) para recuperar los datos que habían sido digitados erróneamente.

Para los datos erróneos restantes utilizamos el método de imputación estándar “k-vecinos más próximos” (*standard imputation method k-nearest neighbors* en inglés), denominado “imputación” en el paquete de software R[4]. La imputación se realizó para las mediciones erróneas de UUA y contenedores, así como para el número de páginas faltantes de una UI.

## 6. Conclusiones

Hemos descrito la ponderación de la muestra del Archivo Histórico de la Policía Nacional de Guatemala y los principales retos para calcular estos pesos. Dejando de lado lo extraordinariamente poco tradicional del procedimiento de muestreo requerido por la estructura del Archivo, confiamos en que hemos logrado formular mediciones representativas. Las lecciones aprendidas a partir de los retos expuestos en este documento ya se están implementando en el muestreo que está en proceso actualmente en el Archivo. Más importante aún es que las muestras de la iteración 10 ya no son susceptibles al traslado de documentos dentro del Archivo. En esta iteración se han muestreado aproximadamente unos 12,000 documentos más y esperamos simplificar el procedimiento de ponderación resultante y, por lo tanto, conseguir que nos consuma menos tiempo. Ya no tendremos problemas por probabilidades de selección desconocidas y esperamos que las ponderaciones sean mucho menos variables. Estamos convencidos de que estos documentos enriquecerán nuestra comprensión del papel desempeñado por la Policía Nacional en la violencia del conflicto armado interno de Guatemala durante los años 1960 a 1996.

### Referencias

- [1] Guzmán, D., Guberek, T., Shapiro, G., Zador, P. (2009), “Estudiando Millones de Documentos Rescatados: Plan de Muestreo para el Archivo Histórico de la Policía Nacional de Guatemala”, en *JSM Proceedings*, Sección de Métodos de Investigación Pericial (Survey Research Methods). Alexandria, VA: Asociación Americana de Estadística (*American Statistical Association*, por su nombre en inglés).
- [2] Price, M., Guberek, T., Guzmán, D., Zador, P., Shapiro, G. (2009), “Análisis Estadístico del Archivo Histórico de la Policía Nacional de Guatemala: en Busca de Documentación sobre Violaciones de Derechos Humanos”, en *SM Proceedings*, Sección de Métodos de Investigación Pericial (Survey Research Methods). Alexandria, VA: Asociación Americana de Estadística.
- [3] Potter, F. (1988), “Procedimientos Periciales para Controlar Ponderaciones Muestrales Extremas”, en *Proceedings*, Sección de Métodos de Investigación Pericial (Survey Research Methods). Alexandria, VA: Asociación Americana de Estadística. 453-458.
- [4] Hastie, T., Tibshirani, R., Narasimhan, B, Chureference, G. (2009), “Imputación de datos sobre *microarrays* – un paquete R”, <http://cran.r-project.org/web/packages/impute.index>.

Traducción del inglés al español: Carla Villagrán, Traductora jurada  
Traducción revisada por Beatriz Bejarano, Benetech.

---

---

**Análisis Estadístico del Archivo Histórico  
de la Policía Nacional de Guatemala: en Busca de Documentación  
sobre Violaciones de Derechos Humanos**

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

**Resumen**

El presente trabajo ofrece un primer acercamiento estadístico al Archivo Histórico de la Policía Nacional de Guatemala (AHPN). Se presentan estimaciones preliminares del nivel de conocimiento que ciertos grupos dentro de la Policía Nacional tenían sobre actos violentos registrados en documentos institucionales. Con base en una muestra probabilística multietápica de documentos del AHPN, se han calculado estimaciones ponderadas.

Como se ha mencionado en los documentos anteriores, la estructura física del Archivo presentaba muchos retos, lo que resultó en un complicado cálculo de probabilidades. El análisis estadístico adoptó los contenedores de los documentos (etapa 2) como unidad primaria de muestreo (UPM). Se comparó la proporción del total de los documentos del Archivo de los cuales cualquier grupo específico tenía conocimiento con la proporción de documentos *de interés* de los cuales un grupo tenía conocimiento. Se estima que aproximadamente un 20% de los documentos del Archivo contiene información sobre los hechos de interés.

**Palabras clave:** muestra multietápica, diseño muestral complejo, violaciones de derechos humanos, Archivo Histórico de la Policía Nacional.

**I. Introducción**

Cuando en 2005 se descubrió el Archivo Histórico de la Policía Nacional de Guatemala (AHPN), familiares de víctimas, organizaciones de derechos humanos e historiadores vieron renacer sus esperanzas ante la posibilidad de que estos millones de documentos pudieran contener información valiosa sobre los crímenes cometidos durante el conflicto armado interno. Inicialmente hubo escepticismo sobre si en el Archivo existían aún documentos de interés para una investigación de violaciones de derechos humanos o si, por el contrario, tales documentos habían sido extraídos de este acervo documental. La interrogante sigue siendo si el Archivo fue depurado o no. Sin embargo, el presente trabajo ofrece un primer acercamiento estadístico al Archivo y confirma que, aun si posteriores análisis revelaran la existencia de vacíos estratégicos en su fondo documental, éste contiene abundante información de utilidad para aprender sobre el violento pasado del país.

Este estudio incluye documentos escritos o archivados entre 1960 y 1996, período del conflicto armado interno en Guatemala. Se estima que de un total de 31,227,378

---

\*Megan Price, PhD., es estadística del Grupo de Análisis de Datos sobre Derechos Humanos de la organización Benetech (*Human Rights Data Analysis Group, HRDAG*, por su nombre y sigla en inglés). Tamy Guberek y Daniel Guzmán son consultores de HRDAG en Benetech [www.hrdag.org](http://www.hrdag.org), [www.benetech.org](http://www.benetech.org). Paul Zador, PhD., es Estadístico Principal de Westat. Gary Shapiro, recientemente jubilado, fue estadístico principal de Westat durante la mayor parte de su participación en este estudio. Ambos, Shapiro y Zador, son miembros del Grupo Voluntario de Intereses Especiales de la Asociación Americana de Estadística.

documentos correspondientes al período de interés, un 20% contiene información acerca de hechos de interés tales como muertes, desapariciones y detenciones. Este informe describe los hallazgos iniciales relacionados con el contenido del Archivo y presenta una comparación del nivel de conocimiento de documentos de interés con el de conocimiento del total de los documentos correspondientes a subconjuntos específicos de la Policía Nacional. En el apartado 2.1 se definen los términos “conocimiento” de un documento y documentos “de interés”.

Uno de los objetivos de este trabajo consiste en plantear preguntas dirigidas a los investigadores e historiadores del Archivo, que son los expertos locales en la materia. Esperamos que estos resultados les ayuden en la investigación en curso, la cual se sumará al trabajo previamente realizado por otras instituciones con el objetivo de develar el pasado violento de Guatemala; como es el caso de la Comisión de Esclarecimiento Histórico, el Proyecto REHMI de la Iglesia Católica y numerosos informes de ONG, todos los cuales carecen de información proveniente de la Policía Nacional [1].

El descubrimiento del Archivo Histórico de la Policía Nacional de Guatemala, la selección de la muestra del Archivo y la metodología utilizada en el proyecto muestral, así como los cálculos de las ponderaciones que resultaron del método de muestreo fueron abordados en nuestros dos informes previos [2,3]. Cabe recordar que la estructura general de la muestra fue multietápica y que se construyó sobre niveles topográficos dentro del Archivo: ambientes, contenedores, últimas unidades de agregación (UUA) y unidades de información (UI). Todos los documentos contenidos en una UI seleccionada pasaron a formar parte de la muestra.

En el apartado 2 de este documento se describen las variables que se codificaron a partir de los documentos muestreados y su pertinencia para los objetivos de la investigación. El numeral 3 presenta los resultados estadísticos. El 4 ofrece algunos elementos para la discusión acerca de la pertinencia de los resultados y algunos de los retos estadísticos que enfrentamos al realizar los cálculos. El capítulo 5 recoge las conclusiones del estudio. Finalmente, en el apartado 6 se prefiguran algunos análisis para futuros estudios acerca del AHPN.

## **Métodos**

- **Codificación de los documentos**

La muestra se diseñó con el propósito de respetar, en la medida de lo posible, las condiciones y estructuras de archivo que se encontraron en el AHPN. Debido a que la Policía Nacional frecuentemente relacionaba y archivaba juntos muchos documentos individuales sobre un mismo tema, como si fueran una sola unidad, se decidió muestrear unidades de información (UI) en vez de documentos individuales. Luego, la información se codificó a partir de cada uno de los documentos que conformaban una UI muestreada. Los datos utilizados en este análisis estadístico provienen de documentos desagregados a partir de unidades de información (por ejemplo: cada registro en el análisis del conjunto de datos corresponde a un documento).

El esquema de codificación se diseñó para recoger información de los documentos con el fin de responder a dos preguntas específicas de investigación: ¿qué podemos aprender acerca de las estructuras de mando y control de la Policía Nacional a partir de estos documentos? ¿Puede contribuir el análisis de patrones y magnitudes de las posibles violaciones de derechos humanos mencionadas en los documentos del Archivo a la comprensión del pasado violento de Guatemala, en particular al de las desapariciones? Analizaremos los resultados iniciales relativos a estas preguntas en el capítulo 4.

Las variables codificadas incluyen información general acerca de cada documento, como la condición física, el tipo de documento y la fecha. Otras variables fundamentales para el análisis contienen información acerca de lo que llamamos “actores a nivel de documento”, tales como los autores, los destinatarios directos e indirectos y las características de estos actores, como por ejemplo la unidad policial u otra filiación a la que pertenecían, así como el rango de cualquier individuo relacionado con la creación o transmisión del documento. El último nivel de información que registramos es el del contenido de los documentos, específicamente si los documentos incluyen datos de hechos de interés para la investigación. Los documentos se consideran “de interés” si contienen información acerca de cualquiera de los hechos incluidos en una lista de 11 que están potencialmente relacionados con violaciones de los derechos humanos (ver Tabla 2).

Dado que una pregunta de interés primordial en este estudio es quién en la Policía Nacional tenía conocimiento de diversos hechos, a partir de las variables a nivel de documento descritas arriba se crearon variables indicadoras adicionales. Por ejemplo, para definir posiciones de mando dentro de la Policía Nacional se agruparon los actores a nivel de documento de acuerdo con los rangos al interior de la jerarquía policial. Ocupar una posición de mando en la estructura policial se consideró equivalente a tener conocimiento de un documento si la persona que ocupaba un puesto de mando era la creadora, receptora (directa o indirecta) o mecanógrafa de cualquier elemento del documento.

La Policía Nacional estaba dividida también en unidades y subunidades (estructuras). Tal como nos sugirieron nuestros colegas de Guatemala, reunimos las cerca de 50 estructuras en 13 nuevas categorías, según el área de trabajo de cada unidad. Se consideró que cada una de estas categorías tuvo conocimiento de un documento si cualquier persona perteneciente a esa unidad fue su autora o lo recibió (directa o indirectamente).

Dado que la información contenida en los documentos del Archivo es de naturaleza fundamentalmente cualitativa, creamos un vocabulario para todas las variables que debían codificarse. El vocabulario controlado fue decisivo para ayudar a mantener altos niveles de confiabilidad entre los codificadores, cuyo trabajo era transformar la información cualitativa en unidades cuantitativas.[4] Se calculó la confiabilidad entre codificadores y esta resultó, de manera constante, superior al 75% a lo largo de las nueve iteraciones de la muestra.



## 2.2 Cálculo de las estimaciones

Utilizando el software estadístico R, se hizo uso de un paquete de nombre ‘Survey’ para calcular todas las estimaciones descritas en los siguientes epígrafes.[5] El paquete incluye funciones para crear ‘objetos’ para el diseño muestral basados en el agrupamiento y la estratificación de la información, el cálculo de ponderaciones y un método de muestreo. Hacia el final de la novena iteración habíamos realizado lo que era, básicamente, un censo de los ambientes que componen el Archivo; por lo tanto, para fines de la estimación de la varianza, la segunda etapa, los contenedores, fueron considerados la unidad muestral principal para este análisis. El paquete ‘Survey’ provee también una función para calcular los pesos replicados; en nuestro caso se utilizó el método JK1[6]. Los pesos replicados se utilizaron entonces para estimar la varianza de cada estimación puntual reportada en este informe.

- **Resultados**

Este apartado presenta las estimaciones correspondientes al nivel de conocimiento que la Policía Nacional tenía sobre los hechos de interés, utilizando como base el muestreo multietápico. En tanto este informe por sí solo no es suficiente para emitir conclusiones definitivas con respecto al papel que desempeñó la Policía Nacional en el pasado violento de Guatemala, creemos que estas estimaciones plantearán importantes preguntas para los expertos en la materia, ayudarán a orientar investigaciones futuras acerca de la estructura de mando de la Policía Nacional y contribuirán a asegurar el interés y los recursos para futuros estudios.

### 3.1 Estimaciones globales

La Tabla 1 presenta las estimaciones globales relativas a los documentos del Archivo basadas en la muestra multietápica. Debido a que se muestrearon unidades de información en vez de documentos directamente, los investigadores se encontraron con que las unidades de información podían contener documentos pertenecientes a años previos o posteriores al período 1960-1996 y, las UI podían incluir, como mínimo, un documento comprendido en ese lapso. Una UI también podía contener un documento sin fecha de creación junto a otro documento con una fecha válida. Las fichas de identificación son el típico documento que no tiene una única fecha de creación sino una fecha para cada entrada que se agregaba a la tarjeta. Ya que los dos tipos de casos mencionados anteriormente (fechas fuera del rango o documentos que no contienen una fecha válida) pueden estar incluidos en una UI elegible, en la Tabla 1 se presentan dos totales: uno de los documentos *relacionados* con el período de la investigación, y otro de los documentos *correspondientes* al período de interés.

La última fila de la Tabla 1 presenta una estimación del número total de hechos mencionados en los documentos del Archivo. Nótese que esta estimación no toma en consideración la posibilidad de que un mismo hecho haya podido mencionarse en varios documentos. En el futuro se planea deduplicar los hechos contenidos en la muestra y ofrecer una estimación de hechos únicos registrados en los documentos del Archivo.

**Tabla 1:** Estimación de los Documentos en el Archivo

	Estimación (EE)	95% IC
<b>Total relacionado con 1960-1996</b>	34,971,341 (5,263,488)	(24,654,904, 45,287,778)
<b>Total correspondiente al período 1960-1996</b>	31,227,378 (5,109,612)	(21,212,538, 41,242,219)
<b>Total de documentos de interés</b>	6,285,487 (1,058,066)	(4,211,678, 8,359,297)
<b>Proporción de documentos de interés</b>	0.20 (0.02)	(0.16, 0.24)
<b>Número total de hechos</b>	12,079,401 (2,275,698)	(7,619,033, 16,539,769)

### 3.2 Hechos de interés

La Tabla 2 es una lista desagregada de los 11 hechos de interés codificados a partir de los documentos. Presenta, en orden decreciente, la estimación del número total de documentos que contienen información acerca de cada tipo de hecho, así como de la proporción que representan con respecto a los documentos de interés.

La categoría, documentos *de interés*, se refiere a documentos que contienen información acerca de hechos incluidos en la presente lista. Los documentos pueden contener información sobre más de un hecho; por lo tanto, la sumatoria de los totales de la Tabla 2 es mayor que la estimación del número total de documentos de interés referida en la Tabla 1. También sucede que cualquier hecho específico puede haber sido reportado en múltiples documentos y la frecuencia con que los hechos se mencionan puede depender de muchos factores, como el tipo de acto y la fecha del documento.

**Tabla 2:** Desagregación de hechos de interés

	Total (EE)	Proporción (EE)
<b>Detenciones</b>	4,658,367 (904,385)	0.741 (0.04)
<b>Muertes</b>	1,365,218 (345,939)	0.217 (0.04)
<b>Denuncias</b>	572,299 (165,704)	0.091 (0.02)
<b>Abuso físico y psicológico</b>	439,666 (189,983)	0.070 (0.03)
<b>Intimidación</b>	302,113 (99,268)	0.048 (0.02)
<b>Desaparición</b>	223,749 (62,035)	0.036 (0.01)
<b>Exhibición personal</b>	65,519 (27,956)	0.010 (0.004)
<b>Interrogatorio</b>	30,171 (22,163)	0.005 (0.004)
<b>Ingreso ilegal a propiedad privada</b>	29,484 (12,408)	0.005 (0.002)
<b>Abuso sexual</b>	14,102 (4,984)	0.002 (0.001)
<b>Vigilancia</b>	1,307 (1,199)	<0.0001 (<0.0001)

Cabe destacar que incluso en el caso de hechos relativamente difíciles de identificar, como las desapariciones, que significan sólo el 3.6% de los documentos con hechos de interés, el número total aproximado de documentos (en este caso 223,749) puede resultar relevante.

El conocimiento de los documentos a lo largo del tiempo por parte de diversas estructuras policiales se describirá en los apartados siguientes; pero vale la pena mencionar aquí el conocimiento que una estructura específica tuvo sobre los hechos desagregados. Los cuerpos departamentales de policía, cuya función era patrullar las áreas fuera de la ciudad de Guatemala, registran una cantidad altamente desproporcionada de información acerca de hechos específicos de interés: estos cuerpos estaban informados acerca de aproximadamente el 75% (EE = 17.1) de los abusos físicos y psicológicos documentados en los registros policiales, del 54% (EE = 13.2) de las muertes y del 45% (EE = 34.6) de los ingresos ilegales a propiedad privada. Estas tasas de conocimiento acerca de hechos específicos de interés son más altas que las de cualquier otra estructura, con excepción de las unidades encargadas de la investigación, las cuales tenían conocimiento del 90% de los interrogatorios (podría intuirse que las unidades de investigación pudieran tener conocimiento de prácticamente todos los interrogatorios; no obstante, la interpretación de este dato estadístico queda todavía planteada como una pregunta abierta para nuestros colegas del Archivo). Esto suscita preguntas importantes para los especialistas en la materia que tienen a su cargo la investigación en Guatemala: ¿estas altas tasas de conocimiento por parte de los cuerpos departamentales de policía se debían a que tenían una vasta cobertura del país? ¿Se podría explicar acaso porque el interior del país fuera más violento que la ciudad? ¿Es posible que la sede central del Archivo en la ciudad haya sido depurada de documentos de interés antes de su descubrimiento en 2005?

### 3.3 Los documentos a través del tiempo

A continuación comparamos la proporción de documentos *de interés* acerca de los cuales un grupo tuvo conocimiento con la proporción de *todos* los documentos acerca de los cuales un grupo tuvo conocimiento a través del tiempo. Esta comparación, partiendo de una línea base, es importante porque podría parecer que un grupo tuvo un mayor conocimiento de documentos de interés durante un período específico de tiempo; pero esto puede, o no, deberse a un incremento en el número de *todos* los documentos acerca de los cuales el grupo tuvo conocimiento. En otras palabras, sospechamos que la diferencia en los patrones de conocimiento de diferentes clases de documentos ofrece un mayor potencial para la identificación de áreas en las cuales podría encontrarse información adicional relacionada con violaciones sistemáticas de los derechos humanos.

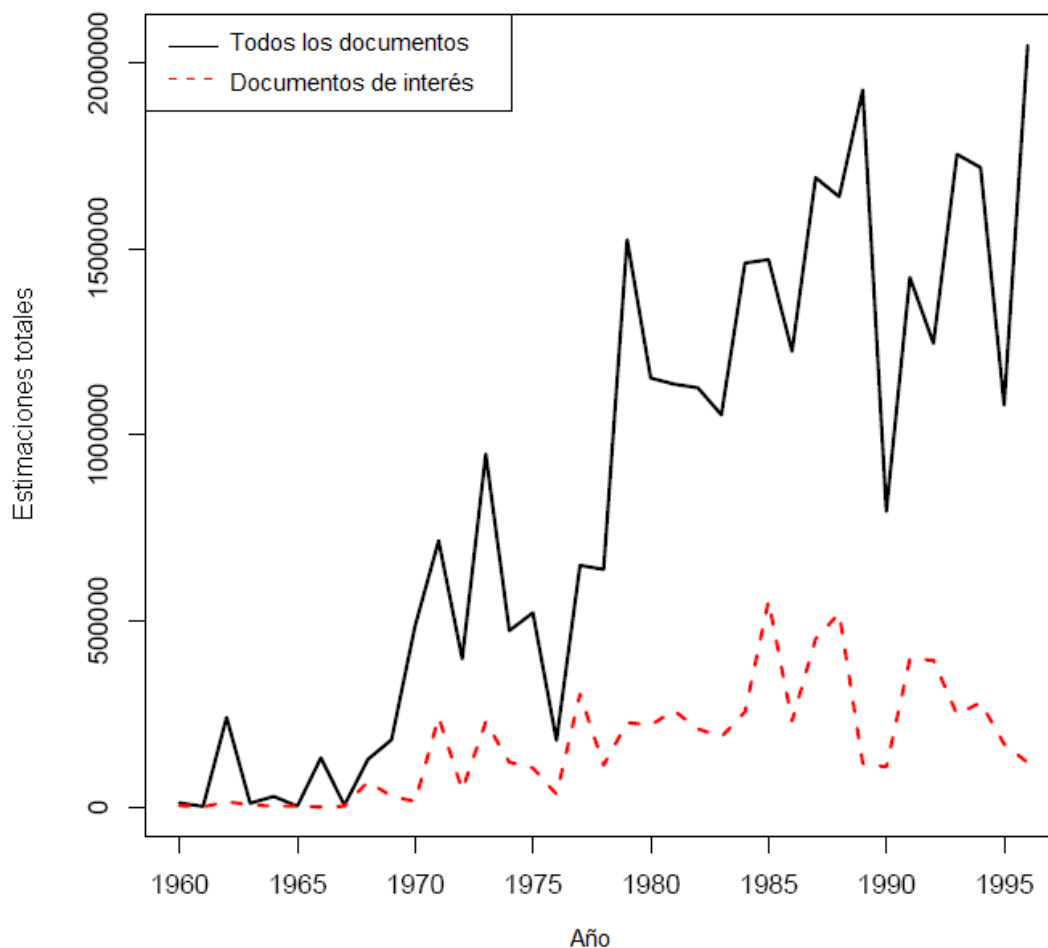
Cabe subrayar que las fechas utilizadas en estos análisis son aquellas en que se creó el documento y no las fechas en que ocurrieron los hechos. Análisis posteriores compararán los patrones de las fechas de los documentos con las fechas de los hechos. También es importante resaltar que aproximadamente el 11% de los documentos de la muestra (902 de 8,162) no tenían fecha (por las razones descritas en el inciso 3.1).

Finalmente, también es importante mencionar que aunque al principio del período de interés (los primeros años de la década de los 60) las estimaciones parecen estar muy

cercanas a cero, esto es simplemente el producto de mayores estimaciones en los años posteriores; las estimaciones correspondientes a los primeros años de los 60 son de magnitud una o dos veces menor que las estimaciones posteriores. Estas estimaciones tempranas también se basan en una muestra mucho más pequeña que las estimaciones posteriores y debería tenerse presente que estas mediciones tienen estimaciones de varianza mucho más grandes (excluidas de las gráficas para facilitar la detección de patrones).

La Gráfica 1 muestra que a través del tiempo la Policía Nacional produjo una cantidad cada vez mayor de documentos (línea continua); sin embargo, hubo un incremento mucho menor en el número de documentos de interés (línea discontinua) creados a lo largo del mismo período de tiempo.

**Gráfica 1:** Todos los Documentos, 1960-1996, versus los Documentos de Interés, 1960-1996



En el inciso 3.4 se puede observar este mismo patrón entre algunas estructuras (Dirección General, particularmente).

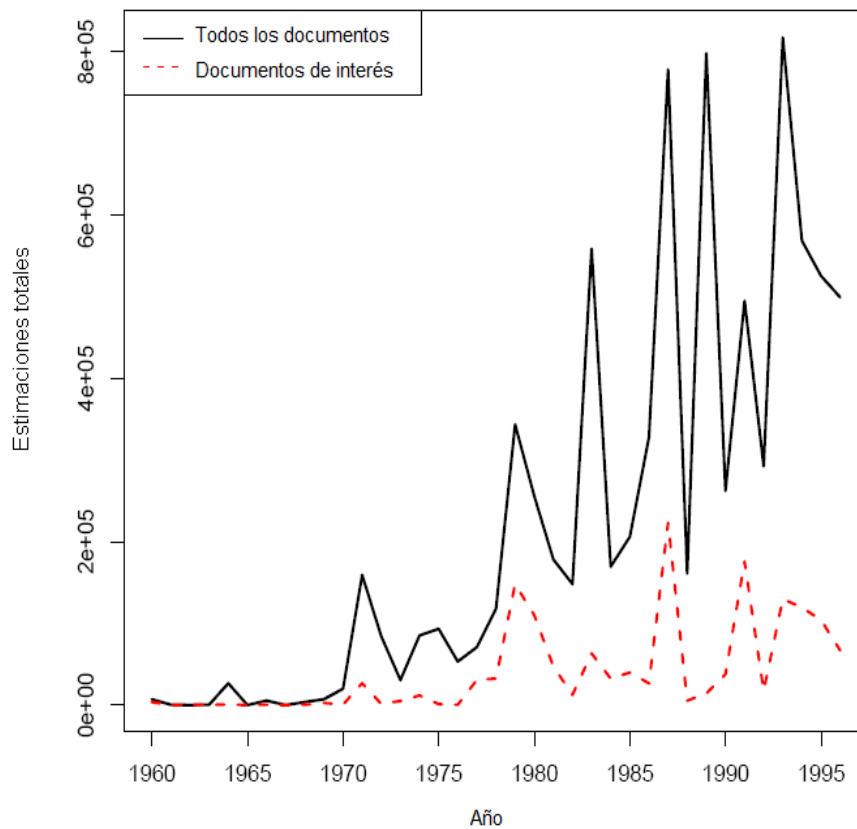
También es importante destacar que el patrón de documentos *de interés* de la Gráfica 1 difiere del patrón de violencia generalmente aceptado para la historia de Guatemala. Las razones potenciales para ello, así como las preguntas que esta diferencia plantea, se abordarán en el apartado 4.

### 3.4 El conocimiento que las estructuras de la policía tenían respecto de los documentos de interés

Las estructuras incluidas en este informe fueron escogidas por el conocimiento cualitativo de su papel en la represión y por los marcados patrones observados en los datos. Algunas veces, como en el caso de la Dirección General de la Policía (ver Gráfica 2), este patrón coincide con el patrón general contemplado en la Gráfica 1. Otros son únicos, como puede observarse en las Gráficas 3 y 4.

Similar al comportamiento señalado en la Gráfica 1, la Gráfica 2 indica que, en general, a lo largo del tiempo, la Dirección General de la Policía (DIGE) tuvo conocimiento de un número creciente de todos los documentos (línea continua), pero de un número relativamente estable (con contadas excepciones) de documentos de interés (línea discontinua).

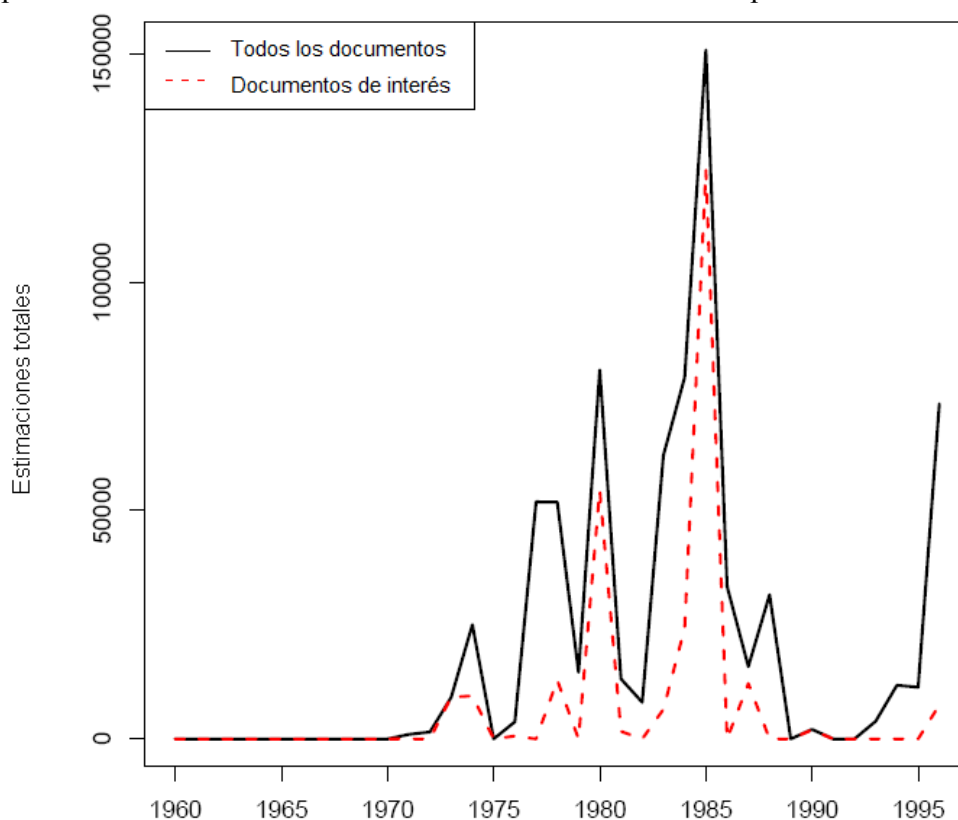
**Gráfica 2:** Todos los Documentos Conocidos por la DIGE, versus los Documentos de Interés Conocidos por la DIGE



La Unidad de Radiopatrullas era la dependencia designada para patrullar las calles de la ciudad de Guatemala y se sospechaba que participaba en actividades de vigilancia. A pesar de tener conocimiento de una proporción relativamente pequeña de todo el Archivo, la Gráfica 3 muestra que, tanto en 1980 como en 1985, esta estructura experimentó importantes incrementos en el conocimiento de todos los documentos (línea continua) y, específicamente, de documentos de interés (línea discontinua).

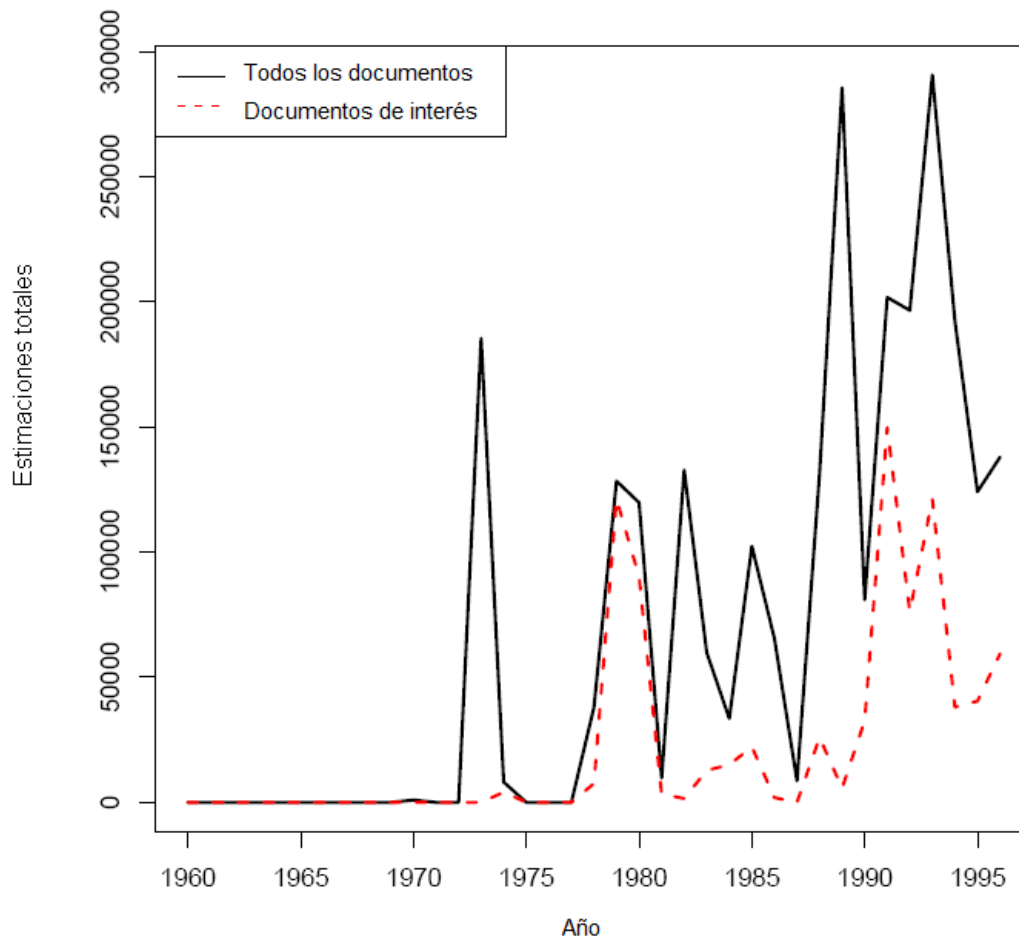
La Gráfica 4 ilustra dos detalles interesantes acerca del tipo y la cantidad de documentos sobre los que tuvo conocimiento el Centro de Operaciones Conjuntas (COC), la estructura que trabajaba directamente con el ejército de Guatemala. Primero, en 1973 demuestran un incremento drástico en el conocimiento que tuvieron de todos los documentos (línea continua), pero sin un incremento correspondiente en el conocimiento de documentos de interés (línea discontinua). Luego, en 1979, sufrieron otro fuerte incremento en cuanto al conocimiento, en esta ocasión también en relación con los documentos de interés. Los años subsecuentes (1982, 1985, 1989) muestran similitudes con 1979,\* es decir, súbitos incrementos en el conocimiento de todos los documentos sin los correspondientes incrementos en cuanto a los documentos de interés.

**Gráfica 3:** Conocimiento de Todos los Documentos por parte de la Unidad de Radiopatrullas versus el Conocimiento de Documentos de Interés por la Misma Estructura



\***Nota de la Traductora:** Dada la redacción del texto, pareciera que en esta ocasión los autores quisieron hacer referencia a 1973, año que se estaría contrastando lo sucedido en 1979, pero el documento original consigna esta última fecha y por lo tanto así se tradujo.

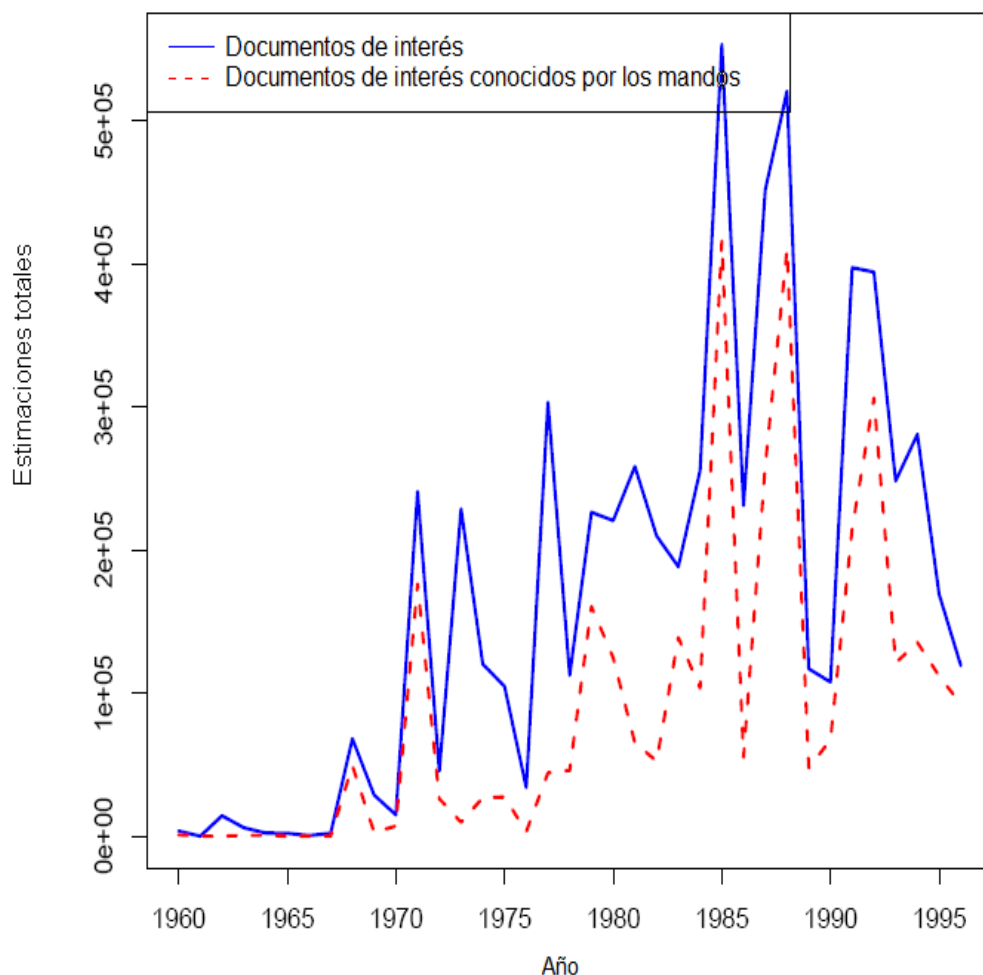
**Gráfica 4:** Conocimiento de Todos los Documentos por parte del COC versus los Documentos de Interés Conocidos por el COC



### 3.5 El conocimiento de los altos mandos sobre los documentos de interés

Una de las motivaciones principales de esta primera serie de estimaciones ha sido entender qué información específica conocían los altos mandos de la Policía Nacional. Nótese que la línea continua en la Gráfica 5 representa una serie de datos diferente de las de gráficas anteriores; la Gráfica 5 compara el número de documentos de interés con los documentos de interés que conocían los altos mandos de la Policía. Esta gráfica indica que, en casi todos los años, los altos mandos tuvieron conocimiento de la mayor parte de los documentos que incluyen información sobre hechos de interés. Las autoridades policiales no pueden alegar ignorancia respecto de las operaciones relacionadas con hechos de interés que llevaba a cabo la Policía Nacional.

**Gráfica 5:** Todos los Documentos Conocidos por los Altos Mandos versus Documentos de Interés Conocidos por los Altos Mandos

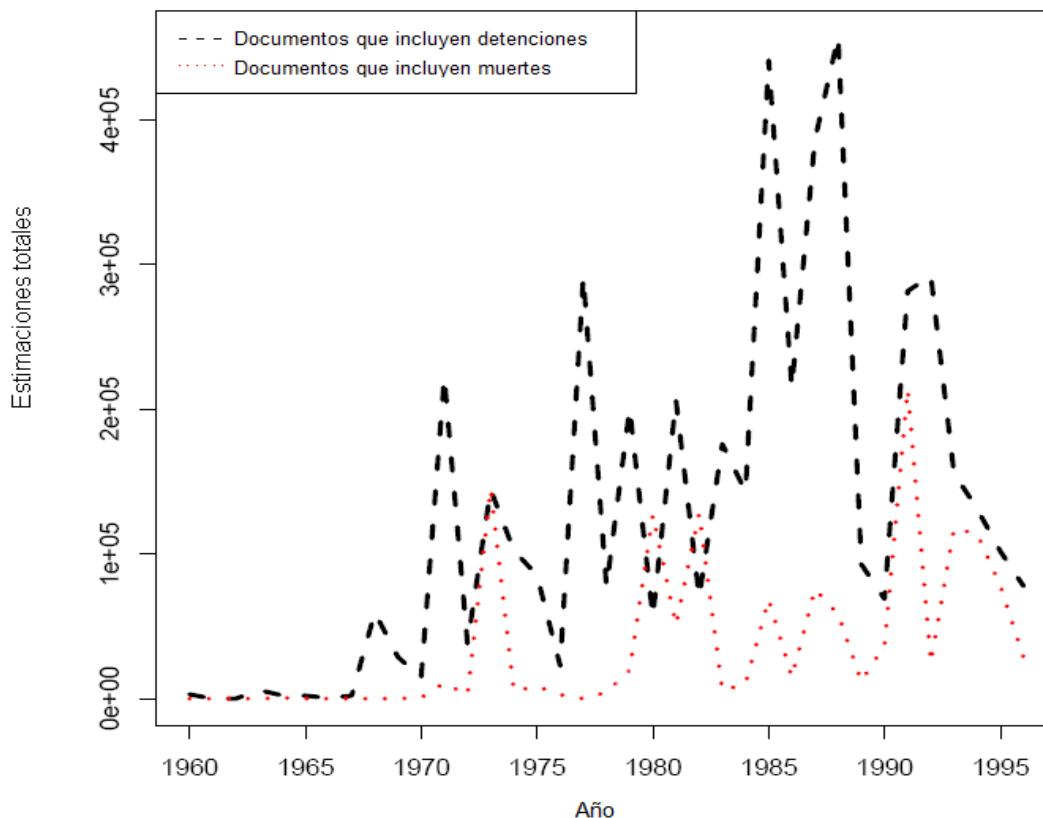


### 3.6 Los hechos desagregados a través del tiempo

Finalmente, la Gráfica 6 recoge un patrón particularmente interesante registrado en 1980 y 1982. En ambos años hay más documentos que incluyen información acerca de muertes que de detenciones. Si se toma en consideración que una de las actividades principales de una fuerza policial es detener individuos, resulta bastante impactante que en estos dos años se registre un número mayor de documentos relacionados con muertes. Ciertamente, se requerirá investigar estos dos años más a fondo.



**Gráfica 6:** Todos los Documentos con Información sobre Muertes versus Documentos con Información sobre Detenciones



#### 4. Discusión de los Resultados

##### 4.1 La magnitud de las posibles violaciones a los derechos humanos

En tanto que una primera estimación acerca del contenido del Archivo, creemos que nuestra investigación ha sido exitosa al señalar que este acervo documental provee una rica fuente de información relacionada tanto con los patrones como con la magnitud de posibles violaciones de los derechos humanos.

Tal como se demuestra en la Tabla 2, probablemente haya cientos de miles, y en algunos casos, millones de documentos con información acerca de muertes, detenciones, desapariciones, intimidaciones y abuso físico y psicológico. Por ejemplo, la estimación de 223,749 documentos con información sobre desapariciones tiene el potencial de ofrecer una riqueza de información que podría permitir conocer el paradero de muchas personas que fueron desaparecidas durante este capítulo de la historia de Guatemala. Sin embargo, aunque algunos otros hechos de interés (enumerados en la Tabla 2) caen dentro del ámbito del conocimiento habitual de la policía, por sí sola la magnitud de información que contiene el Archivo ya amerita continuar investigando.

## 4.2 ¿Quién sabía y hasta qué punto?

Debido a la gran magnitud de documentos con hechos que contiene el Archivo, el patrón de los hechos que la Policía Nacional documentó a través del tiempo también aporta importantes insumos sobre la historia de Guatemala durante este período. Es particularmente importante contrastar los patrones observados en la Gráfica 1 con los patrones de violencia que históricamente se le atribuyen a este período de tiempo en Guatemala. A pesar de que no todos los hechos de interés que figuran en la Tabla 2 son hechos de violencia, ciertamente incluyen varios hechos violentos como muertes, desapariciones y abusos físicos. Diversas fuentes históricas dan cuenta de una campaña contrainsurgente en el área rural entre 1981 y 1983. La Comisión de Esclarecimiento Histórico (CEH) encontró que el ejército de Guatemala cometió actos de genocidio contra la población maya durante estas campañas.[1,7,8]. Resulta preocupante que la Gráfica 1 no muestre un incremento en el número de documentos de interés durante este período, como podría esperarse si la Policía Nacional hubiese estado documentando el número de muertes como consecuencia de actos de genocidio. Las mismas fuentes dan cuenta de un incremento en la violencia urbana en 1979 y 1980, período durante el cual mucha gente fue desaparecida por la fuerza. Una vez más, este patrón no se refleja en la Gráfica 1. Este segundo comportamiento es especialmente relevante debido a que los registros de hechos de violencia de la Policía Nacional se enfocaban principalmente en las áreas urbanas.

Aunque no esperábamos que la documentación de la policía concordara exactamente con los patrones de violencia, sí hubiéramos esperado que durante períodos ampliamente conocidos por su mayor violencia, la documentación de la policía acerca de hechos de violencia hubiera incrementado. Por lo tanto, resulta sorprendente que el patrón de comportamiento de los documentos con hechos de interés esté desvinculado de los patrones de la violencia conocida en Guatemala.[1] No pasamos por alto la posibilidad de que se le ordenara a las diversas dependencias de la policía no registrar cierto tipo de violencia sistemática; o que esta documentación, si existió, fuera sistemáticamente depurada. Tal como se mencionó en el apartado 1, esta posibilidad, así como la manera como se podría comprobar esta idea a modo de hipótesis, siguen siendo preguntas sin resolver.

## 4.3 El reto estadístico

Tal como se describió en anteriores documentos de esta serie [3], los complicados cálculos de ponderación y la fuente de altos niveles de variabilidad resultante representan retos al comparar las estimaciones que contiene este documento. Sin embargo, los objetivos primordiales de este trabajo consisten en hacer un resumen de los contenidos del Archivo, presentar estimaciones acerca de los patrones de conocimiento y ayudar a identificar más áreas de interés para futuras investigaciones. Somos optimistas en cuanto a que las lecciones aprendidas a través de este estudio serán de utilidad para mejorar los cálculos de las ponderaciones en futuras iteraciones y por ende reducir una de las fuentes principales de variabilidad, con lo cual mejorará nuestra capacidad para comparar estimaciones clave.

Otra fuente específica de varianza que esperamos remediar en futuros análisis es el tratamiento erróneo dado a las unidades auto-representadas. Tal como se mencionó en Shapiro [3], las ponderaciones altamente variables fueron provocadas, en parte, por la decisión de llevar a cabo el muestreo utilizando probabilidades proporcionales a las dimensiones de los contenedores, lo cual fue consecuencia de la variedad de sus tamaños.

En los pocos casos en los que los contenedores se auto-representaron (por ejemplo, un solo contenedor en un ambiente) la unidad primaria de muestreo adecuada hubiera sido la última unidad de agregación (UUA). De forma análoga, en casos en los que había sólo un número reducido de contenedores, o bien se debieron haber utilizado las últimas unidades de agregación como las unidades de muestreo primarias, o bien se debió haber considerado el uso de un factor de corrección para poblaciones finitas. Sin embargo, aunque no haber llevado a cabo estas medidas analíticas puede haber resultado en una leve sobreestimación de la varianza, no creemos que hayan sido frecuentes los casos en los que sólo hubiera uno o unos pocos contenedores.

## 5. Conclusiones

A pesar de los numerosos retos físicos que la naturaleza del Archivo presentó para el diseño de la muestra, consideramos haber aplicado una rigurosa técnica de muestreo probabilístico a una excepcional fuente de información para la investigación en derechos humanos. Tenemos la esperanza de que este primer paso, que resume los contenidos del AHPN, no sólo llame la atención hacia una rica fuente de información, sino que además ayude a otros investigadores a realizar indagaciones más específicas en el futuro.

## 6. El trabajo futuro

Los resultados presentados aquí son una pequeña fracción de la investigación que es posible realizar con tan abundante información. Futuros análisis no sólo incorporarán variables adicionales que ya existen en nuestro conjunto de datos (fecha de los hechos en adición a la fecha de los documentos, información sobre autores y receptores, y decisiones de políticas), sino que además incluirán los 12,000 documentos adicionales que se han codificado en la iteración 10. Las iteraciones muestrales subsiguientes se apoyarán en las lecciones aprendidas a partir de este estudio para poder continuar simplificando los procedimientos de muestreo y ponderación. Finalmente, esperamos realizar futuros proyectos de muestreo más específicos y enfocados en períodos de tiempo más delimitados y/o en estructuras de especial interés a partir de los patrones que se han identificado en este trabajo.

### Referencias

1. Comisión de Esclarecimiento Histórico (CEH), (1999). *Memoria del Silencio*, Volumen I, página 50, párrafo 95.
2. Guzmán, D., Guberek, T., Shapiro, G., Zador, P. (2009), “Estudiando Millones de Documentos Rescatados: Plan de Muestreo para el Archivos Histórico de la Policía Nacional de Guatemala”, en *JSM Proceedings*, Sección de Métodos de Investigación Pericial (Survey Research Methods). Alexandria, VA: Asociación Americana de Estadística (*American Statistical Association*, por su nombre en inglés).
3. Shapiro, G., Guzmán, D. y Zador, P. y Guberek, T. (2009), “Ponderación de la Muestra del Archivo Histórico de la Policía Nacional de Guatemala: Retos y Problemas Inusuales”, en *JSM Proceedings*, Sección de Métodos de Investigación Pericial (Survey Research Methods). Alexandria, VA: Asociación Americana de Estadística.
4. Silva, Romesh (2002) “Acerca del Mantenimiento y Medición de la Fiabilidad entre Codificadores al Documentar Violaciones de Derechos Humanos a Gran Escala”. *Actas de las Reuniones Conjuntas de Estadística de la Asociación Americana de Estadística, la Sociedad Biométrica Internacional (ENAR y WNAR, por sus siglas en inglés), el Instituto de Matemática Estadística y la Sociedad de Estadística de Canadá*.
5. Lumley, T. (2009) “Encuesta muestral: análisis de encuestas con muestras complejas”. *Paquete R versión 3.16*.
6. Quenouille, M. (1949) “Aproximación a las pruebas de correlación en series de tiempo”, *Revista de la Real Sociedad de Estadística (Journal of the Royal Statistical Society, por su nombre en inglés) B.11*, 18-84.
7. Ball, P., Kobrak, P., y Spierer, H.F. “Violencia de Estado en Guatemala, 1960-1996: Una Reflexión Cuantitativa”. *Asociación Americana para el Avance de las Ciencias Sociales y Centro Internacional para la Investigación en Derechos Humanos (CIIDH)*.
8. *Proyecto Interdiocesano de Recupeación de la Memoria Histórica (REHMI). Guatemala (1998). “Nunca Más”. Guatemala, ODHAG.*

Traducción del inglés al español: Carla Villagrán, Traductora jurada  
Traducción revisada por Beatriz Bejarano, Benetech.