Chapter 7

The International Center for Human Rights Investigations: *Generating Analytical Reports*

Herbert F. Spirer

Introduction

In this case study, I review my work in conducting the analysis of the data and generating the graphs and tables for the joint International Center for Human Rights Investigations (CIIDH) and American Association for the Advancement of Science (AAAS) report on Guatemala (Ball, Kobrak and Spirer, 1999). The purpose of the report was to use statistics in conjunction with historical analysis to tell the story of state violence in Guatemala from 1960 to 1996. The published report of 154 pages contains 42 graphs, 9 tables, and numerous statistics appearing in a text that largely reflects the information in the figures. Despite the small number of graphs and tables in the final report, it was informed by many hundreds of figures, analyses, and statistics, created over a nine-month period.

I give a summary of the lessons learned from this work, and make recommendations to help others working on similar projects. The project organization, analytical tools, and working relationships used on this project are generally related to those used by industrial analysts. In view of the growing use of large-scale datasets in human rights, I expect that with time the human rights field will develop its own approaches to data analysis. This paper is intended to be a contribution to that developmental process.

The statistical methodology (described below) used in the CIIDH/AAAS Guatemala project work is straightforward and well established; neither sophisticated nor novel methods were used. Because of the need to maintain the highest standards of credibility, the dominant issue in the statistical analysis was the avoidance of error and control of the process of generation and use of analyses. For that reason, my focus in this case study is to show how we met that need.

I believe that we were effective in meeting the standard of credibility necessary for a human rights report that established state responsibility for political violence. Other workers in this field should be able to use knowledge of this case study to achieve the same standard, and may do so more efficiently. In the section **Lessons Learned**, I review the lessons learned on this project, make recommendations, and discuss how those lessons could be applied in future projects.

There was considerable similarity between the process of generating analytical reports for the CIIDH and CEH projects as carried out by the analysts. Accordingly, the analysts for these two projects, Eva Scheibreithner and myself, jointly prepared recommendations for future large-scale human rights data analysis. These recommendations appear in Appendix 1 of Chapter 10, Data Analysis Recommendations.

Preliminaries

Data Analytical Methods in the CIIDH Report

Descriptive statistics

Descriptive statistics are measures that summarize and describe the overall characteristics of a dataset. In general, this refers to a set of well-known statistics derived from the data set that describe one or more variables. For this project, the descriptive statistics are the number of observations, mean, variance or standard deviation, minimum, maximum, median, and the sum for each of the variables in the dataset. I used descriptive statistics in this project largely as a means of detecting, correcting, and avoiding error.

Exploratory Data Analysis

Exploratory Data Analysis (EDA) is "about looking at data to see what it seems to say. It concentrates on simple arithmetic and easy-to-draw pictures. It regards whatever appearances we have recognized as partial descriptions, and tries to look beneath them for new insights." (Tukey, 1977:

p. v.) EDA uses the statistical measures of descriptive statistics, and in addition a number of other methods that involve creative analysis and interpretation. These methods include tabulations and crosstabulations, time series plots, scatterplots, transformations of variables, autocorrelation analysis, regression analysis, difference analyses, and others. Generalizations drawn from EDA may be extended to a larger universe, but cannot be given meaning in terms of mathematical probability.

Inferential statistics

In statistical inference, the analyst generalizes from sample data to make probability-based statements about the larger universe from which the data were obtained. These probability statements are usually expressed in hypothesis test results or as confidence intervals. For example, in the CIIDH report, Appendix 5, Monthly Seasonal Variation Analysis, I use a hypothesis test to infer that the observed monthly seasonal pattern of killings and disappearances is essentially certain to have been caused by an organized plan.

The Data

The CIIDH database is a relational database consisting of cases culled from press sources, documentary, and direct testimonies. CIIDH team members collected over 10,000 cases from newspapers, by reading every newspaper published during the 36-year period of armed conflict in Guatemala. Four thousand additional cases came from documentary sources such as the archives of several Guatemalan non-governmental organizations and the publications of the Justice and Peace Committee of the Guatemalan Church in Exile. Members of the CIIDH team directly collected over 5,000 testimonies for inclusion in the database.

We define a case as the information given by a single source (press report, or interview, or document) concerning violations that are reported as having happened at a particular time and place. "Violations" are instances of violence, including killings, disappearances, torture, kidnapping, and injury. "Victims" are people who suffer violations. A case may be simple (one victim who suffered one violation) or complex (many victims, each of whom suffered many different violations). In the CIIDH analyses, the unit of analysis is almost always the violation.

The basic data with which I worked were contained in four flat datasets (two-dimensional tables of information without established relationships to other tables), each with variables chosen from a common set of variables. Complex Structured Query Language[1] (SQL) queries and extensive programming produced these datasets with variables selectively chosen from the listing of variables shown in the data dictionary of Appendix 1. Unfortunately, the variables did not keep the same definitions in all data sets.

The four basic datasets were denoted by ctanon, ctnmd, rtanon, and rtnmd as indicated in Appendix 1. In this terminology, the prefix "ct" denotes complete, in that these are the data net of overlaps among data sources (interviews, documents, and periodicals). The prefix "rt" denotes reduced, in which the source categories "other" and "non-CIIDH interviews" were folded into the "documents" category.

The suffix "anon" indicates that the dataset consists of both anonymous and named violations for which victim identification exists, and the suffix "nmd" indicates that the dataset consists only of precisely named violations.

I also worked with four additional datasets in which only killings appear with additional variables to describe the nature of the killings and the size of the group in which they occurred. These datasets carry the additional suffix "k". All datasets were followed by "v" with an integer (1, 2, 3,…) suffix to indicate the version of the dataset. By the completion of the report, the version number had reached 16. Table 1 is a summary of the datasets.

Table 1. Datasets.

Data set name	Description
ctanon	Complete, anonymous plus named

[1] Structured Query Language is a computer language used to retrieve, update, and manage data.

rtanon	Reduced, anonymous plus named
ctnmd	Complete, named
rtnmd	Reduced, named
ctanonk	Complete, anonymous plus named, killings
rtanonk	Reduced, anonymous plus named, killings
ctnmdk	Complete, named, killings
rtnmdk	Reduced, named, killings

Data Control Documents

Data control document refers to the data dictionaries, dataset descriptions, variable position dictionaries, and derived dataset descriptions.

By *data dictionary*, I mean a tabulation of the names of each field, the values that can appear in each field, and a verbal description of the meaning of each field variable. Some data dictionaries include the dataset list. However, for my purposes, I separated the descriptions of the data from the description of the files. As mentioned earlier, Appendix 1 is the data dictionary for this project.

The *dataset description* includes as a minimum the name of the dataset, the number of records, and in this case, the number of violations. In some cases, I also included the number of killings. Appendix 2 is an example of such a description.

By *variable position dictionary* I refer to a summary by dataset of the columnar position of variables, which implicitly shows whether a variable is in a particular dataset. The need for this control document was because of the use of Excel for statistical analysis. Appendix 3 shows a variable position dictionary.

The *derived dataset description* includes as a minimum the names and brief description of datasets derived from the source datasets discussed under Background, The Data. It may also include Excel versions of the underlying basic data that were received as a file in xBase format, with the extension dbf. The derived dataset description also includes the source data, where relevant, and comments. Appendix 4 shows part of a typical derived dataset description, with my original footnoted comments.

I updated all of these documents with successive versions.

I used control documents on this project for two purposes. First, I needed them to keep track of the rapidly growing number of files and versions and the field names and values, which were changed during the analysis phase. Second, they played a role in checking for error. As will be discussed in detail later, every new version and revised configuration for a working file was tested with respect to its predecessor. When I could directly predict the expected effect of a change, I used these documents to verify that the expected changes occurred or to explain their absence.

Checking for Data Integrity

Statement of the Problem

There are many challenges to the integrity of the data. I considered every transition involving a dataset a potential source of harmful alteration of the data. Transitions are events in the transfer, conversion, and use of the data. Most of these events occur in the use of data in any form of analysis. For example, the analysis of data can reveal inconsistencies, outliers or suspicious results that result from errors in the working data set. These errors must be corrected and thereby result in new versions of the working datasets.

However, as described in detail in this section, most of the transitions are the result of the overall methodology of this project. Throughout the analysis period, this project was a work in progress, with a strong research component. The results of a particular analysis could reveal fea-

tures that were not anticipated, calling for a recoding or revised query of the dataset. Consequently, transitions were frequent.

For this project, the general sequence of transitions was as is shown below in steps 1-12:

1. Patrick Ball (PB) creates a file.
2. PB transmits the file to me (HS) in a .dbf or .xls format, whichever was convenient for one or both of us as an e-mail attachment. We had to use both formats because of initial conflicts in Excel versions. My actions would then be to:
3. Download the file in native form and archive it.
4. Convert the file to Excel format and archive it.
5. Make a working copy of the file.
6. Filter, reorder, consolidate, summarize, and otherwise manipulate the data to facilitate a desired analysis.
7. Perform the desired analyses.
8. Transmit the results to PB.
9. Create the graphs.
10. Transmit the graphs to PB.
11. Revise the graphs in accordance with format and analytical needs through joint exchanges with PB and Paul Kobrak (PK).
12. Transmit the graphs as attachments by e-mail.

The likelihood and form of the data integrity challenge at a transition is dependent on the transition and the circumstances. For example, I cannot recall an instance in which we found an error resulting from download transmission or format conversion (2-5, 8, and 12, above). However, I only know that these transitions were error-free because I was checking the results. I had many errors – often minor -- develop in the other transitions, which were detected and corrected. Our concern for even minor errors was to avoid the possibility of any negative effect on our credibility.

There were also challenges to the integrity of our results that relate to handling the data. For example, Excel apparently has internal instabilities, or as yet undocumented capacity limitations. On a number of occasions I returned to a workbook several days or weeks after creating graphs and found that the graph had disappeared or that formatting features were altered. I never had this problem in a small worksheet. Archiving the original data and any revised datasets that entered into analysis is essential. However, this action is another transition where the integrity of the dataset itself is in jeopardy from the failure to archive the latest version, or the inadvertent deletion of a file.

Throughout the process described above, I carried out different levels of checking, as I judged appropriate, as discussed in the next section.

Verification Methods

My approach to verification is based on applying descriptive statistical methods to the dataset or pair of datasets (before a transition and after a transition). By definition, summary statistics reduce the information content of the data to facilitate an understanding of the whole set. I show the descriptive statistical measures used for numerical and categorical variables listed in Table 2.

Table 2. Descriptive statistics used for verification

Numerical	**Categorical**
One-way tabulation	One-way tabulation
Record count	Record count
Crosstabulation	Crosstabulation
Extremes (high, low)	
Mean	
Median	
Sum	

For a single dataset, I look for reasonableness in the values. For example, if a dataset contains the variable SEXO for gender, a one-way tabulation should show some number of males and females, which may be coded "m" and "f". What other value might reasonably appear in the tabulation? If there has been agreement on the representation of unknown gender values as d (for *desconicido*), then we expect some number of d's to appear in a one-way tabulation. If I find no d values but a number of –1's, then I suspect that there may have been a change in the assignment of unknown values in this dataset. Of course, this would have to be reconciled.

But if **both** –1's and d's appear, then something is seriously wrong. It may be miscoding or a more fundamental problem. Or perhaps, the tabulation includes blanks. What might be signified by a blank, a missing value that was not properly coded or entered, an input error, or a blank record (which may reflect a serious error)?

With two datasets – one before and one after – I look for a reasonable comparison in the values. If there are two datasets, and the second is one in which records have been removed from the first dataset described above, then only m, f, or the missing data value should appear, and in no case with a higher count than in the first set.

Extreme values of numerical variables (maximum, minimum) can be a symptom of a problem. If there are a large number of numerical values, a one-way tabulation is usually more confusing than revealing. Extreme values may be outliers in the sense that they either are unreasonable or differ greatly from the normal range of deviations. For example, although –1 might be used as a missing value indicator for ages, what do we make of a –2 also appearing in the dataset? Is a maximum age of observation of 95 an error?

Comparison of the median and mean values is a quick way to determine skewness of the distribution of numerical data. To carry out this comparison, the analyst needs a sense of what the distribution of the data is, or should be, or how it would be changed by some transitional step using before and after comparisons.

The sum of columns is a simple check and it is easy in Excel to maintain sums of numerical fields at the bottom of the dataset. I monitored sums and record counts on a continual basis while working with a particular dataset. Using the sum on a continuing basis is a process that has its own problems, because of the automatic selection of data by Excel for certain procedures, and my own errors.

Many of the desired analyses are crosstabulations, and in themselves provide a basis for checking the dataset integrity. While I infrequently made crosstabulations as a check on a dataset, I almost invariably compared marginal totals in crosstabulations to the values produced by independent one-way tabulations.

It is tempting to think of automating these checks and verifications to reduce the dependency on human intervention. Without automation, some person has to make a conscious effort to carry out the check. But with automation, you may have another source of errors and lose the judgmental insight that can only come from knowledge of the data and what its attributes should be, or are

most likely to be. Accordingly, during the project, I used only the Excel built-in functions (where appropriate) to obtain values for the verifications discussed above. However, as will be discussed in Lessons Learned, I also used an intermediate approach.

In the final analysis, human intervention is critical. In one case, the routine checks suggested a possible incorrect coding. To track this down, I visually scanned approximately 10,000 records by observing the "play" of the patterns on the screen as I scrolled rapidly through the dataset. Using this method, Patrick Ball found a coding error, traceable to key entry at the source.

There is no substitute for vigilance and scrutiny.

Examples of errors

The following are a few examples of errors that were detected in the process of analysis. Some related to problems existing in the database or the query process. These have significance for project personnel other than data analysts and therefore have general interest and applicability in the management and implementation of the information system. However, the overwhelming majority of errors were the results of my own actions, occurring on a continual basis and which, by and large, can only be generalized to the need for each individual analyst to work constantly at avoidance, detection, and correction of error.

Early in the project, time series plots showed a midyear peak in violations with a clear, pronounced peak of violations in the sixth month, June. At first we were concerned only with revealing this pattern, but attempting to find out why such a pattern should exist led to the investigation of the coding process by which violations were assigned to a particular month. When the precision of the date of the violation was one year (that is, the violation could only be placed somewhere or at some time within a particular year), the violation was arbitrarily assigned to June. This resolved the problem of giving it a date, and would not affect any analysis of annual patterns. However, when the data were summarized by month across all years, the number of violations in June was improperly inflated by violations that could have happened at any time during the year.

When analyzing the patterns of collective and individual killings that required the use of named datasets, I routinely summed killings by individual and obtained the maximum and minimum values in the column of sums. A minimum of zero would indicate the presence of a zero due to one or more entry errors, corruption of a cell, or records that should not have been in the data set. A maximum above 1 would indicate miscoding, entry, or corruption errors. Two different cases were uncovered by this check:

1. In the early phases of analysis, I found instances where an individual was reported as suffering more than one death. This anomaly resulted from more than one source of data reporting an individual's death. This problem was traced back to duplicate reporting leading to miscoding.
2. In another case, the same individual was reported as killed by the same source at different dates. This is a genuine error, but I found only one.

In the data description associated with a dataset and the data block associated with an analysis, the number of violations is reported. The dataset rtanonkv7 contains only killings and hence, its violations total should have been the same as the count of killings in its source dataset, rtanonv7. Observation revealed that it was not the same, 34,747 compared to 34,659, a difference of 85! While this is an error of only 0.2%, we could not overlook it for reasons of credibility and because it might reflect larger compensating errors. On examination, Patrick Ball found that those 85 death records were reported as more than one of the three death killing categories -- cadavers, individual, and collective. His new program brought the two totals into agreement.

Early in the analysis, a one-way tabulation of ages in the named dataset showed ages of 0 and -1. Both values had been used to represent missing values of age. Conflicts in the number of missing values found at the same time were traced back to a revision of the coding process that caused the loss of the ages of 540 people (out of about 10,000, depending on the dataset).

Performing the Analyses

Statement of the Problem

In the data analytic aspects of this project, our goal was to describe, summarize, and explore the data. By and large, our mission was not to infer some parameter from a sample but to reveal the facts inherent in our data. The broader interpretation of these facts derived from an incomplete

coverage of the actual events must – as it did – come from the interplay and conjunction of the quantitative knowledge gained from the data and the equally relevant anecdotal and qualitative knowledge of subject matter experts.

Our data could not be obtained by probability sampling, which would have enabled the use of inferential statistics and its related disciplines of statistical hypothesis testing and confidence interval estimation. However, we did use probabilistic approaches to evaluate the apparent monthly pattern (Ball, et. al, 1999: Appendix 1).

Thus, most of the tools that are usually called "statistical methods" in the educational process and in much research did not apply to the analyses used in the body of the CIIDH report. The challenge in this project was to apply simple methods to complex, large-scale datasets in such a way that the voice of the data is heard and understood by both the knowledgeable members of the project team and the lay audience of CEH, researchers, and the interested public.

Methods of data analysis

Accordingly, we used **summary statistics, tables** and **graphs** as our primary tools of analysis. In Excel, graphs are called "charts," reflecting the orientation of Excel to business applications. Most statistical programs (e.g., Stata) call them graphs, as we do in this case study.

Our use of tables did not extend beyond the two-way crosstabulation. In our analyses, we used graph formats (e.g., logarithmic axes) and types (e.g., scatterplots) that we did not present in the final report. In fact, with few exceptions, the tables and graphs appearing in the report are fully described in the AAAS/HURIDOCS handbook, *Data Analysis for Monitoring Human Rights* (Spirer and Spirer, 1993). One exception is the **comparative histogram** that relates absolute and relative rates of killing by age (Ball, et. al., 1999: Figure 16.2), another is the time series plot of percent of victims by age that uses stacked line plots (Ball, et. al., 1999: Figure 11.4).

Since readers of this paper may want to relate our approach to the formal discipline of statistics, we reiterate that we have used the tools of *descriptive statistics* -- describing, presenting, and summarizing data to reveal or gain a better understanding about the processes that created the data. Exploratory Data Analysis (EDA) is a related set of techniques for understanding, analyzing, and presenting data, its structure and systematic patterns (Tukey, 1977). Easily understood by non-professionals, these methods have much to offer in the adversarial human rights environment. Their effectiveness has been demonstrated, as in (Hoaglin and Velleman, 1995: 277):

> Our examination shows that approaches commonly identified with Exploratory Data Analysis are substantially more effective [than a long list of advanced model-fitting methods] at revealing the underlying patterns in the data and at building parsimonious, *understandable* [my emphasis] models that fit the data well.

Graphs: The Visual Display of Information

Our approach to the visual display of our analyses conforms to Tufte's standards for Excellence in graphical representation (Tufte, 1983: p. 13):

…. Graphical displays should

- show the data
- induce the viewer to think about the substance rather than about methodology, graphic design, the technology of graphic production, or something else
- avoid distorting what the data have to say
- present many numbers in a small space
- make large data sets coherent
- encourage the eye to compare different pieces of data
- reveal the data at several levels of detail, from a broad overview to the fine structure
- serve a reasonably clear purpose: description, exploration, tabulation …
- be closely integrated with the statistical and verbal description of a data set

Some of the goals above can be achieved through formatting. Accordingly, in the final version of the report, the three authors agreed to a spare, uniform format. There are no gridlines and no frames on any of the graphs, as these embellishments add nothing to comprehension.

In time series plots, I omitted point markers, used ticks sparingly, and only major values are labeled and associated with ticks. Appendix 5 shows a typical time series plot. Appendix 6 shows that a spare style does not mean that complex relationships are not portrayed.

Not all of Tufte's recommendations can be achieved through formatting. Graphical presentation and analysis are interdependent. Table 3 shows the frequency distribution for the types of graphs used in the final report. This is by no means a summary of the graphs used during the analysis, but indicates the types of graphs that are likely to be used in an analytical report on human rights violations for this type of audience.

Table 3. Frequency distribution of graph types

Type	Number of dependent variables	Number of graphs
Time series plot	1	18
Time series plot	2	6
Vertical bar chart	1	6
Time series plot	3	2
Horizontal bar chart	1	2
Horizontal bar chart	3	2
Vertical bar chart	2	2
Stacked line plot	3	1
Stacked bar chart	2	1
Histogram	1	1
Comparative histogram	2	1
TOTAL	—	42

Analysis: The What, How, and Who

Analysis for this project was an iterative process, inseparable from the creative interaction of two and sometimes, three persons. For any consideration of the analysis process, keep in mind that the figures and many textual statistical references of the report are a small fraction of the total number of tables, summary statistics, and graphs that were produced during the analysis process.

As the preface of the report itself states, Patrick Ball designed the analyses, and I carried them out. However, this was not a rigid hierarchical process. As is often the case in analysis, my instructions might be as vague as "see what the relationship is between X and Y," or as precise as "make a bar chart for X and Y for A and B, with A on the left and B on the right." While the general flow of instructions was indeed from him to me, there was interaction in both directions, one analysis leading to another in an ongoing, often iterative process. I did not always confirm expected relationships, and unexpected results were frequent occurrences.

What kinds of decisions did we have to make? Given that an analysis led to a significant result (in the sense that it was worth passing on to the reader), how can we most effectively present it to the reader? Sometimes, issues of modeling were involved. Modeling is integrated with presentation – presentation for us and for the reader. For example, the annual number of killings of women is a highly skewed distribution. When analyzing such a distribution, an analyst's first instinct is to

transform to the logarithm of the variable. This transformation makes it possible to view all values without an indecipherable cluster at the low end of the axis. If the transformed distribution is normal, have we learned anything about the process?

In this case we know that the number of killings has a skewed distribution because of factors pertaining to the actions of the state and the skewed high end results from the actions of the Laugerud García and Ríos Montt regimes. If no comparison is being made (for example, between the number of killings of males and females, where the high ends differ by a large factor), there is no good reason for using the logarithmic transformation. A logarithmic transformation is often used to normalize a skewed data distribution in order to use the methods associated with normal distributions; in this case there is no need for such a transformation except to make the scale of values visually tractable.

The relevant model for the time series of violations is analogous to the standard model of industrial quality control. In this model, many sources of variation common to all data points (called "common causes") accumulate to give a background level of random variation. In terms of this model, the time sequence of killing in Guatemala has a "background" level due to common causes. One or more significantly large deviations from that level would be denoted as due to "systematic causes." The analyst then searches for the systematic cause, which in this case usually is the imposition of state policies.

Control and Traceability

Which analyses have been or are to be carried out? What scientific questions do we want to answer? What is the status of analyses in process? Who is responsible for particular data or results? What is the reference identification for a particular analysis? These are the questions that pertain to control of the analysis.

My analysis control documents followed the progression of the project from exploration to preparation of a final report. In the first phase, Patrick Ball was proposing exploratory analyses and I was producing them. This was an interactive process as is reflected in the control document. By and large, his proposals were received in text e-mail messages and it was my problem to keep track of them and respond either with the result or an additional query. Appendix 7 shows the control document used in this phase, the *Reconciliation of Instructions.* I include a fragment of this document to reveal both the format and the interactive nature of both the process of analysis and of tracking instructions and results, both intermediate and final. The instructions shown in this document are taken from e-mail or verbal instructions.

The second phase was after we incorporated Paul Kobrak into the process and we determined the final structure of the report. We now had to maintain tight tracking of the progress and in particular, tracking the ongoing revisions of both chart and figure numbers. The control document for this phase was the *Figure List*, a typical version of which is shown in Appendix 8.

An analysis results in a table or graph that we would ultimately identify by a figure number in the final report. However, many hundreds of analyses were performed. We recognized early in the project that there was a need to associate a minimal set of data with each analysis. These factors were that:

- Figure numbers were context sensitive and in a state of flux until completion of the final report.
- The creation of a table or graph and assignment of a figure number was another transition that might produce error.
- The dataset version used in the analysis was a moving target due to ongoing revisions.
- Excel in itself produced no record of the source dataset (workbook, worksheet) or the process of analysis.
- Certain summary data was relevant to every analysis (e.g., number of violations included in the analysis). For example, the number of violations or other count of units entering into the analysis is essential to our evaluation of the results as we passed from draft to final copy.
- The analyst identification is needed to determine source of analysis for questions or revisions.

- The date of the analysis provides a control for error. For example, the date of the analysis must be consistent with the date of the source dataset. Also, the date reveals whether the latest desired version of the analysis has been performed by reference to the Figure List.

Thus, Patrick Ball proposed, and I agreed that I would associate an *informational data block* with each analysis. This block was attached to every analysis until the final report was prepared. It took several iterations before we fixed on a standard format, shown below:

Date of analysis [*date*]
Analyzed by: [*analyst*]
Records included: [*count of the records used in the analysis*]
Violations included: [*count of violations used the analysis*]
File Reference: [*workbook(s), worksheet(s)*]

The data block count of the records used in the analysis was not necessarily the total included either in the source dataset, the workbook derived from it, or the particular worksheet. It was the number used to perform the particular analysis. Of course, I made errors in the data blocks and had to apply the same constant verification as in the case of the datasets themselves. However, in the long run, these data blocks proved to be invaluable in verification and in finding a way among the many dozens of subsidiary workbooks and sheets when I needed to revise or verify and analyze.

Backup

I started with a simple backup strategy. I backed up my work locally on removable disks and each week mailed a complete compressed copy of my project files on a 100MB ZIP disk to the AAAS for archival storage. Since we finished the project and have been able to create full electronic archives at the end, this aspect of the project could be considered a success. We can trace any analysis in the final report to a figure including a data block, and hence, reconstruct the original analysis.

However, my inconsistent directory structure and file naming conventions made this more difficult than it should have been, as will be discussed in the following section, Lessons Learned. These problems came in part from the fluid nature of the project, which was essential to a creative process.

Lessons Learned

In a successful project such as this one, the retrospective issue is to set the stage to carry out successful projects in the future. By showing what we did in the preceding sections, I hope that others will get guidance for their own future work. In this section, I specifically target functions and methods that worked well, and those that did not work well, in order to make recommendations that can be applied both by others and us in similar large-scale human rights data analysis projects.

Large-scale analysis of human rights data rarely occurs in the same environment twice; it is much closer to social science research than industrial statistical analysis. A common issue in applying lessons learned to recommendations is a tendency to introduce central control, uniformity, standard procedures, and conformance to rules as a way to improve efficiency and effectiveness. This is a valid approach in situations where **control** is important. On the other hand, freedom of action and tolerance of diversity is vital to **creativity**. I regard the establishment of the appropriate balance between these two poles is the major administrative and personal challenge that we face. My own preference at this stage is to lean toward promoting creativity. As a minimum, each contributor should have a unique individual approach to resolving the common problems – but in such a way that other team members can access and comprehend his or her work.

Another common and general issue is self-discipline. If you set up a rule for naming files or a procedure for backups, and so forth, stick to it. This is not easy when trying to get new answers to new problems under time pressure, but it is precisely those circumstances when lack of discipline will hurt the work the most.

Our lessons learned and related recommendations are summarized in Table 4, following. A more comprehensive jointly authored set of recommendations for data analysis, based on both the CIIDH and CEH experiences appears in Appendix 1 of The Guatemalan Commission for Historical Clarification: *Generating Analytical Reports*, by Eva Scheibreithner.

Table 4. Summary of lessons learned and recommendations

Entity, Function	Lesson	Recommendation	Issues
Data dictionary	Valuable to analyst	Ideal would be a common data dictionary, used and updated by all who create variables	If common, who is allowed to make entries? Who is required to make entries? Should this be networked (private web site)?
Directory structure	A rational project structure would help everyone. Backups from other team members would be comprehensible on sight.	Agree on a project directory structure for common use.	Will a common structure serve all? Can a single structure be used throughout the duration of the project?
File name system	Patrick Ball's dataset naming rules worked well. Mine quickly became unsatisfactory. It was good only for a small-scale project, here no better than sequential serial numbers.	Use appropriate file naming rules that will be understandable to all.	Can satisfactory rules be set at start of project?
Field names	Ambiguity in field names is treacherous	Don't use the same field name for different variables even if appearing in different datasets.	Self-discipline.
Control documents	I can't work without the control documents described in this report	Some people need control and some don't. Do what fits you.	Finding approaches to shared documents that are mutually satisfactory to team members working together.
Update of control documents	If you don't keep your records updated, you may be sorry.	Don't end up being sorry.	Self-discipline.
Errors, transitions	Human, machine, program, transmission errors happen	At every stage, be vigilant and scrutinize.	Self-discipline. See also, Facilitating error checking and verification, below.
Backup	Backup of files is a Good Thing	Have individual and project backup system.	Present system seems satisfactory; is it good enough for the next project?
Software	Different software, different versions lead to unnecessary inefficiencies and errors. I had to switch both computers and software versions to match his. These transitions caused a number of problems.	Have team members working together use the same programs and versions from the start.	Agreeing on software and versions at project start. Cost of upgrading team members' resources. Site licenses and project-owned hardware: reasonable approaches?

Audit trail of procedures	No way to know what series of edits and operations have been performed in Excel	Don't use Excel for analysis in large-scale data projects.	Which programs to use for statistical analysis? Is Stata, for which AAAS has license, the statistical software of choice? Can we get adequate graphical output from Stata?
Graph objects	Encapsulated Postscript Graphics would have been a lot easier to work with in final report than Excel pictures.	Make sure that the statistical software produces graphical files that facilitate report production.	Is Stata good enough? Do we have to consider other alternatives?
Analysis	It is a nuisance to have to integrate a revised version of a dataset in Excel.	It would be a Good Thing to be able to re-run a set of analyses on a new data set without concern for changes in number of variables or records.	Choose and agree on a statistical program that will do this.
Variable formats	It is not a Good Thing to define numerical variables to have textual values.	Don't do it!	Be careful with variable definitions.

Facilitating Error Checking, Maintaining an Audit Trail, and Updating

Stata and other statistical programs (e.g., BMDP, SPSS) have functions that make error checking and verification of datasets semi-automatic and more reliable. In addition, they produce logs of the edits and operations, outputs, and instructions used to achieve results. In Stata, the commands, *codebook, describe, list,* and *inspect,* have value in error checking and verification. The Stata commands *log* and *edit* enable the analyst to maintain a log of the steps in producing a given result. I will give a brief summary of what these Stata functions can do for the analyst. This is not the place to give a tutorial on Stata, since we do not have agreement that it will be the software of choice.

codebook examines the variable names, labels and data to produce a codebook describing the data. You can determine the pattern of missing values, automatically obtain summary numerical statistics for continuous variables and tabulations for categorical variables. (Stata, 1997: v. 1, p. 151-4)

describe produces a summary of the contents of the dataset. You can list the variable names in a compact format (Stata, 1997: v. 1, p. 206-10).

list displays the values of the variables (Stata, 1997: v. 2, p. 335-7).

inspect produces a summary of numeric variables. It reports the number of negative, zero, and positive values; the number of integers and non-integers; the number of unique and missing values; and a miniature histogram (Stata, 1997: v. 2, p. 271-3).

log enables the user to maintain a log of both commands and results (Stata, 1997: v. 2, p. 341-2).

edit adds to the log changes made to the data in the Stata editor (Stata, 1997: v. 1, p. 251).

A Stata log can be converted to a procedure that can be applied to new versions of a dataset without modification, facilitating upgrading to a new version.

No software is without problems, and I am not proposing that Stata (or any other program) is a panacea. If the user does not have a commitment to vigilance and scrutiny to the extent described above, these or similar procedures will be unused or poorly used. If the user does not exercise judgement, intuition, and creativity in dealing with the data, the choice of software will mean little. If logs are poorly edited and maintained or cannot be properly accessed and associated with results because of poor control and naming conventions, they will be of little value.

For example, when the user is engaged in the disorderly but creative process of searching for the appropriate data configuration and analysis, the Stata log can become large and cluttered with false starts. To reduce this log to the final set of procedures can be time-consuming. For tracking logs by name, the Stata team itself has nothing better to offer than date stamping. This is not good enough for our work. At the minimum, we would have to produce logs that are subsets of the daily log. For the individual and possibly for the project, a systematic organization of final procedures will be needed.

Stata has problems and bugs, and there is an active list server in which users share experiences. The descriptive commands discussed above are sensitive to the ways in which data are defined in the data set, and can give incomplete and confusing results. As always, the functions, advantages, and disadvantages of using different software must be evaluated to find the balance appropriate to the job at hand. It may well be that for all the problems and inefficiencies, the software and methods used on the AAAS/CIIDH project were somewhere near the optimum, given that the project goals were met in a timely fashion.

Conclusion

In this report I reviewed my work on the CIIDH/AAAS report, summarized the most important lessons learned, and made recommendations for work of this nature on future projects.

I know that this summary will help me to do a better job on the next similar analysis project. I hope that it will also help others, and in that spirit, close with this quote:

> After spending many years with the Estonian Cancer Registry [I] now think more intensively about data quality than about the application of refined statistical and cartographic methods to data analysis.[2]

[2] From a book review of *Global Geocancerology: A World Geography of Human Cancers*. *The Scientific American*, Feb. 1987, pp. 27-31.

Appendix 1

AAAS/CIIDH Data Dictionary

Data Dictionary			**Appears in file:**			
Field name	**Meaning(s)**	**Values –meaning**	***ctanon***	***ctnmd***	***rtanon***	***rtnmd***
AGE	Age of victim	integer >= 0 -1 – missing value		x		x
C_NMD	Number of victims having age and gender	integer = 1		x		
C_TOT	Number of victims including not named and without age or gender values	integer = > 1		x		
CERTFECH	Precision of dates	1 – day 2 – month 3 – quarter 4 – semester 5 – year 6 – decade 7 – season, no year 8 – no idea of date	x	x	x	x
CIV_CIV	Civilian perpetrators present	0 – not reported	x	x	x	x
EST_EJR	Army involved	0 – no 1 – yes		x		x
EST_EJR	Number of violations in which Army involved	integer = > 0	x		x	
EST_PAC	Civil patrollers involved	0 – no 1 – yes		x		x
EST_PAC	Civil patrollers involved	integer = > 0	x		x	

EST_POL	Police involved	0 – no 1 – yes		x	x	x
EST_POL	Number of violations in which Police involved	integer = > 0	x			
FW	Number of violations for a given case	integer >= 1	x		x	
MON_VLCN	Month in which violation occurred	1 – January … 12 – December	x	x	x	x
MONYEAR	Month and year of violation	59-03 to 95-12	x	x	x	x
OVERKILL	Presence of additional indignities to a victim either killed or being killed	0 – not reported 1 – reported present	x	x	x	
P94_NAC	1994 national census population for municipio of birth	integer > 0 -1 = missing		x		x
P94_VLN	1994 national census population for municipio of violation	integer > 0 -1 = missing		x		x
PAR_PAR	Paramilitary present at violation(s)	0 – not reported 1 – reported present	x	x	x	x
PRES_URN	Guatemala National Revolutionary Union present at violation(s)	0 – not reported 1 – reported present	x	x	x	x

REGIME_N	Regime in which violation occurred as identified by President	01 – Ydígoras Fuentes 02 – Peralta Azurdia 03 - Méndez Montenegro 04 – Arana Osorio 05 – Laugerud García 06 – Lucas García 07 - Ríos Montt 08 - Mejía Víctores 09 – Cerezo Arévalo 10 – Serrano Elias 11 - de León Carpio	x	x	x	x
REGION	Homogeneous geographical region	01 – Occidente 02 – Costa Sur 03 – Verapaces 04 - Petén 05 – Oriente 06 – Meseta Central	x	x	x	x
SEX	Gender of victim	F – female M – male d – unknown	x	x		x
SVNUM	Serial number of named victim	sv concatenated with 7 digit integer except: SV0050217 (input error)		x		x
TYPE_SOU	Type of source of case	DOC - documentary ENT – CIIDH interview PER – periodical			x	x
U_R	Type of area	r – rural u – urban d – unknown	x	x	x	x

URN_URN	Guatemala National Revolutionary Union involved	integer = > 0	x		x	
URN_URN	Number of violations in which Guatemala National Revolutionary Union involved	0 – not reported 1 – reported present		x		x
V_DEPTVL	Department of violation	department no se sabe - unknown	x			
V_DPMU	Department and municipo of violation	department, municipio otro – no department, other municipio no se sabe, no se sabe – department and municipio unknown	x			
V_IND	Ethnic category of victim	Desconocido – unknown Indigenous Ladino	x	x	x	
V_MUNINA	Birthplace of victim	municipio name no se sabe - don't know		x		x
V_MUNIVL	Place of violation	municipio name no se sabe - don't know otro – other	x	x	x	x

V_ORG	Victim's organizational affiliation	civ-camp – civilian, peasant civ- ddh – civilian, human rights civ – emp – civilian, employee civ-ind – civilian-indigenous organization civ-otr –civilian, other civ-rel – civilian, religious civ-sin – civilian, labor civ-uni – civilian, university otr-otr – other other pol-mil – political, military pol-pol – political, political	x	x		
V_SEXO	Gender of victim	F – female M – male d – unknown			x	
V_TRAB	Occupation of victim	occupation name no tenia trabaj – unemployed otro – other blank – missing		x		
VLCN	Type of violation	Mu – killing Ds – disappeared Se – kidnapping or illegal detention Hr – injury To – torture	x	x	x	x
YR_VLCN	Year in which violation occurred	1959, …, 1995	x	x	x	x

Appendix 2

AAAS/CIIDH Dataset Description

Primary Data Sets[3]

Filename	No. Records	No. Violations	Killings
ftanonkv8	5,601	34,660	34,660
ftnmdkv8	8,968	8,968	8,968
rtnmdkv8	8,964	8,964	8,964
rtanonkv8	5,585	5,585	34,656
ftanonv8	8,242	43,547	34,660
ftnmdv8	13,917	13,917	8,969
rtnmdv8	13,906	13,906	8,964
rtanonv8	8,205	43,535	34,656

[3] Data sets with the prefix "f" for full were not used in my analysis.

Appendix 3

AAAS/CIIDH variable position dictionary

Variable number gives order in data set[4]

Filename	No. of records	svnum	vlcn	mon_vlcn	yr_vlcn	certfech	muni_vlcn	type_sou	c_nmd	c_tot	age	sex	munitot	hompct	v_ape1	cnt	ape2cnt
ftf2t2v1.dta	17,941	1	2	3	4	5	6	7		8							
ftf2t1v1.dta	14,025	1	2	3	4	5	6	7			8	9					
fintaba.txt	13,821		1	2	3		4	5	8	9	6	7					
ft11.txt	14,025	1	2	3	4	5	6	7			8	9					
ft1.txt	14,025	1	2	3	4	5	6	7			8	9					
ft2.txt	17,941	1	2	3	4	5	6	7		8							
mn4.txt,.xls, .dta	34		1				2						3	4			
mnap.txt, .xls, .dta	410		2				3	1					7		4	5	6

[4] Data sets with the prefix "f" for full were not used in my analysis.

Appendix 4

AAAS/CIIDH derived dataset description

(Fragment shown only for illustration)

Filename	Description	Source	
age distribution.xls	Demographic vertical two-sided plot of male, female age distributions for Guate based on 1994 census.		
ageBOX.gph	Primitive box plot of age distributions by violation	ageBYvln.dta	
ageBOX.wmf	Primitive box plot of age distributions by violation		DEL
ageBOXvln simple.doc	Primitive box plot of age distributions by violation		
ageBYvln.dta	ages in columns, one column for each violation type	rtnmdv3AG	
AgeBYvln.txt	ages in columns, one column for each violation type		
Book1.xls	can't recall		DEL
Book2.xls	dummy block		DEL
censo_a full for source.xls	working copy	censo_a	
censo_a.xls	original	censo_a	
Ch05AgeHistOpen with rtnmdv3A.xls	age histogram for Chapter 05 linked to rtnmdvA.xls	rtnmdvA.xls, histogram template	
D1_VctmDemo.xls	PB South Africa automated graph generator		Look at
Dummy for Time Series.xls	dummy block		
fintaba.dta	working copy	fintaba	
fintaba.xls	working copy	fintaba	
Formatted Sections.xls	Contains standard sections. For example, names of presidentes in chronological order, region names, etc.		

Notes:

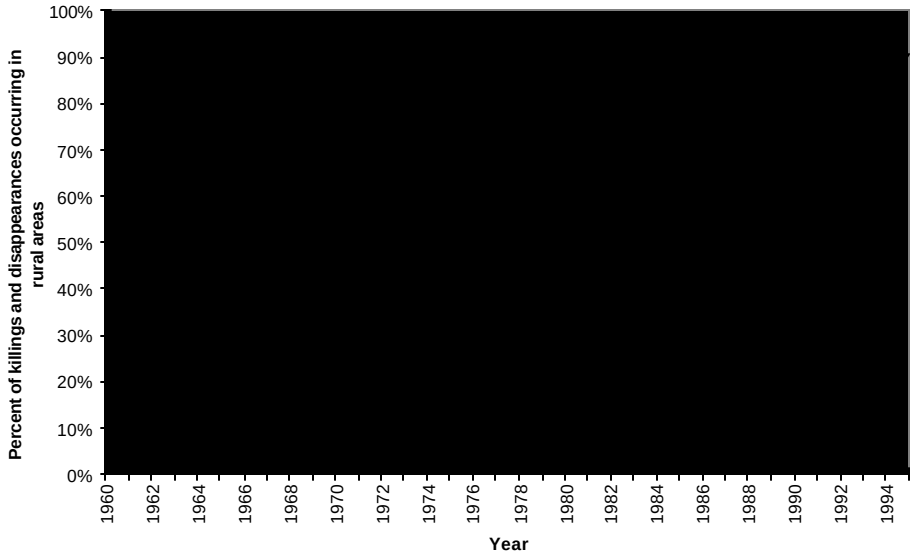
1. If an entry is given in "Source," then the file is derived from an original data set described in Data Set Descriptions.doc.

2. Dummy blocks contain blocks of entries to add to data sets. With these blocks included, Pivot Table will have something to chew on for each year. This makes the year variable continuous and complete.
3. Proliferation of .xls files with same prefix on name was to avoid excessively large files. If I had to do it over again, I would simply number them sequentially.

Appendix 5

Typical AAAS/CIIDH time series plot

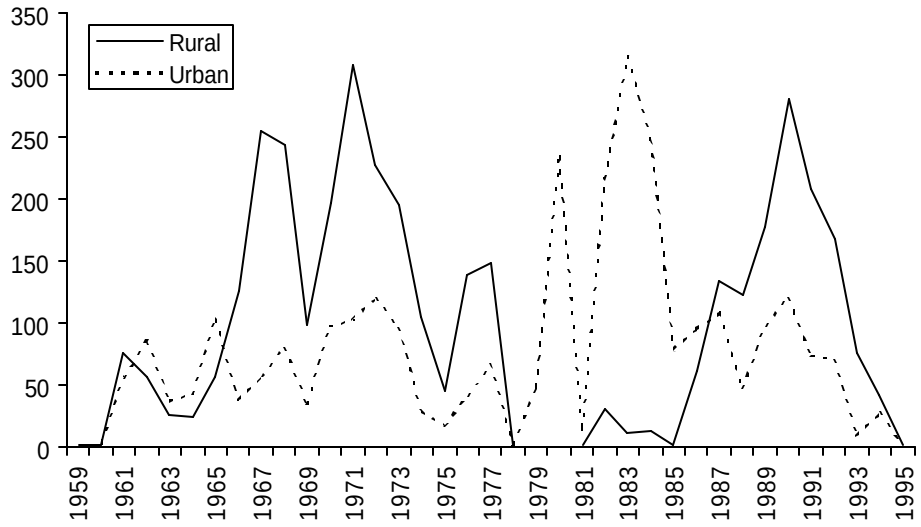
Figure 8.3 Percent of killings and disappearances occurring in rural areas by year, 1960-1995



Appendix 6

Complex AAAS/CIIDH time series plot

Figure 9.1: Number of killings and disappearances reported in the press by area (rural vs. urban) and by year



Appendix 7

AAAS/CIIDH Reconciliation of instructions

(Fragment shown only for illustration)

ID	Date	Chapter	Instruction	Action
01	?	03	Monthly 1st diff sources and violations	not needed
02	?	03 and ?	Time can be broken into segments, 61-78, 79-85, 86-96	Done when appropriate. Use of logarithmic reexpression allows use of all at once when relevant.
03	?	04	monthly first differences for departments, limited to ENT/DOC, for all killing, then COL/IND	Done for annual. Too many months with no violations, too much detail.
04	5 May 98	?	Order of departments by number of killings among three sources and relative proportions	Done in Chapter 4
05	5 May 98	03	Analysis by regime as a kind of time	Done in Chapter 4
06	5 May 98	04	Urban rural differences by regime, separately by violation	Done in Chapter 4
07	5 May 98	04	Killings overall, disappearances overall	Killings done, very few disappearances, so not done
08	5 May 98	03	Compare killings by IND, COL and ENT to PER	Done
09	5 May 98	03	Do regime comparisons at end	Done
10	9 Apr 98	All	Break into numbered sections	Done, needs watching and agreement on style
11	9 Apr 98	03	Numerous position changes of sections marked on draft	Done
12	9 Apr 98	03	Compare one kind of violation by ENT PER DOC sources, a figure for each type of violation	Done

Appendix 8

AAAS/CIIDH Figure List

Figure list as of 24 October 1998

Fig# 24 Oct	Fig# 14 Oct	Fig# 13 Oct	Figure title	Derived from figure, notes	Filters	Dataset	Resp.
01.1	1.1	1.1	Number of killings and disappearances by year, 1960-1995		Net of URNG, certfech<=5	CTanon	HFS
02.1	2.1	2.1	Number of killings and disappearances by year, 1960-1969		Net of URNG, certfech<=5	CTanon	HFS
03.1	3.1	3.1	Number of killings and disappearances by year, 1970-1979		Net of URNG, certfech<=5	CTanon	HFS
04.1	4.1	4.1	Number of killings and disappearances by year, 1980-1989		Net of URNG, certfech<=5	CTanon	HFS
05.1	5.1	5.1	Number of killings and disappearances by year, 1990-1995		Net of URNG, certfech<=5	CTanon	HFS
06.1	6.1	6.1	Number of disappearances and killings, by regime	Figure 3.26, killings only, without means	Net of URNG, certfech<=2	CTanon	HFS
06.2	6.2	6.2	Average monthly number of deaths and disappearances, by regime	this is the other half of Fig 3.26; ordered by regime	Net of URNG, certfech<=2	CTanon	HFS
06.3	6.3	7.1	Number of killings and disappearances by month, July 1979-June 1984		Net of URNG, certfech<=2	CTanon	HFS
07.1	7.1	8.1 & 8.2	Number of killings and disappearances by year and source (press vs. documentary vs. interview)	Fig. 3.7; cut off vertical axis at 700, leaving the DOC and ENT peaks off the graph	Net of URNG, certfech<=5; PER and DOC	RTanon	HFS
07.2	7.2	8.4	Number of killings and disappearances by regime and data source	Fig. 3.31, PER/DOC/ENT	Net of URNG, certfech<=25; PER, DOC and ENT	RTanon	HFS

References

Ball, Patrick, Kobrak, Paul, and Spirer, Herbert, 1999. *State Violence in Guatemala, 1960-1996: A Quantitative Reflection.* Washington: American Association for the Advancement of Science and Centro Internacional por Investigaciones en Derechos Humanos.

Hoaglin, David, and Velleman, Paul, 1995. "A Critical Look at Some Analyses of Major League Baseball Salaries." *The American Statistician* (August 1995), pp. 277-285.

Spirer, Herbert, and Spirer, Louise, 1993. *Data Analysis for Monitoring Human Rights.* Washington, DC: American Association for the Advancement of Science and HURIDOCS.

Stata, 1997. *Stata Reference Manual.* Release 5. College Station, TX: Stata Press.

Tufte, Edward, 1983. *The Visual Display of Quantitative Information.* Cheshire, CT: Graphics Press.

Tukey, John W., 1977. *Exploratory Data Analysis.* Reading: Addison-Wesley Publishing Co.