Chapter 10

The Guatemalan Commission for Historical Clarification: *Generating Analytical Reports*

Eva Scheibreithner

Introduction

The Initial Problems

I started working with the CEH in July 1998, when its investigation had almost come to an end, joining the database team eight months before the final report was released. Thus there were many situations which led to problems that I could have prevented had I been continuously engaged full-time in statistical analysis from the start of the project. In this report I explain the problems faced in producing the needed statistical analyses and how I found solutions. While my solutions were not always the most efficient, I found effective solutions that I could execute under the circumstances.

Before my employment, no one person had worked full time on the statistical analysis. Many different people obtained statistical output, graphs, etc., with the hardware and software that became my workstation. As every person has his/her own way of maintaining logical order, there was no orderly basis from which to work. No one had kept a permanent record, so I didn't know what had been produced, to whom it went, from where it was produced, etc. The source files and Excel worksheets had names based on no logical system that I could ascertain. Every person working with the files used his/her own system of naming and archiving files. The statistical outputs had neither uniform layout nor titles. Every graph looked different, although some of them were the same! I found it possible to identify them only by close examination. No one could find any specific result; hard copies weren't kept and there were no explanations or details describing the process, the variables, or the abbreviations used.

In the beginning I didn't know what to do with the variables. The CEH defined three different types of victims: *individual identified victims* (VICT_IND), *collective victims* (VICT_COL) and *anonymous victims* (VICT_AN). I found Excel files with three columns, corresponding to these three variables. I added another column summing the three to get the total of violations and I got different totals when repeating the process. Some days later, the programmer checked my computer and verified my suspicion that the sums were wrong because my computer had a virus. He asked me about my additional column for sums and then he told me that the variable VICT_COL already included VICT_IND, and the variable VICT_AN was the sum of all three types of victims. I never got the right totals as they had already been included. This is typical of the kind of problems I faced at the start.

The first two weeks I spent looking through the files and worksheets, trying to find a way to organize the structure. Finally I concluded that it was easier and quicker for me not to use the existing structures and to start with a completely new system of archiving, organizing and naming. I archived the old structure and started working with new updated files and structure.

I believe that the integrity, transparency, and safekeeping of the data and results are important. By transparency, I mean that if I was not available and another person was hired to do the statistical analysis, that person would be able to rapidly continue from where I stopped. I designed my system accordingly. I tried to make my system so easy-to-use that even a person who had never before worked with that structure or topic could follow my work in less than two days of study. Although this was not necessary, I know that anyone else who accesses my archives in the future will be able to follow my work without difficulty.

Importing New Data: From Programmer to Statistical Calculation

Specify New Input Files

Input files for statistical calculation were in dBase Data Format, with the extension ".dbf." Accordingly, unless it would create confusion, I use the acronym "DBF" to denote input files obtained for me by programmer queries on the source database.

To specify new DBFs, I needed to know the exact needs of the investigator: for what reason, why, how does he/she want the information? Not unreasonably, the investigators often didn't know exactly how to express themselves in their requests for information and sometimes they asked for impossible or useless information. However, we were almost always able to work it out and once the required information was identified, I passed the description of the needed blocks of data to the programmer, who provided the DBFs.

In the beginning I had some problems specifying the new datasets needed. The programmer wanted to help me and gave me more than I asked for, so that I could use the files again later (his idea). But for me it wasn't very useful. I wanted only the data I requested and had to cut out the additional, unwanted information. In fact, it's easier to work with small files; small files lead to less trouble with Excel, which performs more reliably with smaller blocks of data.

For example, I needed the disappearances for individual victims in Guatemala City between 1980 and 1982. So I asked for the data blocks for violation, region, year, and individual victims. I received, in addition, the gross violation, the collective victims, the anonymous victims and sometimes the subregions, for which I had no use.

Import New DBFs

Using the Excel function Data>ImportData, I imported the DBFs created by the programmer from his computer directly into my Excel workbook. Note that this function (Data>ImportData) appears only after an Open Database Connectivity[1] (ODBC) link to an external data source has been established. In the absence of such a link, this menu item does not appear.

Check New DBFs

The programmer always made the first check of the new DBFs. When the DBFs passed the first check I always made a rough second check after import. This second check consisted of comparing the total violations of the categories of victims (one, two or all three categories) with the up-to-date overview. The up-to-date general overview was made with every new update of the database. It contained the new totals of cases in the database (of all three categories of victims), the totals of violations (for the same categories), the five main violations (selected in the beginning by the CEH commissioners) with their totals for the same three categories. This information was used to inform -- in a short and exact way -- all the personnel working with the CEH who used the database output (commissioners, central team, investigators, database staff, etc.).

Thus, the last up-to-date general overview showed a total violation count for individual victims of 15,233. All the new DBFs that I imported from the programmer had to match with this number. Usually I checked all the three categories of victims (which meant always the individual victims - VICT_IND -, the individual victims plus collective victims - VICT_COL -, and the total of victims - VICT_AN).

Update Data, Import New Data to Existing Excel Files

To update data from the existing Excel files, I used the function Data> UpdateData within the Excel program. As mentioned earlier, this is possible only after an ODBC link to an external data source has been established.

To assure correctness, I updated existing files in a way that was a bit more complicated than necessary. I "artificially" checked every file, every worksheet and manually updated the worksheets, checking before and after the update. I could have done this much faster with a macro. However, I would still have had to check every worksheet of every file if the process succeeded. I

[1] Open Database Connectivity is a Microsoft standard using drivers to access database files in a variety of formats.

would have had to assure that the result was correct if I had made calculations or manually added columns as well. In addition, I would have needed an updated record of all files with all the sheets I wanted to update.

Check New and Old Excel Files After the Import or Update of Data

I used a process to check the files that may appear time-consuming but which I felt was necessary. I created a checking form (Figure 1, below) to standardize the process. I checked all the updated data for the main questions: what, where, when, who, to whom and a special *star-question*, a complex specific question involving the context (an example is given in the appendix).

First I compared the main violations. I required that all the totals agreed with the general overview. Then I looked for more specific details, comparing all the Excel files, one to the other, always using three different attributes.

Figure 1. Checking Form Example

EXCEL FILE	WHAT	WHERE	WHEN	WHO	TO WHOM	STAR QUESTION
Name	compare main violations to general overview, gender, age	using three regions	three different years	Military	URNG	complex question in context

Generally I spent more time checking the information than making graphs or other statistical outputs. Sometimes the process took weeks. Checking and updating about 40 different files, each with at least eight sheets, and verifying with the checking form was time-consuming. But in the end I found mistakes that had been made by people in the database chain (programmers, typists, and analysts), confirming my belief that checking is imperative and cannot be neglected. Don't automatically trust what you see on the screen!

Update Data

Before an update in the database I always discussed with the programmer the DBFs I needed updated. There were DBFs for which we had no further use, and with every step developing new possibilities, we decided not to update old DBFs (those with no further use). When the programmer got my list of still-useful DBFs, he updated them, and, when completed, passed the checked list to me. By the time the CEH report was finished, only about 20% of all the DBFs created in the whole process had been updated.

In November 1998 we updated the database. This was a busy period for statistical analysis. The investigators were finishing their reports, which created a high level of demand for graphs and statistics. I had a list of about 60 files to update. When I gave this list to the programmer asking him to update the DBFs, he was concerned about the amount of work required in view of the total workload on the system. So I took back the list and reviewed it. I eliminated another 25 files and in the end we updated only about 35 DBFs. The programmer's reaction was correct, as it is time-consuming to update files, and such time should not be spent on files that aren't going to be used in the future.

Final Update

The final update was a more extensive process to be done in a limited time. I still had many files to check, which I had not eliminated. After the normal update check I checked all the files for "white cells" (cells without information). I then passed them on to the programmers to have the cases checked individually to see if there was indeed no further information to enter in the cells or if an error had been made. If needed, the correction was made and we then completed the update to final form.

To be prepared for every possible request I still kept many of the files until the final update. If I had eliminated more of them I wouldn't have had to spend so much time checking them. After checking I did a "white-cells-check," looking for cells without any entries for age, gender, violation,

region, etc. Usually the white cells errors were typing errors, but on occasion some of programming was at fault. I still looked for the outliers, e.g., age=260, etc., and found some I had overlooked previously. As we had a special flag for "massacres," I also checked the case number files for "massacre." We flagged a case as a "massacre" if it had at least five executed victims. We could detect some "massacres" which had not been marked as "massacres" before, and others which were flagged incorrectly. I had two files with the case numbers, so I could pass the case numbers I filtered with these checks on to the data processors to check the cases again. These errors were both typing and analysis errors.

Output from the Database: Answering the Requests of the Investigators

Kinds of Analysis

Generally the information provided by the database was descriptive statistical information, easy-to-read graphs and easy-to-understand figures (examples of some graphs are given in the appendix). We primarily did calculations based on violation, but some special calculations were based on victims. In general, there was no analysis based on cases. Usually we only produced statistical output based on the information the CEH collected.

I produced some graphs based on victims for the chapter on indigenous identity, to show the percentage of individual victims identified for their ethnic characteristics. There were two exceptions. I made some graphs based on cases with the key word Massacre, in particular a time line graph for the areas of military operation. In addition, I made some graphs from the information provided by the military (e.g., how many military commissioners they had recruited in the different regions in the years of the armed conflict) that had not been collected by the CEH.

Kinds of Graphs, Figures, Numbers

The investigators used the graphs we produced to find the tendencies and confirm or disconfirm the hypothesis. Thus, graphs generally had only a few details, and because of this they were easy to read and understand.

Once the recommendation group asked me for a special bar chart. They had made a special codification with only a sample of cases (randomly selected) with instructions to create graphs to analyze and show the tendencies. I made a colored bar chart with the standard error lines. The next day another investigator from a completely different section came to our office to say that we were making useless graphs. He had found the colored bar chart and made a copy for himself. However, since the blue color lines changed to black in the copy, he couldn't identify the standard error lines, so I stopped using colored graphs and used different shades of gray for bar charts and different styles of lines for time lines.

Checking Graphs, Figures, Numbers

Every output of the database had to pass a final check on the layout and the information calculated. I used a special unified layout for all outgoing information. It contained the basic information, e.g., the date of the last update and the total number of cases from the last update, so that the investigator always knew from which set of information his output had been made (an example is given in the appendix). Occasionally, I forgot to change information within the macro that did the automatic formatting. So graphs went out with the old information. The investigators, who sometimes did not find my error, would return claiming that the graph looked different from the last one, but that the total number of cases were the same. At first I was confused and uncertain, but then I realized that I had simply forgotten to change the numbers. After this experience I began checking outgoing material even more carefully.

Lessons Learned

Problem	Alternative used	Lesson learned

Copies from graphs originally printed in color, which led to unidentifiable graphs.	Only black and white graphs used.	It is better to prevent the problem than to trust that everybody will know that copies usually are only black and white.
Output has to go out with some basic information for identification and checking purposes.	Layout of the sheet for providing the basic information which every output received.	Carefully check every outgoing graph to see if all the variables of the layout are changed and updated.
There was no check after me.	Had to check even more carefully.	It would be better to have someone else as a security check.

The Statistical Analysis Program

We used Excel for several reasons. One reason is its primary advantage, which is that it is widely used and it was easy to find a person who knew the program. Another advantage is that the interface is easy to understand and one can develop the ability to use it rapidly.

Unfortunately as we found over time, Excel has many disadvantages. Among them are: loss of graph layout (i.e., formatting would change without apparent cause); update data had to be checked carefully; large data manipulation sequences are difficult for Excel to manage; there are frequent crashes of the computer because of working with too much data, and there is no record of what was done to get the result shown in the worksheet.

Administration: Keeping a Record of the Output

Administration of the Excel Files

I made a new branch of the file directory for every new step or development within the programming process. Thus, the Excel files show their relational structure within the file directory tree, so it's easy to identify the steps of programming.

At the completion of the CEH report process, it was instructive to see the file directory tree. It revealed the whole history of the programming process. First we had different files only for massacres or only for no massacres, then we had the combined files where I could search inside the file the different violations; for example, within massacres or outside of massacres, etc.

Administration of the Outgoing Graphs

The outgoing graphs and figures were registered within two different archives. First they all had their registration numbers included in the title, which consisted of a letter and numbers. Then they were registered in the book of registration and then on the visits control sheet that I maintained. I made copies of all the outgoing graphs to record them within two different file records: one by the type of statistics, e.g., all the bar charts, and the other one for the topic or variable, e.g., all the material covering "massacres" or "department." Thus I always had several ways to find graphs and figures.

I spent almost as much time administering the output as checking it. I had a permanent horror of losing some graph while under pressure to produce results and not being able to find the graph again. So I devised a special registry system. Every type of chart got its letter: B=bar chart, L=time line, T=table, etc. The letter was followed by the indicator for the topic or main characteristic of the graph: 1=department, 2=children, 3=massacres, etc. And then I gave each output a serial number. For example, the registration number B2.32 indicated that this was the thirty-second bar chart on children. I manually kept the records in the registration book, where every type+topic combination had its own page: e.g., I had a sheet for B1, a sheet for B2, and a sheet for B3, etc. In this registration book I kept the date information, the title, where the data came from (file and worksheet name), where the output went (name of the investigator), and notes (e.g., if the graph was an updated version from a former graph).

Lessons Learned

In this section, I discuss both the lessons learned and their implementation.

There was considerable similarity between the work for the CIIDH and CEH projects. Accordingly, the analysts for these two projects, Herbert F. Spirer and myself, jointly prepared recommendations for future large-scale human rights data analysis that appear at the end of this paper in Appendix 1, Data Analysis Recommendations.

Problem	Solution	Issues
No permanent person working on statistical analysis, which led to unique outputs, and inconsistent ways of archiving and naming different layouts.	Have the same person work on statistical analysis and output from start to finish.	If not possible to have same analyst(s) throughout project , establish a uniform logical structure at the start.
No records were kept.	Because of the considerable effort that would have been necessary to recover and reorganize the materials stored under the former inconsistent structure, I started with a new recording system.	If it is too time-consuming to restructure the existing material and if you are still at the beginning of the statistical analysis, create a good new system and take the loss of former material.
There was no detailed information about how the data were processed before they were used in the statistical analysis	Immediate detection and correction of mistakes resulting from of misunderstandings concerning what was in the input data.	Start by asking for all the details you need to know for working with the data (former calculations, what the variables mean, how are they calculated, etc.).
Sometimes the DBFs provided by the programmer contained too much information.	Discussed with the programmer until we found a middle way: I received only the blocks I wanted (plus a few more)	Specify exactly the needs for producing the statistical output. This means exactly specifying the variables requested.
Data has to be checked before using for statistical analysis.	I designed a large checking system with a first rough total check on import of the data and another specific widespread check afterwards.	You can never check too much. It's not so important how, but the important thing is that there are checking steps.
Updating only the minimal number of files used to meet the needs of investigators.	There were always too many files to update. This led to a long updating process.	Eliminate the files with no further use, as there will be new ones as the archive always grows.
Data after the final update has to be as completely checked as possible and cleaned.	Extensive checking methods for the final update.	There may always be some mistakes that you will overlook, but it's always worth trying to eliminate all error.
We found mistakes when checking data.	We found mistakes from typists, analysts, and programmers.	Detecting errors is necessary and positive, but it does not mean blaming someone! Errors happen.
The investigators complained that they didn't receive what they had been told they would receive, because the investigator receiving the output didn't really know what statistical output would look like.	Alleviated by having one full-time analyst (me).	The statistical analyst making the output is the one receiving the requests, and should explain at the beginning how the investigators should make their requests and what they can expect to get.

The person receiving the requests from the investigators was not the same person as the one doing the statistical analysis.	Constantly working to stay in touch.	Only statistically skilled persons should receive requests.
The person handing the output over to the investigator was not the same person as the one doing the statistical analysis. Investigators didn't receive explanations on what the output was about, how it was calculated, where the data had come from.	Constantly working to stay in touch.	Set up system so that analyst physically gives analyses to users. Analyst should explain the meaning of outputs. May not be needed with statistically knowledgeable users.
When I started work, other persons without statistical understanding were obtaining statistical outputs. There was no control over the outgoing information. It wasn't statistically checked, so mistakes went out, and incorrect records and different layouts frustrated investigators	I was able to correct this situation, but could not undo the problems of the past.	Only one qualified person produces statistical analysis to maintain control and records. Or, if more than one person produces statistical analysis, one person has to check everything for statistical correctness and maintain the records.
Investigators haven't been educated in reading graphs and understanding statistics, no explanations were provided. Investigators deduced incorrect explanations of the figures in their chapters, even to the point of misunderstanding the meaning of the title of the graph. Also, investigators misinterpreted analytical findings and made hypotheses that did not correctly reflect the analytical findings.	I held a class and tried to inform everyone about statistics.	Periodic workshops for investigators on the use and interpretations of basic statistics, explanation of the basic graphs.
Many people working on the project had problems using statistical reasoning. This is quite common where there has been no training in statistical methods. For example, people can confuse statements such as "20 percent of the women in Rabinal were assaulted" and "20 percent of the women who were assaulted were from Rabinal."	See above.	The path from the producer to the final user of the statistics should be a short as possible to guarantee a correct result in the final version. Each added intermediary is a potential source of error or confusion, especially if they are not fully qualified. Education for everyone is important; statistical reasoning can be unfamiliar to people.
Programming develops within the process. The ability to identify the different steps must be provided.	Identified within the file directory tree, every step another branch.	Better to have everything written down and recorded in a logical structure.
Output must be easily identified.	Manually using a registration number and recording in the registration book.	Registration is necessary, but my way was very "artificial" as it was kept manually in books.

Output must be easily and quickly retrievable.	Double file record archiving system to provide the possibility to look for the output by two different criteria.	Copies of output are very useful for examples, and as proof and replacement, if needed later

Implementation of Lessons Learned

The following table reviews some of the specific actions that I took in order to put learned lessons to work during the project. The positive effects of these actions are also shown in the table.

How it was	Positive effects
I always put updated graphs and output in visible places in the team's offices to keep the team informed. As almost everybody working in the database offices had been living and working in Guatemala before the CEH started working, they were knowledgeable about the history and actual situation.	As the whole database team is involved in the process providing the data for statistics, they know the whole chain and are interested in knowing what's at the end. This led to a better identification with the group, refreshed their energy and strength, and reduced the widely held distance to statistics. They also made their own personal hypotheses and interpretations leading to interesting discussions in the team.
We produced a general updated overview with every database update, where I changed some expressions into more understandable words before handing it over to the commissioners as part of our agreement to keep them informed.	It was necessary to use common interpretations of technical terms to make the overview more understandable and easy to read. Then the commissioners, the central team and all the people working with statistical output received an easy-to-read overview periodically and were pleased that they were included in the process and could understand what they received.
The output had the same layout and basic information.	Led to a professional impression by the investigators and made any one graph more official.
I started with one three hour workshop, inviting all the investigators, commissioners, team leaders, etc. I explained the main graphs used up to this moment; the data processor explained the different variables and terms used and the programmer talked about the lists of cases provided by him.	The small audience that attended that meeting appreciated the effort and reported that they had learned a lot. From this favorable experience came my idea of periodically providing basic workshops in statistical reasoning.
I started keeping records of the visits from the investigators (unfortunately only for two months), noting their concerns and wishes.	It was easier to prevent misunderstandings and to reproduce acceptable materials later for the same person.

Appendix 1

Data Analysis Recommendations

By Eva Scheibreithner and Herbert F. Spirer

Introduction

As part of the process at the Experts' Meeting, we jointly reviewed our experiences and lessons learned and have integrated them into this set of recommendations for data analysts who will be carrying out similar missions in the future.

We make some recommendations that are explicit statements of procedures that we believe should be followed to maintain the integrity of the data while producing analytical results that faithfully report on the findings of the project. Such recommendations are those required for Verification.

We make recommendations that are general and meant as guidance to the analysts. They are for control of datasets, choice of statistical program, chart standards, an output identification system, and education. In these cases, we hope and expect that analysts will recognize the validity and value of our guidance and use it to formulate their own procedures and practices that are consistent with the context in which they are working. Such recommendations are those concerning Graph Standards.

Control of Datasets

As we have discussed, avoidance of error is critical in the analysis stage to maintain the credibility of the final results. We have found that the following requirements are the minimum needed to assure this freedom from major sources of error.

- The statistical analyst must maintain a current data dictionary. This data dictionary must contain as a minimum, the variable (field) name, the meaning of the variable, and a list or verbal description of the values that can appear in the corresponding field for the variable.
- The analyst must also maintain a cross-reference table of files and variable (field) names so that the analyst and others will know which variables appear in which datasets and which datasets contain a given variable.
- To avoid confusion among different versions of a dataset with a given name, the analyst should use a separate directory (folder) for each version, numbered in accordance with the sequence of the version. If database personnel produce these datasets through queries and store the datasets in directories, they should organize the datasets in this manner.

Choice of Statistical Program[2]

We used Excel in performing our analyses and both of us found it to have problems as described in this paper. In addition to our statistical issues with this program, it had the disadvantage of limited graphic output capability. This latter limitation caused significant problems in the production of the reports. Those problems could have been avoided by the use of Encapsulated PostScript files.

Encapsulated PostScript (EPS) is a standard format for importing and exporting graphic files in all environments. The EPS file is included as an illustration in other PostScript language page descriptions and can contain any combination of text, graphics, and images. Unfortunately, not all PostScript-enabled printers are able to print the EPS files, creating a hardware or software issue that must be resolved to facilitate the analyst's work.

In addition, Excel does not produce a log recording the actions taken by the analyst and the use of Visual Basic macros for this purpose is dangerous. Unless the analyst is diligent in keeping records, in the absence of the analyst, other personnel on the project or outside auditors may have

[2] These observations also apply to the statistical work of the TRC.-PB.

no way to recreate the analysis, except to try to repeat the process. Unfortunately, the analyst cannot recover the actions taken to produce a result from that result except by reverse engineering.

We believe that the use of a particular program should not be dictated, and the analyst needs the freedom to choose a program consistent with experience, abilities, and preferences. Balancing of costs and benefits will lead to the best choice of a program. These considerations include the skills and knowledge of the analyst. In a particular context compromise may be necessary.

Accordingly, we suggest the following as desirable goals.

- The graphical and tabular output will be in the form of Encapsulated PostScript files.
- Either the analyst or the program (preferred) will produce a detailed log of the actions taken in manipulating the dataset to produce results.
- The analyst will use standard programs to make it easier for replacement analysts to check the work rather than exotic programs or those not widely known.

Graph Standards

Choosing the appropriate graph to display information is a combination of technology and art, essentially a creative process. To give specific rules is to stifle that creativity and in the long run, will lead to results of limited value. Our approach to the visual display of our analyses conforms to Tufte's standards for Excellence in Graphical Representation, quoted in the paper The International Center for Human Rights Research Investigation, in the section, Graphs: The Visual Display of Information.

In addition to that general guidance, we recommend that:

- The purposes and needs of the data analysis be met in large part by strategic use of the following types of graph: univariate time series plot (time line), overlaid time series plot, vertical bar chart, horizontal bar chart, stacked bar chart, and histogram.
- The analysts avoid pie charts, which can be difficult to interpret and are often misleading.
- The analysts strive to avoid clutter, which means, among other things: use ticks, but don't use gridlines, don't set charts in visible frames, and don't use markers unless there is a clear need.
- Any tables be spare, and without clutter. There are a number of examples of such table layouts in the CIIDH report (Ball, Kobrak, and Spirer, 1999, pp. 70, 119, 122-3, 128-130).

Verification

The need for verification derives from both the human and machine elements at work in the process of statistical analysis. Among the sources of error are:

- programmer errors in preparing the datasets
- analyst errors in doing the analysis
- program faults inherent in the current version of the analysis software
- consequences of computer crashes
- hardware limitations, inherent in the hardware and possibly unknown to the analyst
- key-entry errors, which can occur at any stage from the initial to the final output

The ideal situation is when none of these errors occur. Analysts, programmers, and others can with experience and motivation, reduce the number of errors generated, but they can never eliminate them. No software is ever completely bug-free, and hardware is prone to both inherent flaws and degradation. Thus, to have credible analytical results, we need a verification process for detecting errors. To this end, we recommend the following to statistical analysts:

- Have programmers producing working datasets supply totals and extremes for all numerical dataset variables as a part of each version.
- Use these totals and extremes as a check on the changes from the prior version of the dataset.
- Check the dataset as received from the programmer.
- Base checks on Table 5, following. The analyst should maintain the summary described in this table. If the analyst uses a program generating a log and allowing the use of stored programs, such as Stata or SPSS, a summary will be automatically retained.

Table 5. Summary of recommended checking methods.

File name	Check	Units of analysis	
	Totals		
	Extremes		
	*-questions		
	Key tabulations of categorical variables		

Note: A *-*question* is some question about the data that will provide a check of context, such as "what proportion of women were disappeared in the month of…?"

- Check the dataset as received.
- Check the dataset at every critical transition. When in doubt, check the dataset at every change. Checking means comparing totals, extremes and *-questions for before and after values.
- Be skeptical, vigilant, and scrutinize constantly.

Output Identification System

To track graphs through the information management process, each graph or tabular output must have an informative, unique identification. With such an identification, which we call the *Graph Identification Number* (GIN), unambiguous reference to outputs can be used in communication among project members and any final or intermediate result can be tracked to its source. In this section, we propose a format for the GIN, a system for tracking the production and transit of outputs, a simple rationale for archiving outputs, and a method for tracking the subsequent changes in a given dataset.

- The GIN is structured as follows: TT-FF…FF-VV-DATE, where TT is the mnemonic output type designator (TA for table, BV for vertical bar chart, TS for time series plot, etc.), FF…FF is a variable length mnemonic for the title, VV is the two-digit version number of this particular output, and DATE is the date produced.
- Outputs are archived by GIN and by category.
- To provide an audit trail and the ability to access users, the analyst should maintain a graph tracking system as shown in Table 6, below:

Table 6. Structure of recommended graph tracking system.

GIN	Create date	Title	Source dataset	Recipient	Notes

- Since modifications of the dataset produced by the analyst to obtain particular outputs produce new internal data sets, successive versions can be tracked by the use of upper case suffixes; and versions from which outputs are produced by lower case prefixes. Hence, BRTANONV14A could represent the data set obtained by filtering out all violations except killings (RTANONV14A) and the specific subset of that data set used to create the second variation of a bar chart as BRTANONV14A.

We note that the use of **some** identification and tracking systems is the key part of our recommendation. Here we recommend a particular system based on our experiences, but other contexts may call for other approaches.

Education

The outputs of statistical analysis are a major part of the end result of a large-scale human rights data project. They represent the physical realization of the logical process of drawing meaningful conclusions about the data. To come to that point, a great deal of interaction among team members is needed. Since most team members will not have had either education or experience in statistical and analytical reasoning, we recommend that education in these topics be included in the project plan and execution.

Education of the type we discuss will have the benefit of more effective, efficient work, and better relationships among project team members. We make the following recommendations, understanding that their implementation will depend on the context and issues of resource limitations.

- Educational objectives are (1) how to interpret graphs and tables, (2) methods of descriptive and exploratory data analysis, (3) the meaning of statistical statements, (4) how to read titles and notes, (4) how to work with absolutes and percentages, and (5) how to work with conditional statements about data.
- Project management should decide on what is best for the given project; whether the educational process should involve all team members, or functional groups, workshops or classes, or continuing and periodic or episodic sessions.
- Because of the serious problems in how "statistics" is taught in schools, many people are averse to the subject and it is usually necessary to mandate attendance.[3] Team members should know that practical data analysis often bears little relationship to the content of conventional first statistics courses.

The amount of time required of team members for the educational process should be strictly limited. Because much of the education in these methods will take place in the workplace, workshop time can be limited to less than eight hours throughout the project.

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.

[3] Teachers of statistics know of these problems. These issues are discussed at almost every meeting of the profession. Unfortunately, there continues to be a difference between what teachers teach in the first course in statistics and what these same people do when working in the field.