WP. 13

ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Dissemination of Statistical Commentary
(Geneva, Switzerland, 4-5 December 2003)

Topic (iv): How to integrate statistics from different sources and subject-matter areas to produce analysis that would be of interest to a wide audience

SHOWING THE BIG PICTURE: EXAMPLES OF BLENDING DATA FROM DIFFERENT SOURCES

Supporting Paper

Submitted by the United States of America[1]

[1] Prepared by Marianne Zawitz, U.S. Bureau of Justice Statistics; Laurie Brown, U.S. Social Security Administration; Colleen Blessing, U.S. Energy Information Administration; and Renee Miller, U.S. Energy Information Administration.

Showing the big picture:
Examples of blending data from different sources

By Marianne Zawitz, U.S. Bureau of Justice Statistics; Laurie Brown, U.S. Social Security Administration; Colleen Blessing, U.S. Energy Information Administration; and Renee Miller, U.S. Energy Information Administration

Statistical agencies often release data from just one collection at a time—narrowly focusing on what they can measure and how they measure it rather than on what people want to know or other contextual factors. But stories that interest people are usually about things that affect them in their daily lives and often require several data sources to provide a complete picture. They answer the journalistic questions of who, what, where, why, and when and are written from the audience's perspective using their vocabulary without methodological or technical terms.

Blending data sources, particularly in a nontechnical presentation, can be difficult. Issues arise concerning data quality, definitions, harmonization of findings, coverage, currency, and differences in methodology. This paper will discuss how several U.S. statistical agencies have handled some of these issues when preparing interesting stories.

Simple Explanations of Complex Methodologies

How can data be presented to help nontechnical users accurately answer questions without overburdening them with metadata and methodology?

Most data collections include extensive metadata such as data dictionaries and a description of the methodology used to collect and analyze the data. Those metadata are often included in single survey publications as an independent, and often quite extensive, section or appendix. However, most nontechnical audiences are not the least bit interested in this information and would never refer to a methodology section. The problem becomes more acute when data from more than one source are used to display the big picture. Such presentations demand that some of the metadata be displayed. The challenge is how to provide users with enough information to allow them to make informed judgments about the data.

One technique is to provide clear, short descriptive pieces of information at the point of presentation—for example, a side bar or a short paragraph with appropriate definitions or explanations. An introductory summary of the collection methods used is also helpful. The following example from *Report to the Nation on Crime and Justice* (BJS, 1988) is the first section from a chapter on criminal offenders:

How do we know who commits crime?

Three major sources provide information about the kinds of persons who commit crimes:

- **Official records** compiled by police, courts, jails, and prisons have the advantage that they offer information on the more serious crimes and criminals. However, these records are limited to only the crimes and criminals that come to the attention of law enforcement officials.

- **Self-report surveys**, in which people are asked whether they had committed crimes, can provide more complete information than official records about crimes and criminal whether or not they are detected or apprehended. But there is the danger that people will exaggerate, conceal, or forget offenses. Many self-report surveys are limited to people who are in correctional custody.

- **Victim surveys**, such as the National Crime Survey obtain information from crime victims including their observations of the age, race, and sex of assailants. Victim surveys give information not only about crime reported to the police but also about unreported crimes. A disadvantage is that crime of stealth (such as burglary, and auto theft) victims seldom ever see who committed the crime. Also, many victims of crime fail to tell interviewers about being victimized by relatives and other nonstrangers.

Source: ***Report to the Nation on Crime and Justice, Second Edition***, Bureau of Justice Statistics, U.S. Department of Justice, 1988

The subsequent text then uses the terms *official records, self-report surveys,* and *victim surveys*. This approach eases the reader into the limitations of the data before they are presented, thus allowing the reader to make judgments as they read. This example also shows how several sources are required to show the big picture, because one source alone may be misleading.

That same report also reduces the technical appearance of the document by providing general source references at the end of each chapter and on all graphics and tables. More detailed documentation is provided in a separate document or technical appendix. Hypertext as used on the Web allows for a similar drilling down to more detail without unnecessarily burdening the main text for the nontechnical reader.

Similarly, the Bureau of the Census also uses boxes to present short, descriptive information at the point of presentation. For example, in the report *Poverty in the United States: 2002*, an accuracy statement appears in a box on the first page, which tells the user that the estimates in this report are based on interviewing a sample of the population. Boxes then appear on subsequent pages to describe new racial groupings and the official poverty measure as they are introduced.

The Social Security Administration's Office of Policy has taken a slightly different approach with its data publication *Income of the Population 55 or Older* (see http://www.socialsecurity.gov/policy/docs/statcomps/income_pop55/), choosing to publish a companion *Income of the Aged Chartbook* (see http://www.socialsecurity.gov/policy/docs/chartbooks/income_aged/). The chartbook summarized some of the original tabular data into a short text description—with a clearly stated main point—and a chart.

Definitions and Measurements

What to do when sources use the same term for different things?

When statistical reports focus on a single data collection, the definitions involved are self-contained. However, with integrated presentations, either in a single publication or across multiple publications housed on a single Web site, conflicts between various definitions become more apparent. For example, until the advent of the Web, efforts within the U.S. Energy Information Administration (EIA) to standardize data definitions had met with mixed success. As electronic dissemination made EIA's products accessible to a much broader audience, concerns about potential customer confusion increased, and in 1998 EIA embarked on an effort to reconcile the multiple data definitions. EIA chartered an official cross-organizational team, the Common Data Definitions Team (CDDT), in February of 1998 that met on a weekly basis from March of that year until the fall of 2001 when the team completed its work. The team found that there were many terms, such as "coal," for which multiple definitions existed because the definition developers thought that different levels of detail were appropriate. These definitions were not contradictory; they evolved because the definition developers were writing for different audiences and thought that different aspects of the definition were appropriate to stress. To harmonize the different definitions the team decided to:

- **Begin definitions with a generic statement.** Whenever possible, the team began definitions with an overall generic statement that was intended to serve as common ground for data users. More specific information was then provided as needed for a better contextual understanding of the terms.

- **Limit supplementary descriptive information.** The team made every effort to limit definitions to the minimum amount of information required to uniquely define the terms. In some cases, additional information was considered helpful. In those cases, a supplemental section of the definition, beginning with the italicized word, *Note*, was provided in order to include this additional information. These notes covered such information as references to specific instructions to survey respondents, caveats about limitations of the data for data users, or specific information needed for a complete understanding of a term.

Furthermore, there were some terms such as "crude oil," where analysts were actually using the same term to represent different concepts. In fact, EIA had seven definitions of crude oil. Some definitions included "lease condensate" (a mixture consisting of pentanes and heavier hydrocarbons, recovered in lease separation facilities) and others did not. This was an important distinction for the upstream analysts who were concerned with estimating oil reserves, but not for the downstream analysts concerned with market issues, such as price. To conform to common usage of the term, the team defined crude oil to include lease condensate and then created and defined another term—crude oil excluding lease condensate (shown in EIA's Reserves publication).

The definitions developed by the team are now in use throughout EIA and are presented in a detailed glossary on the EIA Web site (see

http://www.eia.doe.gov/glossary/glossary_main_page.htm). The standardization helps users to find the correct data and understand exactly what is being measured.

A different approach is taken by the National Center for Health Statistics. In their glossary, they present different definitions for a term based on the survey in which it appears. For example, the following definition is given for the term "hospital discharge" (see http://www.cdc.gov/nchs/datawh/nchsdefs/discharge.htm):

Hospital Discharge

The H**National Health Interview Survey**H defines a hospital discharge as the completion of any continuous period of stay of one night or more in a hospital as an inpatient. According to the H**National Hospital Discharge Survey**H a discharge is a completed inpatient hospitalization. A hospitalization may be completed by death or by releasing the patient to the customary place of residence, a nursing home, another hospital, or other locations.

SOURCE: *Health, United States*

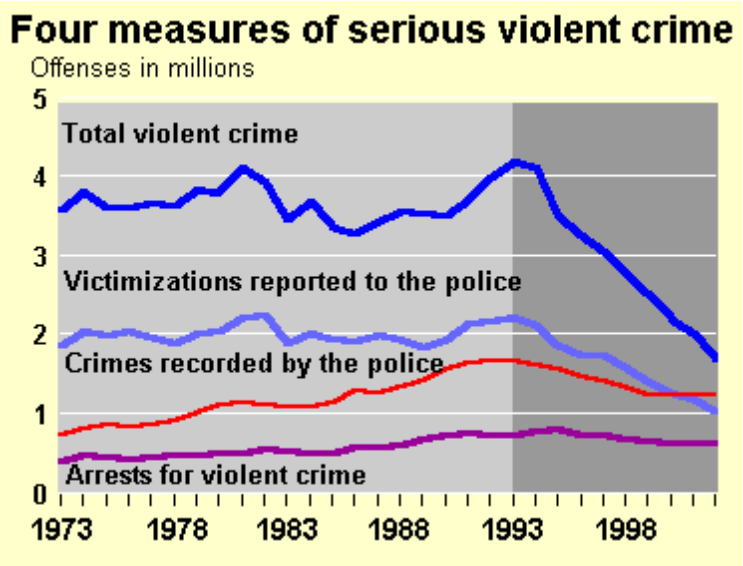
How can different measures of the same item be compared?

Often, more than one data source covers a particular topic such as multiple, overlapping surveys or surveys overlapping with administrative records. In those situations it is important recognize the commonalities and differences among sources, be aware of the relative accuracy of measures across sources, and present the data in such a way that the user does not feel deceived or confused.

For example, in the United States, there are two measures of crime, the Uniform Crime Reports (UCR), collected by the Federal Bureau of Investigation (FBI) from reports by law enforcement agencies, and the National Crime Victimization Survey (NCVS), a nationally representative household survey conducted by BJS. Some of the coverage and definitions differ from one source to the other. For example, the UCR summary system includes eight offenses; for violent crime it includes murder, rape, robbery and aggravated assault. In the NCVS, violent crimes do not include murder, since victims cannot be interviewed, but do include simple assault and sexual assault. In addition, the NCVS does not include any commercial crimes like burglaries of businesses, and it only covers crimes against persons aged 12 or older. The UCR is also dependent on victims reporting to the police and the police reporting to the FBI, whereas the NCVS includes all crimes, whether or not they were reported to the police.

There is great interest in whether or not violent crime is increasing or decreasing. To show the big picture, BJS presents data from both of those sources and adjusts them to make them as consistent as possible (see http://www.ojp.usdoj.gov/bjs/glance/cv2.htm). The following is the resulting chart:

Serious violent crime levels declined since 1993.



The numbers have been adjusted to make them comparable. A factor that accounts for crimes against businesses and against children under age 12 was removed from the UCR. Murder was added to the NCVS, and simple assault and sexual assault were removed from NCVS. In addition to source information, this graphic is accompanied by the following definitions:

The measures are:

Total serious violent crime
The number of homicides recorded by police plus the number of rapes, robberies, and aggravated assaults from the victimization survey whether or not they were reported to the police.

Victimizations reported to the police
The number of homicides recorded by police plus the number of rapes, robberies, and aggravated assaults from the victimization survey that victims said were reported to the police.

Crimes recorded by the police
The number of homicides, forcible rapes, robberies, and aggravated assaults included in the Uniform Crime Reports of the FBI excluding commercial robberies and crimes that involved victims under age 12.

Arrests for violent crimes
The number of persons arrested for homicide, forcible rape, robbery or aggravated assault as reported by law enforcement agencies to the FBI.

As seen in the chart, violent crimes reported to the police as recorded by the FBI and the NCVS show a decline since 1993. And, given the error associated with both series, the current numbers of reported violent crime are indistinguishable.

What should be considered when selecting the most appropriate measure from combined or overlapping sources?

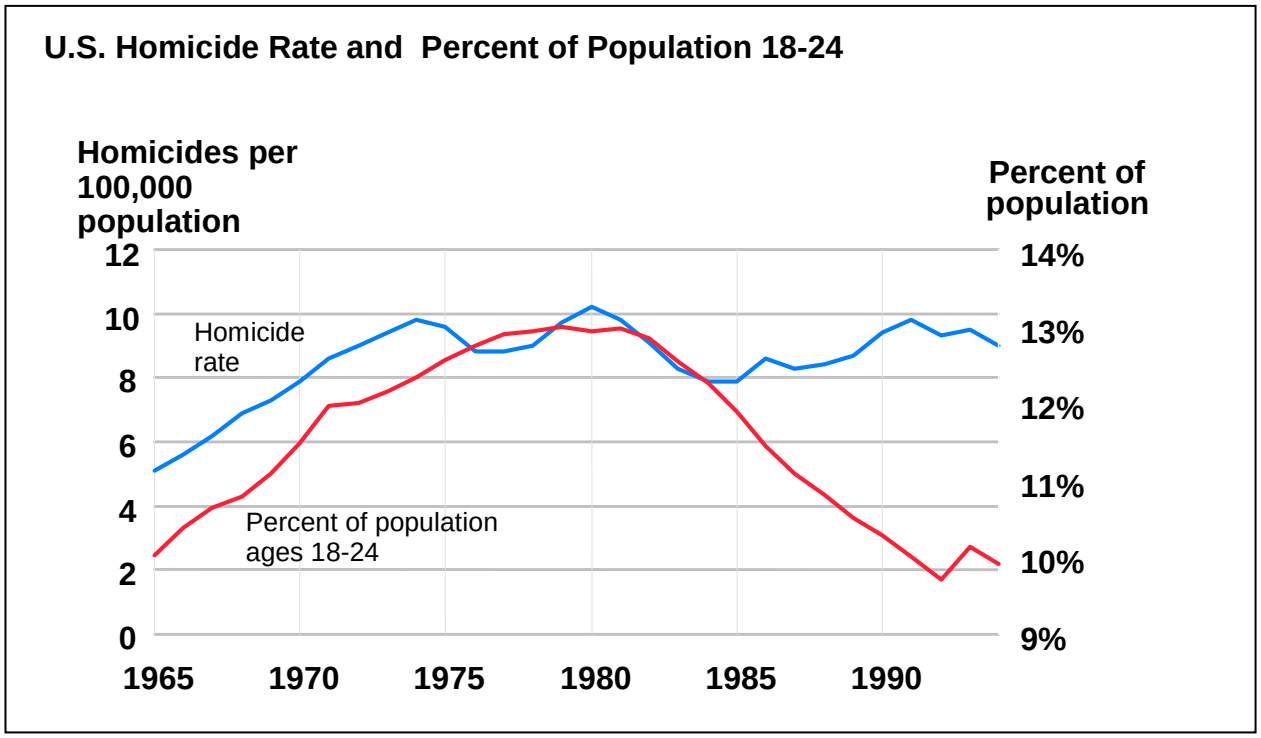
The Social Security Administration's (SSA) Office of Policy also deals with multiples sources—both internally, with multiple administrative sources, and when combining administrative records with survey data. For example, SSA administers benefits under two different programs. The Old-Age, Survivors, and Disability Insurance (OASDI) program, commonly known as Social Security, is best known for paying benefits to retired workers. However, it also has components that cover survivors and the disabled. Coverage is earned through work, and the program is not based on need. It is possible for a single beneficiary to receive multiple benefits—known as dual entitlement—from the OASDI program. For example, a person may receive both a retirement benefit and a survivors benefit. The Supplemental Security Income (SSI) program, on the other hand, pays benefits to the aged, blind, and disabled based on need. Beneficiaries may receive both disability benefits under the OASDI program and payments from the SSI program. Therefore, when reporting data either from the OASDI program alone or in combination with SSI data, it is important to be aware of the difference between a beneficiary and a benefit.

SSA also produces data that are based on a combination of administrative records and survey information. For example, the New Beneficiary Data System (see http://www.ssa.gov/policy/docs/microdata/nbds/index.html) combines administrative data with the 1982 New Beneficiary Survey and the 1991 New Beneficiary Follow-up Survey. While the two surveys cover many items that are not contained in SSA's administrative records, for those items that are, the administrative records are generally more accurate than the same information reported by the respondents. This is an important consideration when selecting measures for further analysis.

Making a Match

How can findings from two or more different sources be presented when the units of analysis are not the same?

Often analysts will try to present two or more variables on the same chart. A common convention is to use two scales when the measures use different units of analysis. The following chart is an example of this type of presentation—presenting the homicide rate and the proportion of the population aged 18–24 on the same graphic:



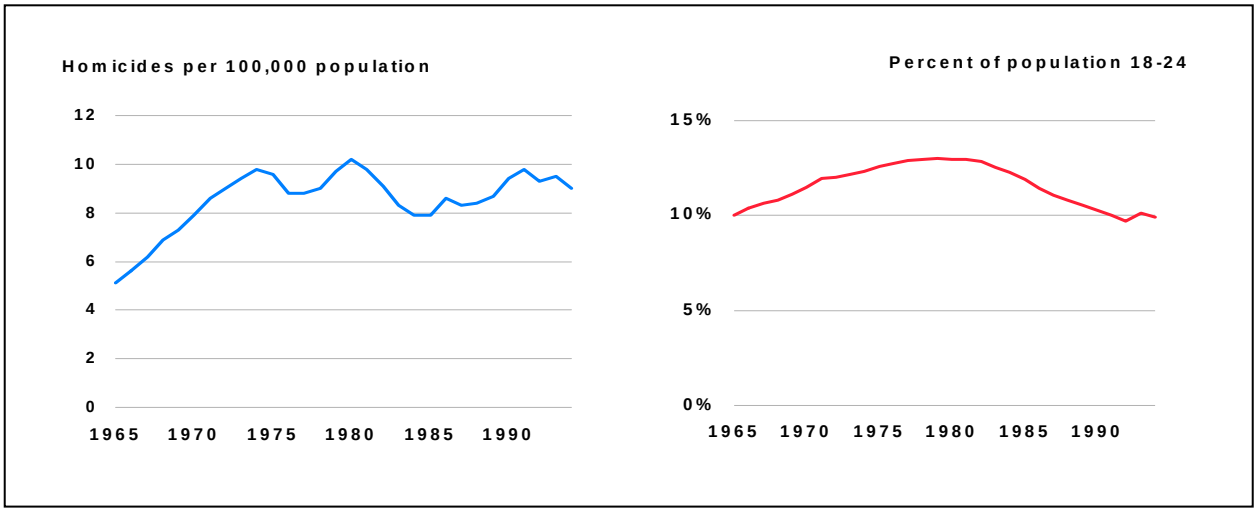
In this example, the author intended to imply causality—that the homicide rate was affected by the size of the population that was in the most crime-prone age groups. On further examination, the reader discovers that the percent of population scale does not go to zero, which looks like the author manipulated the data to get a specific result. In addition, the grid lines are aligned with the homicide rates, making the population line values impossible to discern.

These types of presentations should be avoided. They frequently imply causality when none really exists, and they are very hard for users to understand. Staff members at the Energy Information Administration conducted usability tests on graphics with two scales (Blessing, et al., "Cognitive Testing of Statistical Graphs: Methodology and Results," Federal Committee on Statistical Methodology, November 2003). They tested several different types of graphics with two different axes and found that with the exception of price and volume the two-axis graphics were confusing for users, including many experts. For one of the tests, they concluded the following:

> Complex messages are not easily conveyed through the use of dual axis graphs. In fact, participants can become distrustful of the intent of authors due to the author's ability to manipulate the scales of the axes. Moreover, participants have difficulty in easily perceiving the association between the lines and the axes.

Another way to present data from two different sources using two different units of analysis is to use several small graphics organized to permit comparisons. Tufte refers to this method of presentation as small multiples (Tufte, Edward R., *Visual Display of Quantitative Information*, Cheshire, CN; Graphics Press, 1983, pp. 170-175.) These maintain the integrity of the data without manipulation while permitting comparisons. According to Cleveland's Hierarchy of

Graphical Perception (Cleveland, William S., *The elements of graphing data*, Monterey, CA: Wadsworth Advanced Books and Software, 1985), humans are better at decoding certain elements of graph design than others. Position along common nonaligned scales is second on Cleveland's hierarchy after position along a common scale. The following chart uses those design elements on the homicide and population data previously presented.



It's All a Matter of Time

What to do when the time periods covered do not correspond?

The time periods covered by various surveys frequently differ. Often the most recent period in one survey is several years behind that in another survey. This becomes a major issue when combining data from a variety of surveys into one presentation. The U.S. Bureau of Justice Statistics and the National Center for Education Statistics produce an annual report, *Indicators of School Crime and Safety*, which uses a variety of sources to produce a series of 19 indicators. However, not every indicator gets updated every year. The most recent data available are used for each indicator. The approach used in the report avoids manipulating the data to fit the same time periods and allows the use of the most recent data. A simple presentation in the press release clearly alerts the readers to these time differences:

> "In some cases time periods reflected in the indicators may vary since the report contains the most recent crime and safety data available from a number of separate federally-funded studies. This year's report repeats many indicators from the 2002 report, but also provides updated data on fatal and nonfatal student victimization, nonfatal teacher victimization, the percentage of schools reporting crimes to the police, discipline problems at public schools, and disciplinary actions taken by public school principals."
>
> Source: Bureau of Justice Statistics, Office of Justice Programs, U.S. Department of Justice, Press Release, (http://www.ojp.usdoj.gov/bjs/pub/press/iscs03pr.htm) October 22, 2003.

Conclusion

As these examples have shown, one must consider many different issues when trying to show the big picture. Providing metadata for nontechnical audiences—certainly a challenge in single survey publications—is made even more difficult when sources are being blended. Differences in methodologies, data definitions, time periods, units of measurement, and accuracy all must be considered when presenting blended data. As with so many things we do today, the focus must be on the users. What do users really need to know in order to understand the data? And how can that information best be presented, both in terms of publication design and the use of terms and concepts they will understand in the explanations?