Document Title:	Creating a UCR Utility
Author(s):	Michael D. Maltz ; Harald E. Weiss
Document No.:	215341
Date Received:	August 2006
Award Number:	2004-IJ-CX-0083

CREATING A UCR UTILITY

Final Report to the National Institute of Justice

Michael D. Maltz
and
Harald E. Weiss

Criminal Justice Research Center
and
Department of Sociology
The Ohio State University

June 23, 2006

The overall goal of this project has been to provide the criminal justice community with an easily usable version of the crime data published in the Uniform Crime Reports, collected and published by the Federal Bureau of Investigation (FBI) since 1930. Police departments throughout the United States send monthly counts of crime and arrest data to the FBI, which the FBI then compiles and publishes in its annual report, *Crime in the United States* (CIUS). There are a number of reasons for focusing on monthly UCR data, when most of the attention in the past has been on annual crime data and the extent to which it changes from year to year. First, issues of seasonality can be addressed. While studies of crime seasonality are found in CIUS, they don't take into account local contexts. As an extreme example, larceny in Aspen, Colorado, peaks during the ski season and larceny in Provincetown, Massachusetts, peaks in the summer; averaging these two larceny rates would mask the two effects. Second, not every agency reports its crime every month, meaning that gaps exist in the data. The FBI accounts for these gaps by imputing the missing data, but its imputation procedure doesn't take monthly variation into account. And third, a picture of monthly crime over a long period gives the viewer a clear sense of how the crime problem has changed in that jurisdiction over time; as Tufte (1990, p. 37) states, "*to clarify, add detail*" (emphasis in the original).

This report describes the history of the data collection process, the nature of the problems in dealing with the data, and the steps taken to clean the data and render it useful. It also describes the four products of this project: the cleaned data, a utility for viewing the data, a procedure for adding to the data set as additional years' reports become available, and an analysis of "missingness" in the data.

A Brief History of the UCR

The UCR was developed and implemented in 1929 by the International Association of Chiefs of Police (IACP) and transferred to the FBI in 1930. The primary reason for its establishment was to allay what IACP felt were the public's undue fears of "crime waves," which they felt were manufactured by the press to sell newspapers (Maltz, 1977).

No two states defined all crimes alike, however, making the creation of a uniform record system difficult to achieve. After a number of drafts, the final version of *Uniform Crime Reporting* (IACP, 1929) was published by IACP and distributed to police departments throughout the country, and many police departments began to collect UCR crime data. In 1930 the FBI began to publish the data in quarterly (later annual) editions of *Crime in the United States* (CIUS).[1] The agencies prepared monthly counts of the seven crimes included in the FBI's Index of Crime[2] (also known as Part I offenses) and sent them to the FBI for compilation and publication. Each quarterly report published the crime data for each of the three included months for the reporting jurisdictions, so all four quarterly reports are needed to compile monthly data for the entire year.

[1] The report was originally entitled *Uniform Crime Reports for the United States and Its Possessions*. It was changed in 1958 to *Crime in the United States*, when the reporting became more extensive.

[2] The FBI's Crime Index originally included murder and nonnegligent manslaughter, rape, robbery, aggravated assault, burglary, larceny $50 and over, and auto theft. In 1979 arson was added, but it is not used to any great degree. And all larcenies are now reported, since 1972.

Changes to the UCR. The UCR has undergone three major changes in its lifetime. The first occurred in 1958, when, among other changes, additional data began to be reported;[3] quarterly reports were abandoned in favor of an annual edition of CIUS; and national and state-level estimates of crime were prepared by imputing data from non-reporters (Maltz, 1999).[4] Although the post-1958 CIUS report only provides annual data, the reporting agencies sent (and continue to send) monthly crime tallies to the FBI. With over 18,000 agencies sending data, this is the only reasonable way to present the data in printed form.

Reporting agencies are given an ***ORI*** (an abbreviation of "Originating Agency Identifier") by the FBI, a seven-character string in the form of STcccaa.[5] The first two characters represent the state (AK for Alaska, AL for Alabama, etc. – the same as the postal codes with the exception of Nebraska, for which the FBI uses NB instead of NE). In many States, the next three characters are numbers referring to the FBI sequence number for the county. However, this is not always the case; in eleven States, the county codes do not correspond to the third through fifth digit. See Appendix 1 of Lindgren and Zawitz (2001).

The next two characters of the ORI code correspond to a sequence number for the agency within the county (in some cases these two characters may be letters, like SP to designate the State Police barracks in that county). For example, the Ohio State University Police Department's ORI is OH02527, since it is in Franklin County (FBI sequence number 25), and is agency no. 27 in that county.

ORIs identify agencies that are linked to the FBI through its National Crime Information Center (NCIC), and while all of them are criminal justice agencies, not all of them are police departments. So not all of these agencies have crime data to report.

The second major change to the UCR occurred in the early 1970s, when some states began to require police departments to report their crime data to state agencies, often the state police. This added a step in the error-checking process, augmenting the FBI's own error-checking procedures. In subsequent years most states adopted some form of state-wide data collection, encouraged by funding from the Bureau of Justice Statistics (BJS) and its predecessor agencies. Subsequently a number of states abandoned their state crime data collection programs as federal funding decreased, but most still maintain state-level data collection (Maltz, 1999, p.8).

The third major change, conversion to the National Incident-Based Reporting System (NIBRS), was recommended in 1985 (Poggio *et al.*, 1985) and is still in progress.[6] Rather than an agency reporting the monthly number of crimes (as in the UCR), in NIBRS each individual crime is represented by multiple linked records that detail the characteristics of each victim, each offender, weapons used, property involved, etc. That is, instead of an agency reporting, say, four robberies in January (a single *number*), each of the four robberies would be described in a number of *records*, resulting in a few orders of magnitude more information – and more difficulty in recording.

[3] This included unfounded crimes; Part II offenses such as larceny under $50, lesser offenses such as gambling, liquor law violations, and prostitution; and arrest and agency personnel data.

[4] Gaps in the data occur at the agency level, when a police department misses reporting one or more months, but the FBI imputes only to the state, regional, or national level. The FBI does not impute at the local or county level because it would be misleading (Maltz & Targonski, 2001, 2003).

[5] The FBI's National Crime Information Center (NCIC) assigns 9-character ORI codes to agencies, but only the first seven characters are used for UCR purposes.

[6] We describe it for completeness, since it is not the focus of this project.

According to the Justice Research and Statistics Association (JRSA, 2006), as of August 2005 over one-fifth of the U.S. population is policed by agencies that report using NIBRS (the FBI extracts summary UCR data for these agencies from their NIBRS data). The fact that most agencies now have automated reporting systems to collect such data for their own purposes (e.g., crime analysis) has made this change feasible. In all, over 90 percent of all agencies submit UCR data.

Reporting Gaps. The collection of crime data in the early years of the UCR was far from extensive; many police departments either would not or could not provide reports to the FBI. There was (and is) no penalty for non-reporting, since reporting crime data to the FBI is voluntary.[7] Most agencies, however, do try to comply and provide reports, but may not do so for a number of reasons. In the past few years these reasons have included:

- An agency may have budget problems (and decide it is better to put an officer on the street than dealing with reports).
- Personnel shifts may take a person knowledgeable in the arcana of the UCR away from the reporting system, to be replaced with a person who has yet to learn the intricacies of the UCR.
- Agencies in the midst of computerizing their crime reporting system, especially since NIBRS, may find that the software vendor overpromised on the capability or timeliness of the new system, so no crime data may be forthcoming for a period of time.
- Natural disasters can occur that disrupt the reporting process; e.g., Kentucky lost four years of data due to a flood that wrecked its computer system.
- In addition, small agencies with no crime to report for months on end may decide to opt out of the reporting process.
- Aside from these reasons for non-reporting, there are also cases of mis-reporting; data entry errors may also occur, so individual incorrect entries (e.g., negative numbers, order of magnitude jumps in counts, counts of 999 and 9999 that are used to indicate missing data) have to be excluded.

Policy Implications of Using Incomplete Data. Yet such gaps in reporting did not prevent the UCR from being used in the media and in research studies. For example:

- By November 1930 the FBI was issuing press releases depicting trends in homicide rates (Maltz, 1977, p. 37) – before one year's worth of data had been compiled and while the number and type of reporting jurisdictions still varied considerably from month to month.
- A 1975 study of the effect of executions on the homicide rate used UCR data from the 1930s through the 1970s (Ehrlich, 1975). While the study was hotly contested in terms of the regression model and methods used, no one looked at the data to see to what extent the findings may have been affected by reporting gaps and biases.
- As mentioned earlier, UCR data were used in a study of the effect of laws permitting the carrying of concealed weapons on homicide and other violent crimes (Lott & Mustard, 1997; Lott, 1998). Soon after this study was completed, the potential effect of gaps in the data used by this study was noted (Maltz, 1999, p. 11), and analyzed more closely (Maltz

[7] Some states do require that agencies send crime reports to a central agency. However, there is usually no penalty for not doing so (Maltz, 1999).

& Targonski, 2002), leading to the conclusion that the data could not support the study's findings.

In addition to these policy-oriented studies, numerous research and evaluation studies published in criminology journals have used UCR data under the (often unchecked) assumption that they are sufficiently accurate. Many of these studies have also affected policy, albeit more indirectly.

The FBI has tried to improve the accuracy of the UCR by estimating (imputing) the crime rate to fill in the gaps. However, there are a number of problems with the imputation method they have been using (Maltz, 1999), and new methods are currently being considered; with funding from the Bureau of Justice Statistics, the American Statistical Association has funded our study to develop new imputation methods.

Thus, some of the nation's crime control policies have been based on crime data that are incomplete – and incompletely understood. In the first example cited above, it is apparent that crime trends often were (and still are) based on partial data. In the second and third examples, analyses of these deficient data have been used to convince states to change their policies on execution and on carrying concealed weapons respectively.

However, there has been no concerted effort to compensate for the reporting gaps that have crept into the UCR over the years. To accomplish this, we have compiled all of the extant UCR crime data from 1960 (the earliest year for which computerized data are available) to 2002. We have also developed procedures to clean those data and integrate the data of subsequent years. The next section describes the cleaning process.

OBTAINING AND ORGANIZING THE UCR DATA

Crime in the United States is published annually, as are the UCR data sets on which CIUS is based. Most of the annual UCR data sets can be downloaded at no cost from the National Archive of Criminal Justice Data (NACJD), part of the Inter-University Consortium for Political and Social Research (ICPSR), which is housed at the University of Michigan; see http://www.icpsr.umich.edu/NACJD/ucr.html#desc_al. NACJD has data sets for 1966 through the most recent year. By dealing directly with the Criminal Justice Information Services Division, we were able to extend the data back to 1960, with one notable exception: data for 1962, for state numbers 43-51,[8] "were inadvertently erased during an electronic update of the Uniform Crime Reporting (UCR) Program's Master Files"[9] and are not available.

Availability, however, does not necessarily imply accessibility. The data reside in annual files, so our first task was to combine them into longitudinal files, running from January 1960 to December 2002. The files contain more than just crime data for the seven Index crimes, and we used a great deal of the additional data in our compilation. The data we extracted include, for each agency:

- Monthly crime counts for the seven Index crimes of murder, rape, robbery, aggravated assault, burglary, larceny, and auto theft. We have not included arson since, as the FBI notes (e.g., CIUS, 2002, p. 209), "fewer agencies furnished complete reports for arson than for the other seven offenses making up the Crime Index."
- Monthly crime counts for manslaughter; for forcible and attempted rape (they sum to the Index category); for robbery with a gun, knife, personal weapon,[10] and other weapon entry (they sum to the Index category); assault with a gun, knife, personal weapon, and other weapon entry (they sum to the Index category); for simple assault; for burglary with and without forcible entry (they sum to the Index category); and for theft of autos, trucks, and other vehicles (they sum to the Index category).
- Population for that year. These data are obtained by the FBI from the Census Bureau. What makes this somewhat complicated is that some jurisdictions are in more than one county, so the agency's data includes the population located in (up to) three counties.
- County indicator(s). The FBI-designated number of the county/ies in which the jurisdiction is located.
- Population group. The FBI's classification of population groups is given in Table 1.
- Months for which that agency reported that it submitted crime reports. If an agency sends in a report to the FBI, the date of receipt is noted in the data file as "Date updated." If no such date is given for that month, our default assumption is that no data were sent for that month, which is usually (but not consistently) true.
- Covering agency ORI. If an agency did not submit its reports directly (perhaps because of its small size), it records the ORI of the agency through which it submitted its crime

[8] They are, respectively, TX, UT, VT, VA, WA, WV, WI, WY, AK, HI – DC is included as "state" no. 8.

[9] Email to M. Maltz from the FBI Criminal Justice Information Services Communications Unit (cjis_comm@leo.gov), June 2, 2005.

[10] A "personal weapon" is cited when an offender's hands, arms, fists, feet, teeth, etc. are used in the commission of the crime.

reports. [The FBI indicates this relationship by stating that the instant agency is "covered by" the reporting agency.]

- Identification number of the Standard Metropolitan Statistical Area (SMSA) in which the agency is located. This is included in case users want to obtain SMSA-level statistics.

Table 1. FBI Classification of Population Groups

Population Group	Political Label	Population Range
I	City	250,000 and over
II	City	100,000 to 249,999
III	City	50,000 to 99,999
IV	City	25,000 to 49,999
V	City	10,000 to 24,999
VI	City[a]	Less than 10,000
VIII (Rural County)	County[b]	. . .
IX (Suburban County)	County[b]	. . .

Note: Group VII, missing from this table, consists of cities with populations under 2,500 and universities and colleges to which no population is attributed. For compilation of CIUS, Group VII is included in Group VI.

[a] Includes universities and colleges to which no population is attributed.

[b] Includes State police to which no population is attributed.

Thus, each agency would be represented by 43 x 12 months of data for 25 categories of crime, and 43 years of population, population group, crime reporting behavior, covering agency, county, and SMSA identifier data, or close to 15,000 data elements for each agency. Rather than create a single file with over 18,000 cases (one for each reporting agency), each with these data elements, we decided to create state-level data sets – and even these average over 30 Mbytes in size.

It would be possible to reduce the size of the files by noting that many of the Index categories can be obtained by summing sub-categories, and that the Index itself is the sum of the individual Index categories. Doing this, however, would slow down the data plotting considerably for little gain –the cost of data storage is so low and the speed of accessing data from storage devices is so high that we didn't consider it to be a worthwhile tradeoff.

In the 43 years covered, a number of changes were made to the formats and identification codes used in the files. The first task we faced, therefore, was to extract similar data from dissimilar data files. We accomplished this by generating specialized SPSS syntax files for each year's data file. Since the syntax files follow a specific pattern for the entire year – that is, the data structure and sequence for February is the same as for January – all we needed to do was to determine the sequence for one month and replicate it for subsequent months.[11] This was accomplished by programming the sequence in an Excel spreadsheet and adjusting the column numbers (e.g., where the sequence started and the number of columns in each sequence) so that all of the data would be entered in the same order for each year.

We then extracted the annual data for each state into a separate Excel file, giving us a separate state file for each of the years – the 44th year was withheld to permit us to develop

[11] Those who would try to replicate this task should be aware that the sequences shift from year to year; however, they follow similar patterns.

programs to append the next year to the extant state files, since we fully expect this effort to continue beyond the 2002 data year. The last step in this process was to write a macro in Excel to combine all the year-files for each state into a single multiyear state file.

CLEANING THE DATA

After inspecting the data we noted that certain aspects of the data cleaning process could be done automatically, depending on the type of "missingness" we encountered. Data could be missing because:

- the agency did not exist at the time. Over the past four decades new towns (and police departments) arose, towns became cities, and rural areas assumed their own policing responsibilities.
- the agency did not maintain a crime reporting program for a period of time, but rather contracted that function out to another agency. A "covered" agency is one whose data are reported, but reported by the "covering" agency, often a county-wide agency.
- the agency did not have an entry for "Date updated" in the FBI raw data file. In such cases data may be missing because the agency neglected to send in a monthly report for one reason or another, or the lack of the "Date updated" entry may be in error, since in a number of cases we found data to be present when none was expected (and vice versa).
- the agency decided, as a matter of policy, not to submit data for a period of time. This may be due to the agency's not having any data to report for extended periods of time, or because it did not want to assume the bureaucratic burden entailed by reporting the data; there is no penalty for not reporting the data, since the UCR system is voluntary.
- the person in the agency responsible for collecting and processing the data for the UCR retired or was transferred or promoted. The new person responsible for the UCR may take a few months to learn the ropes, during which time data may not be transmitted to the FBI.
- the agency did not send in data every month, but instead sent it in bimonthly, quarterly, semiannually, or annually. In this case the data seem to be, but aren't, missing, because they are aggregated into another month's data.[12]

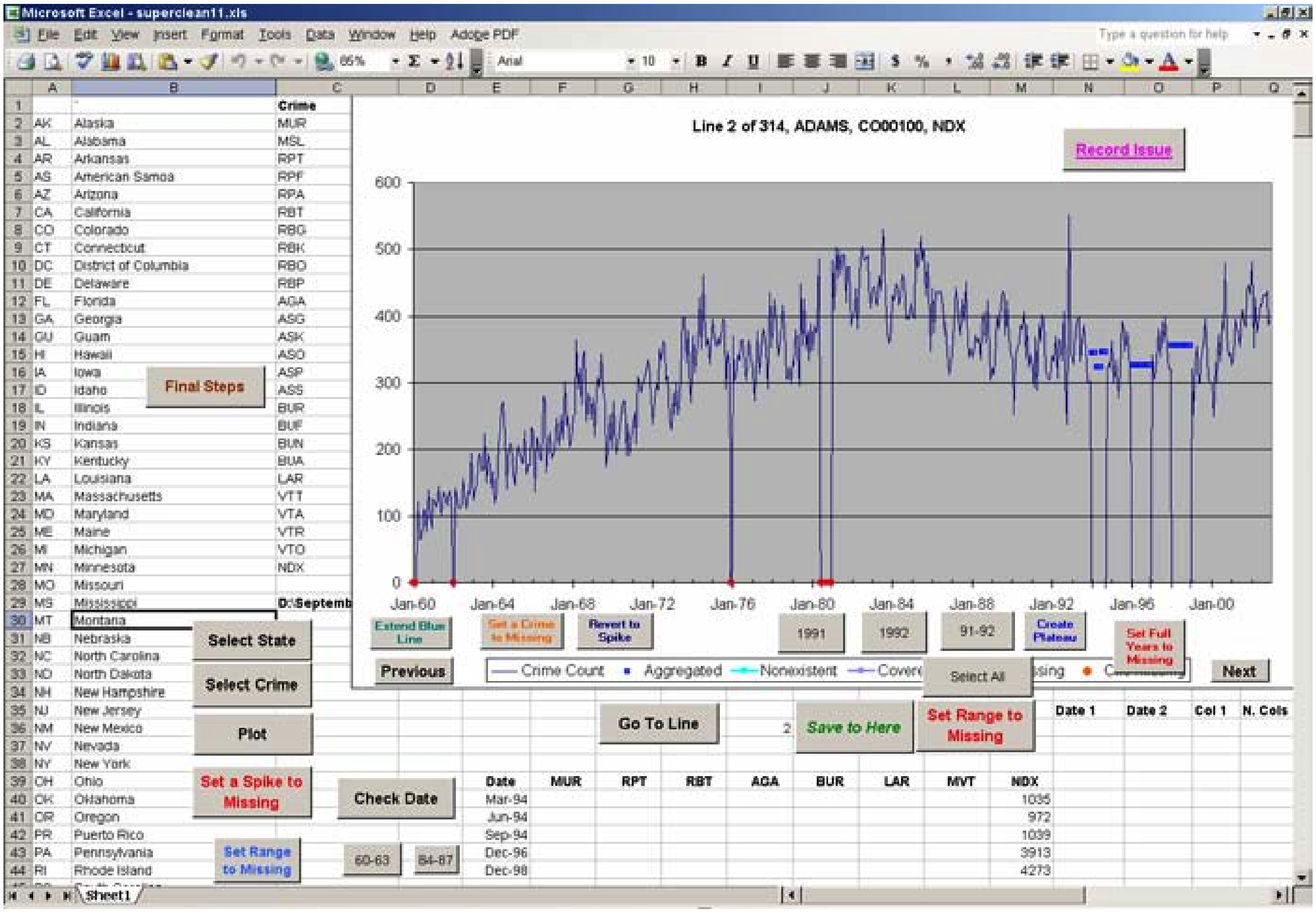
Many of these apparent or true anomalies can be dealt with by using automated search techniques to find and account for them, while others need to be dealt with individually. For example, if an agency has no data for January and February, but does have data for March, in some cases it is due to quarterly reporting, in others due to no crime in the first two months; a visual inspection of the data is therefore necessary to distinguish them.

Attempting to distinguish such cases automatically, by writing an algorithm to recognize when data are aggregated and when not is fraught with problems. We know, because we tried. One has to take into consideration not only the average and standard deviation of the data, but how volatile the agency's data seem to be (is there a strong component of seasonality), its behavior over the long haul vs. the more recent years, and the behavior of other agencies in the

[12] As mentioned, in 1958 *Crime in the United States* began to publish annual reports, and many agencies thus considered the UCR program to be an *annual* reporting program. Consequently, an unknown number of states began to submit reports annually, or (if they submitted data monthly) to correct earlier monthly reports by submitting negative crime counts in subsequent months, to ensure that their end-of-year statistics were correct (John Jarvis, personal communication). This has resulted in negative crime counts. We analyzed the negative count data and decided that we would accept as data counts of -1 to -3, but not -4 or under, which we considered to be coding errors – we did not regard as likely that, for example, -25 crimes in a month was a correction, especially if the average monthly crime count for that agency was of the same order of magnitude.

same county. An algorithm constructed along these lines would still perform less well than a trained observer's eye. This is not to say that a Bayesian decision algorithm can't be trained to replicate the selections of a trained analyst; however, since this process need only be done once, it is much easier to perform the task manually than to write such an algorithm and train it.[13]

One of the guiding principles of our data cleaning process has been to record every change we make to the data so that, if ever a researcher wants to look at the original data set, all that need be done is to check the record of changes and restore the original data.[14] This meant that even the "manual" changes needed to be documented automatically. After a great deal of trial and error (we had to restart the cleaning process four times), we developed a set of macros that would allow an analyst to depict the time series of each crime category to determine whether changes needed to be made. The analyst could then make changes by specifying dates and pressing a button. Figure 1 is a screen shot of a typical time series.



	A	B	C
1			Crime
2	AK	Alaska	MUR
3	AL	Alabama	MSL
4	AR	Arkansas	RPT
5	AS	American Samoa	RPF
6	AZ	Arizona	RPA
7	CA	California	RBT
8	CO	Colorado	RBG
9	CT	Connecticut	RBK
10	DC	District of Columbia	RBO
11	DE	Delaware	RBP
12	FL	Florida	AGA
13	GA	Georgia	ASG
14	GU	Guam	ASK
15	HI	Hawaii	ASO
16	IA	Iowa	ASP
17	ID	Idaho	ASS
18	IL	Illinois	BUR
19	IN	Indiana	BUF
20	KS	Kansas	BUN
21	KY	Kentucky	BUA
22	LA	Louisiana	LAR
23	MA	Massachusetts	VTT
24	MD	Maryland	VTA
25	ME	Maine	VTR
26	MI	Michigan	VTO
27	MN	Minnesota	NDX
28	MO	Missouri	
29	MS	Mississippi	D:\Septemb
30	MT	Montana	
31	NB	Nebraska	
32	NC	North Carolina	
33	ND	North Dakota	
34	NH	New Hampshire	
35	NJ	New Jersey	
36	NM	New Mexico	
37	NV	Nevada	
38	NY	New York	
39	OH	Ohio	
40	OK	Oklahoma	
41	OR	Oregon	
42	PR	Puerto Rico	
43	PA	Pennsylvania	
44	RI	Rhode Island	

2

Date 1	Date 2	Col 1	N. Cells

Date	MUR	RPT	RBT	AGA	BUR	LAR	MVT	NDX
Mar-94								1035
Jun-94								972
Sep-94								1039
Dec-96								3913
Dec-98								4273

Each button activates a different macro. They include:

- "Select State." The analyst selects the cell containing the state from Column A or B and presses this button. This macro closes other state files, opens the one that is selected,

[13] Developing such an algorithm might be worth doing, however, if only to check on our decisions; one of the drawbacks of manual review is that we may overlook situations that an algorithm would not.
[14] This policy was suggested by Marianne Zawitz of BJS, a member of the project's advisory committee.

colors the state box yellow, and depicts the first agency (which are ordered by ORI code).

- "Select Crime." The analyst selects the cell containing the crime from column C and presses this button. The macro then draws the selected time series. The default is NDX (Index crime).

  In addition to displaying the time series, the macro lists the end date of any plateau in the trajectory (designated by a thick blue line) and the total number of Index crimes that plateau represents (in column M, "NDX"). For example, if the Index crime data are reported once a year in December, with 120 crimes for that year, the blue line would show ten crimes for each month of that year. The reason for representing the data in this way (instead of showing the spike of 120 in December) is so that the figure's vertical scale gives the analyst a better visual indication of the trajectory's pattern.

- "Next" and "Previous" allows the analyst to look at the next or last ORI's trajectory.
- "Set a Spike to Missing." In rare instances all of the Index crimes for a month are considerably higher than earlier or subsequent months. It may be that this particular month contains aggregated data for a few months, but those months had already been reported. In such a case we set the entire month's data to "Missing." We can investigate whether a single crime or multiple crimes are overly high by enabling the macro "Check Date" (below).
- "Check Date." To see all of the Index crimes for a sequence of months (e.g., to see if only one crime or a number of crimes are unusually high or low) the analyst can list them in column E under "Dates," highlight them and press this button. This will parse the Index crimes into their individual components (and the total) in columns F-M.
- "Set a Crime to Missing." Occasionally we see a large spike in the Index crime count for a single month, and upon tabulating the individual Index crimes for that month (see the description of the macro "Check Date" above) we see a count of 999 or 9999 for a single crime. Since those numbers are often used to indicate missing values, we replace these spikes with missing values, by selecting the cell indicating crime and date.
- "Revert to Spike." The initial cleaning process searched for potential quarterly, semiannual and annual reporting by checking for monthly reporting patterns of 00X, 00000Y, and 00000000000Z, where X, Y, and Z are nonzero and

  - X occurs in March, June, September or December;
  - Y occurs in June or December;
  - Z occurs in December; and
  - the values X, Y, and Z are significantly above the mean Index crime for that agency.

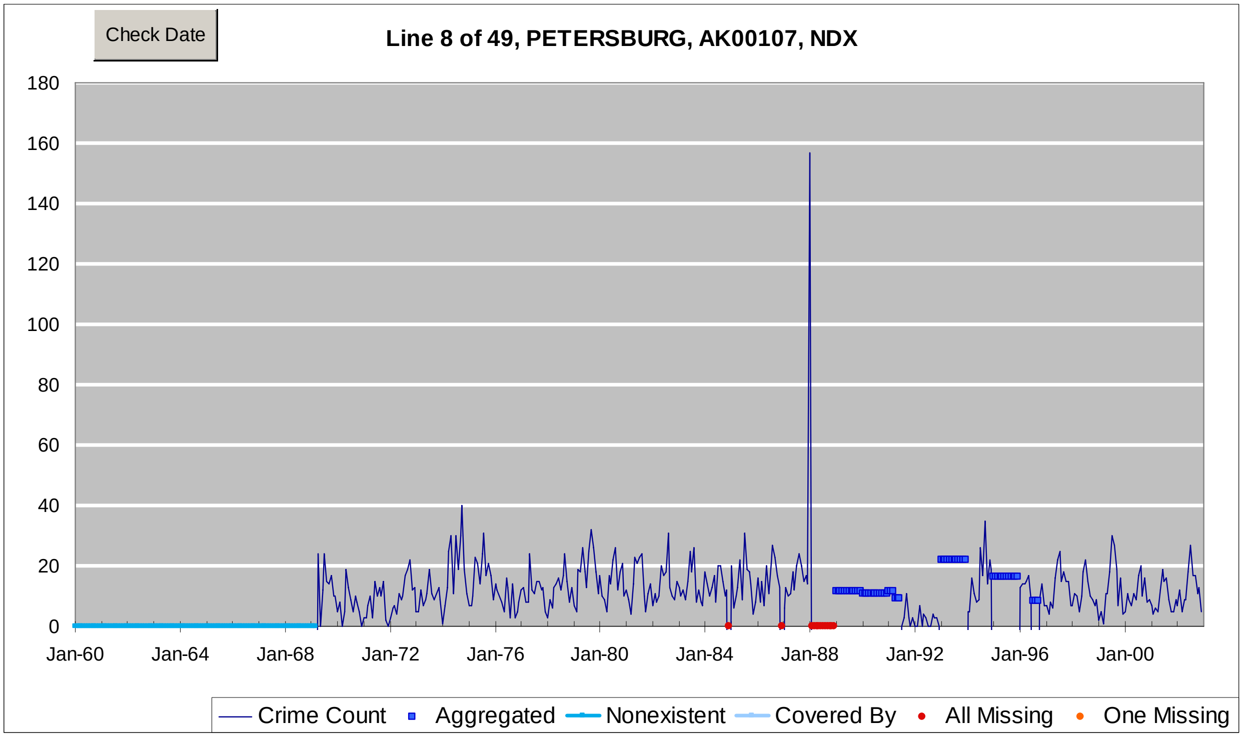
  Since the mean is not always a good indicator of whether the data have actually been aggregated, occasionally the analyst must remove the algorithm-designated plateau and revert it to a string of zeros and a spike.

- "Set Full Years to Missing." This macro requires a start date (month/year) and end date. The analyst highlights these dates and the algorithm sets all of the entries between the two dates to the value that indicates a missing value (-99, cell colored red).
- "Record Issue." When this button is pressed, the analyst is prompted to enter a date in an entry box; this is then followed by a prompt to enter the nature of the issue in another entry box. Very often this box describes an unusual crime count. For example, in one case the sequence of larceny counts was 1814, 1404, 15145, 3033, 2795. We are certain that the middle count was mistyped by the data entry clerk hitting the 4 and 5 keys together, but we are unqualified to make this decision. So we left the anomaly in the data and recorded an issue, which is entered as a comment to the entry for this agency in the list of revisions.
- "Go to Line." An entry box opens and the line number is entered. The current line number is shown in the cell to the right of this button (as well as in the title line of the graph).
- "Save to Here." Pressing this button will save the file and all of the revisions made to date.
- "Set Range to Missing." The analyst enters two dates, and the macro sets the cells between the two dates to the missing value (-99) and changes its color to red.
- "Extend Blue Line." Blue lines represent missing data that do not require imputation, because either agency had not previously been assigned an ORI number (the jurisdiction may not have had a police department until then), or it may have reported its crime data through another agency ("covered-by"). However, these missingnesses are reported annually, but may have ended in the middle of a year. That is, in 1967 an agency may have reported its data through another agency, but stopped doing so in June 1968. In 1967, then, the agency was recorded as having been covered by another agency and in 1968 it was recorded as having reported its own data – *even though the agency was covered by another agency until May 1968*. If the pattern of reporting appears to reflect this, then we extended the covered-by status through May 1968. We treated new agency ORIs the same way.
- "Final Steps." As mentioned earlier, some agencies spread beyond a single county, and to facilitate future county-level analysis we create new records (with the same ORI) in the other counties. The need then exists to allocate the crime counts to the various counties comprising the ORI. We do this by prorating the crime by population in each county, which is included in the original FBI data set. The final step accomplished by the macro called by this button is to copy the same missing codes, ones entered manually during the cleaning phase, to the new records we have created.

As can be imagined, developing these macros and the actual cleaning of the data using them have required a great deal of time, effort, and concentration. Not only that, but we found that we had to start over a number of times, for a number of reasons: problems in acquisition of the original data; macros that had to be revised; and other issues that did not surface until we had already cleaned (improperly) a few states' data. Despite these problems, we felt it best to begin anew each time, to make sure that this time-consuming process would not have to be revisited.

We should also note that the data cleaning is not complete. Some "fine tuning" still needs to be done. There are spikes that need to be explained; as an example, the figure below shows a spike in January and no data for the rest of the year, making us consider the possibility that the data represents the entire (subsequent!) year. In virtually every other case the data for a whole year is reported in December of that year.

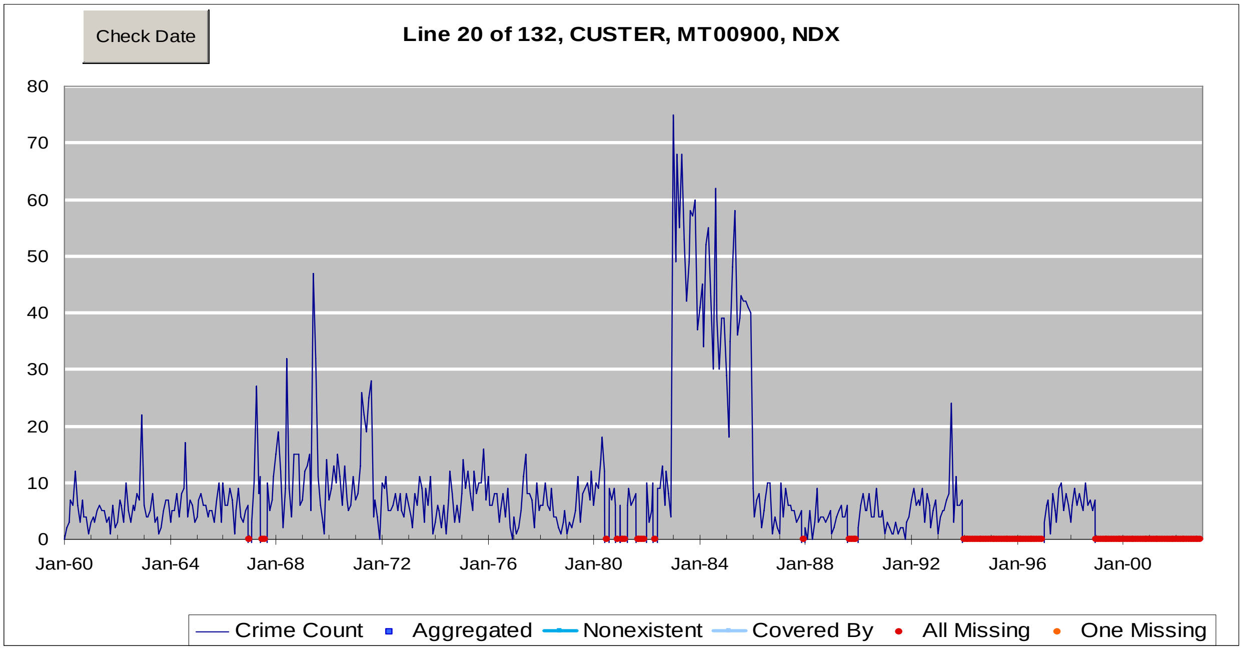
[Going back to the original data, we ascertained that this was due to the way and time that the data for this year were entered. The data for months 2-12 were recorded as having been updated on March 29, 1989, while the data for January were recorded as having been updated on June 23, 1989. We can therefore assume that the January entry reflected the crime data for the entire year.[15]]

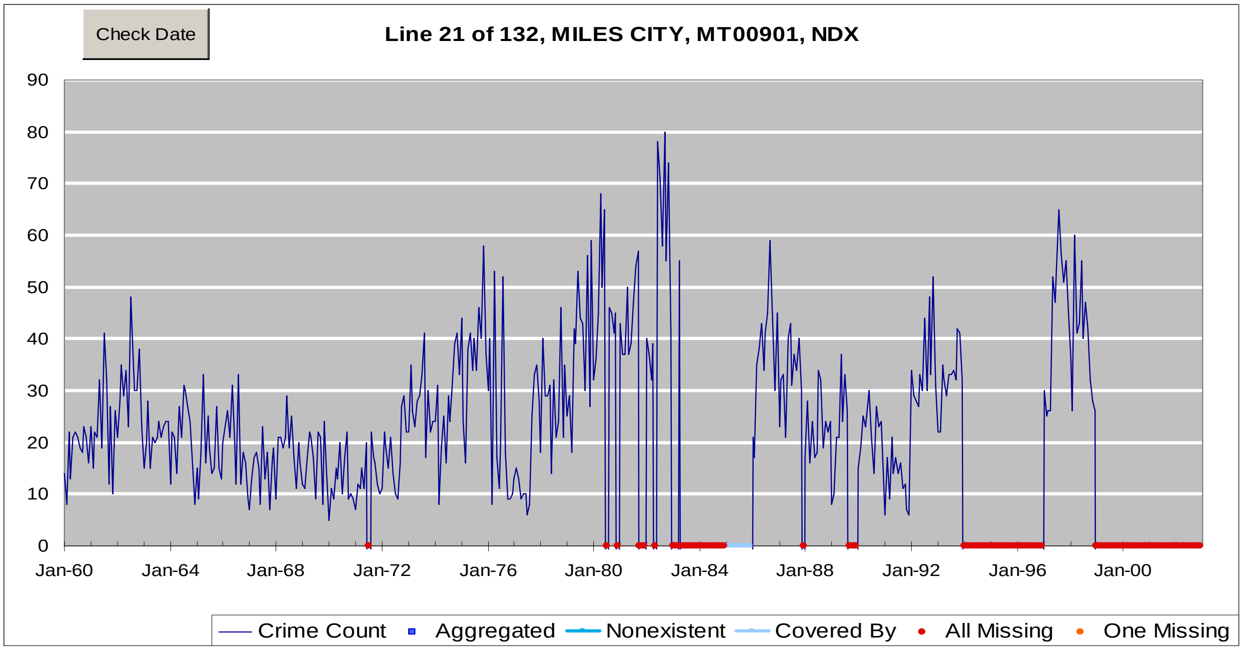


Another cleaning problem relates to consideration of multiple agencies. An example is shown in the next two figures.

As can be seen, it is apparent that the Miles City Police Department reported through (was "covered by") the Custer Sheriff's Police from 1983-85 (except, perhaps for April 1983); however, it was only reported as such for 1985 – see also Maltz & Targonski (2002, 2003). Looking only at one of these two agencies will provide a misleading picture concerning the amount of crime in either jurisdiction. One of the ideas we have come up with to (partially) deal with this is to change the color of the artificially elevated crime count in Custer and similar cases in other states and counties. However, we have not been able to include this in the current version of the data plotting utility.

[15] This points out another data-related issue (Maltz, 1999, p. 21): the data published by the FBI in *Crime in the United States* and the data we have been working with are not exactly the same, since the electronic version of the data has a later publication deadline than the paper version.





We do feel, however, that the data set is sufficiently clean that others can use it – as long as they are made aware of the problems that remain in the data set.

PROMISED PRODUCTS

During the course of the grant we met frequently with our advisory committee.[16] Based on their advice we determined that we would develop four deliverable items to NIJ as our work product: cleaned UCR data for the 50 states and the District of Columbia, in Excel workbooks, and a list of revisions we made to the data so others can determine the steps taken to clean the original data; a UCR charting utility to permit users to plot, over time, the monthly count of crimes for any ORI in the US that has reported crime data; a set of instructions and/or computer procedures to permit the addition of subsequent years of data to the utility; and an analysis of the nature and extent of "missingness" in the data. These four products are included in the package we have provided NIJ. Some words of explanation are in order.

1. **A Clean UCR Data Set**. We decided early in the project that the data would be provided in Excel workbooks. Putting the data into a database program like Microsoft Access (instead of Excel, a spreadsheet program) would reduce the size of the files considerably. There are, however, some advantages to using Excel.

- First, most desktop computers are equipped with this (or compatible) software.
- Second, it can be opened on either PCs or Macs.
- Third, more people know how to work with spreadsheet programs than database software like Access.
- Fourth, Excel data cells can be given different colors (which we use to signify different data "missingness" characteristics), while database programs (such as Microsoft Access) do not include this capability. Coloring the data cells permits the user to get a gestalt impression of the extent and type of missingness in the data.

A disadvantage to using Excel is that it supports only 256 columns; however, this is a limitation of many database programs as well, including Access. This has meant that, since states have many more than 256 agencies (and one state has more than 256 counties), we could not use a single worksheet for each crime category. Instead, we decided to use three worksheets for each crime category: "MUR1" includes the murder data from January 1960 to December 1979; "MUR2" from 1/80-12/99; and "MUR3" from 1/2000 through the latest year. An advantage to this is that only the "XXX3" sheets need to be revised as data for new years are added to the data set.

We used the semiautomated procedure described earlier ("CLEANING THE DATA") to locate and clean the anomalous data points. Whenever a datum was changed, a record of the change was entered on the last page of the state's workbook (the "Revisions" page), so that an analyst can see what was changed and could recreate the original datum by reversing the change.

2. **A UCR Charting Utility**. The charting utility is an Excel workbook, UCRPlot.xls. I***t will not run unless macros are enabled, and it must be in the same directory as the state data***.

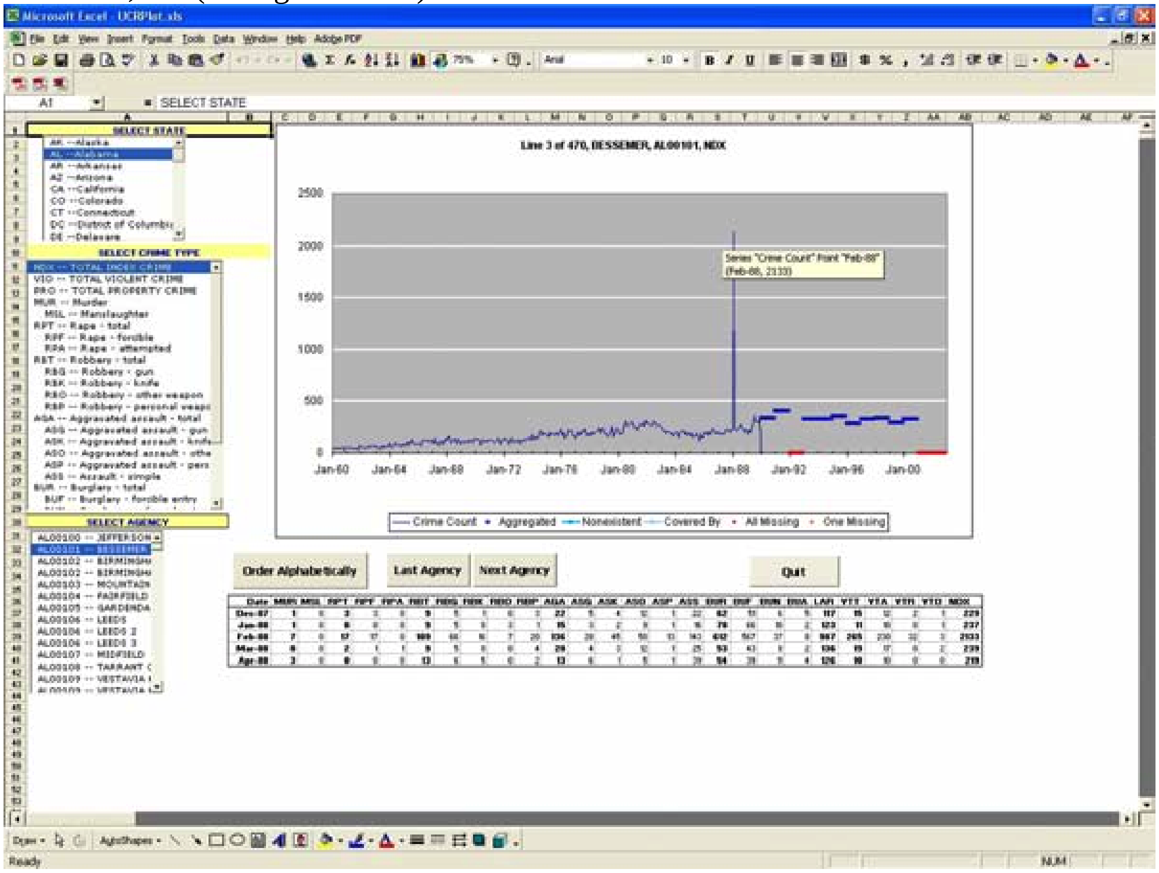
[16] The grant advisory committee consisted of Marc Buslik, Chicago Police Department; John Jarvis, FBI; Christopher Maxwell, Michigan State University and National Archive of Criminal Justice Data; and Marianne Zawitz, Bureau of Justice Statistics. Also participating in the meetings were Patrick Clark and Gerald Gaes, NIJ, and Ramona Rantalla, BJS.

The user initially makes three choices, the state, the crime type, and the agency. Selection of a state brings up the list of agencies, arranged in order of their ORI identifier.[17] The default crime type is Index crime and the default agency is the first listed agency. The agency listing can be changed from ordering by agency ID (default), which is essentially a geographical ordering, to alphabetical ordering, using the button immediately to the right of the agency list.

The next two buttons, "Last Agency" and "Next Agency," permit the user to move through the list of agencies. The plot then shows the selected agency's trajectory for the same crime type.

The next button, "Quit," is self-explanatory; it closes whichever state file is open, as well as the charting program.

To get an indication of the crime count around a given point, the user can position the cursor directly on that point and left-click the mouse. The complete crime data for that date and two months before and after that date will now be displayed below the chart, with the Index crime counts in bold font. To show how this can be used, the user can select the agency Bessemer, AL (see figure below).



Date	MUR	MSL	RPT	RPF	RPA	RBT	RBG	RBK	RBO	RBP	AGA	ASG	ASK	ASO	ASP	ASS	BUR	BUF	BUN	BUA	LAR	VTT	VTA	VTR	VTO	NDX
Dec-87	1	0	3	3	0	9	5	0	0	3	22	5	4	12	1	22	82	51	6	5	117	15	12	2	1	229
Jan-88	1	0	6	6	0	9	5	0	3	1	15	3	2	9	1	16	76	66	8	2	123	11	10	0	1	237
Feb-88	7	0	17	17	0	109	66	16	7	20	136	28	45	50	13	143	632	567	37	8	867	265	230	32	3	2133
Mar-88	6	0	2	1	1	9	5	0	0	4	29	4	3	12	1	25	53	43	8	2	136	15	17	0	2	239
Apr-88	3	0	0	0	0	13	6	5	0	2	13	6	1	5	1	20	54	39	9	4	126	10	10	0	0	219

3. **Procedures for Incorporating an Additional Year's Data**. The steps needed to add to the data set are fairly complicated, and should only be conducted by persons very familiar with SPSS and Excel. The general steps we have used are as follows.

a. Obtain the data from the National Archive of Criminal Justice Data in SPSS format.

[17] This means that, in most cases, they are arranged by county, since most ORI identifiers are so ordered.

b. Follow the instructions given in the Excel workbook Addyear.xls (attached). This file contains a macro (accessed by pressing the "Extract" button) that will extract the relevant data from the SPSS file and create a number of Excel workbooks for each state.

c. Using the macro "Combine" will put these individual files into a single state file which will permit the user to add the data easily to the original state-level file.
NOTE: we have gotten to this step with the 2003 data (and have enclosed the 51 single-year data sets) but ran out of time before we could add this data to the original data set. The steps that still need to be taken to add the 2003 data in are complicated but not difficult. They cannot be joined directly because (a) some ORIs may have been created in 2003, so additional lines would have to be added to the original data set; (b) some ORIs may have extended into another county for the first time, which also would lead to adding a line to the original data set; (c) some ORIs may have disappeared, perhaps having been absorbed into another agency, meaning that no line for it would be found in the single-year data set. All of these steps can be readily accomplished, but developing an automated procedure will have to wait.

d. Use the file "Superclean.xls" to clean the data thus entered. ***NOTE***: the x-axis scale needs to be changed to permit the additional year to be graphed.

4. **"Missingness" Analysis**. We have attached a draft paper, "Analysis of Missingness in UCR Data," that provides a preliminary analysis of missingness. It describes the variation in missingness by state, year, and population.

ADDITIONAL WORK ON UCR DATA

Aside from the remaining cleaning problems described above, there are three additional tasks that should be accomplished to make the data set especially effective in future studies of crime and criminal justice policy evaluation.

- Task 1 is the development of imputation methods to account for the missing data. As mentioned earlier, we received a grant from the American Statistical Association to develop such methods, and are on schedule to complete it by the end of June 2006.

- Task 2 is to combine individual ORIs to permit the calculation of countywide or SMSA crime statistics. While this effort will help to reduce one problem – the "covered-by" issue described on the previous page (because virtually all of the covering agencies are in the same county as their respective covered agencies, so data like that in the two figures above combine properly) – it exacerbates other problems: accounting for different kinds and amounts of missingness in different agencies at different times; developing point estimates of the county crime counts; and generating confidence intervals around the estimates.

- Task 3 is to go through the data sets to look for anomalies in the crime subcategories. For example, in going through the data for Jefferson AL we noted that, while the Jefferson Police Department reported their Index crimes, they apparently did not break them down by subcategory for 1993 and 1994. A person studying, for example, forcible entry burglaries or gun robberies would be led to believe that there were none during that period.

In addition, we noted that for some subcategories a count of "99" appears, but it really reflects a missing value, not a true count. Because of their frequency we have not been able to resolve all of these anomalies during the grant period.

Aside from these tasks focused on cleaning summary UCR crime data and making it more useful and usable, we note that there are two additional data sets that could use similar attention. The first is the agency-level arrest data collected as part of the UCR program. Although the American Statistical Association requested proposals to perform the same cleaning operation on UCR arrest data, there were not sufficient funds to study UCR and SHR imputation as well as arrest data cleaning and imputation. Because arrest data have many more (and larger) gaps than crime data (Maltz, 1999, cover), and because arrests are a reflection of police policies more so than UCR crime data, these tasks are considerably more daunting than the crime-focused tasks.

The second data set that could use similar attention is the NIBRS data set. This activity will be much more complicated, since NIBRS data is a few orders of magnitude more complex than UCR data or survey data. The missingnesses of survey data are relatively well-understood, and useful imputation techniques have been developed for them (e.g., Little & Rubin, 20XX). But surveys ask basically the same questions of everyone, so imputing missing data is relatively straightforward. NIBRS crime records, on the other hand, may all be different in size, in the number of variables, and in which variables are included. Although the utility of NIBRS data cannot be denied, developing data cleaning and imputation methods for this data set would require a considerably greater level of effort than was required for the UCR data.

REFERENCES

Ehrlich, I. (1975). The deterrent effect of capital punishment: A question of life and death. *American Economic Review*, 65: 397-417.

Federal Bureau of Investigation (2002). *Crime in the United States.* FBI, Washington, DC.

International Association of Chiefs of Police (1929). *Uniform Crime Reporting*. New York: IACP.

Justice Research and Statistics Association (2006). Status of NIBRS in the States. http://www.jrsa.org/ibrrc/background-status/nibrs_states.shtml, accessed February 20, 2006.

Lindgren, S. A., and M. W. Zawitz (2001). *Linking Uniform Crime Reporting Data to Other Datasets*. Report No. NCJ 185233, Bureau of Justice Statistics, Office of Justice Programs, U.S. Department of Justice, Washington, DC, May, 2001. http://www.ojp.usdoj.gov/bjs/pub/pdf/lucrdod.pdf.

Little, R.J.A and Rubin, D.B. (2002), *Statistical Analysis with Missing Data*, Hoboken, N.J.: Wiley-Interscience.

Lott, J. R., Jr. (1998). *More Guns, Less Crime.* University of Chicago Press, Chicago.

Lott, J. R., Jr., and Mustard, D. B. (1997). Crime, deterrence, and right-to-carry concealed handguns. *Journal of Legal Studies* 26: 1–68.

Maltz, M. D. (1977). Crime Statistics: A Historical Perspective. *Crime and Delinquency* 23: 32-40. Reprinted in Eric Monkkonen, Ed., *Crime And Justice in American History*. Meckler, 1990.

Maltz, M. D. (1999). *Bridging Gaps in Police Crime Data*. Report No. NCJ-1176365, Bureau of Justice Statistics, Office of Justice Programs, U.S. Department of Justice, Washington, DC, September, 1999. http://www.ojp.usdoj.gov/bjs/pub/pdf/bgpcd.pdf.

Maltz, M. D., and Targonski, J. (2002). A note on the use of county-level crime data *Journal of Quantitative Criminology* 18: 297–318.

Maltz, M. D., and Targonski, J. (2003). Measurement and Other Errors in County-Level UCR Data: A Reply to Lott and Whitley. *Journal of Quantitative Criminology* 19: 199-206.

Poggio, E.C., Kennedy, S.D., Chaiken, J.M., and Carlson, K.E. (1985). *Blueprint for the future of the Uniform Crime Reporting program. Final Report of the UCR Study.* Abt Associates. NCJRS Report NCJ 98348.