

The author(s) shown below used Federal funds provided by the U.S. Department of Justice and prepared the following final report:

Document Title: **Analysis of Missingness in UCR Crime Data**

Author(s): **Michael D. Maltz**

Document No.: **215343**

Date Received: **August 2006**

Award Number: **2004-IJ-CX-0083**

This report has not been published by the U.S. Department of Justice. To provide better customer service, NCJRS has made this Federally-funded grant final report available electronically in addition to traditional paper copies.

Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Analysis of Missingness in UCR Crime Data

Michael D. Maltz
Criminal Justice Research Center
Ohio State University
Columbus, Ohio

Since its inception over 75 years ago, the crime data series collected by the FBI as part of its Uniform Crime Reporting (UCR) Program has been used to understand the nature and extent of crime in the United States and to establish policies aimed at its reduction. For the most part, those who use crime statistics have some appreciation that they are far from exact, but they are used regardless of their deficiencies because they are, in some respects, the only game in town. This is especially true when it comes to looking at the effect of variations in policies within and across jurisdictions.¹

Using the UCR to analyze policies has gotten more popular of late, especially at the county level (Donohue, 2003a, 2003b; Lott, 1998, 2000; Lott & Whitley, 2003; Maltz & Targonski, 2002, 2003). The reasons are twofold: first, since 1977 county-level crime data has been available on the website of the National Archive of Criminal Justice Data (NACJD)²; and second, there is greater variation in demographic characteristics among counties than among either states or cities, thus permitting an analyst to tease out additional relationships between crime and these characteristics.

The county crime data set, however, is severely flawed, and the method used to calculate the county crime rate changed substantially in 1994 (Maltz, 1999). These concerns were virtually

¹ The National Crime Victimization Survey (NCVS) is very useful (and accurate) in looking at *national* trends, but it cannot be used for understanding *local* policies. Because the NCVS represents the crime victimization experience of a random sample of about 45,000 households nationwide, local (and even state) comparisons usually cannot be made.

² The NACJD is part of the Inter-university Consortium for Political and Social Research (ICPSR) housed at the University of Michigan

ignored in many of the studies cited above, because of an assumption that they had little effect on the outcome of the studies. In fact, at present there is no way of knowing whether the effect size found in those studies is of the same magnitude as the errors in the data. This paper is a first attempt to understand the error structure of UCR crime data, as a means of determining the magnitude of these errors. The analysis is of necessity limited and descriptive, since there are no models or theories that relate to the nature and type of missingness in the data.

I. Introduction

When the Uniform Crime Reporting Program was established in 1929-30, only a handful of police departments initially submitted reports to the FBI. Although the UCR program is voluntary, the number of agencies contributing to the UCR grew substantially in the ensuing decades (Maltz, 1977). But it has never reached 100 percent, because of its voluntary nature (Maltz, 1999) and because agencies sometimes have problems that impact their data collection efforts.

Thus, some agencies have never been represented in the UCR; and for agencies that do normally report their crime data to the FBI, some do not always provide all their data. They may neglect to report individual months or consecutive months or even consecutive years, for a host of reasons (Maltz, 1999: 16).³

Under a grant from the National Institute of Justice (NIJ), we have been working on cleaning the UCR crime data to make it more usable for research and policy purposes. This has entailed investigating the nature and extent of “missingness” in the data so that appropriate imputation procedures can be devised to fill in the gaps. One of the unique benefits of the

³ There are also cases in which the reporting is less than complete, due to agencies’ efforts to cut their crime statistics, if not their crime problems (Maltz, 1999, p. 5, fn. 5).

cleaned data set is that, for the first time, monthly crime data will be available for the over 18,000 police departments for over forty years, from 1960 to (almost) the present.

In examining missingness one encounters the difficulty of initially determining whether a datum is missing (the data files made available to us did not have a specific indicator to signify when a datum is missing) or is merely zero. We have used exploratory data analysis (EDA) techniques (Tukey, 1957) to help in that determination.

When a zero occurs in the data set, there are a number of possible explanations. It might, of course, mean that no crime occurred in that month. Or it might mean that:

1. the agency did not (yet) begin reporting data to the FBI because it didn't exist (or did not have a crime reporting unit) at that time;
2. the agency existed, but it reported its crime and arrest data through another agency (i.e., was "covered by" that agency);
3. the agency existed and reported data for that month, but the data were aggregated so that, instead of reporting on a monthly basis, it reported on a quarterly, semiannual, or annual basis;
4. the agency existed and reported monthly data in general, but missed reporting for one month and compensated for the omission by reporting in the next month aggregate data for both months; or
5. the agency did not submit data for that month (a true missing datum).

Our goal has been to distinguish among these different types of missingness. In this paper I describe the characteristics of the data that are truly missing. In particular, I explore the length of runs of missingness and how they vary by state, by year, and by size and type of agency. Obviously, there are a number of other variables that might be investigated: population, county urbanicity (Goodall *et al*, 1998), and crime rate are some possibilities. However, this initial analysis will focus only on how missingness varies by year, state, and size and type of agency.

The typology of agencies I use is the one used by the FBI, and described in Table 1. As can be seen, the typology is fairly crude, and agencies change their group designation as their population changes.

Table 1. FBI Classification of Population Groups		
Population Group	Political Label	Population Range
1	City	250,000 and over
2	City	100,000 to 249,999
3	City	50,000 to 99,999
4	City	25,000 to 49,999
5	City	10,000 to 24,999
6	City ^a	Less than 10,000
8 (Rural County)	County ^b	...
9 (Suburban County)	County ^b	...

Note: Group 7, 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 7 is included in Group 6.

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

^b Includes State police to which no population is attributed.

There is no theory describing the relationship between missingness and other variables, so this paper will provide a descriptive analysis of how missingness varies with these variables, and attempt to explain the relationships *a posteriori*.

The next section describes the data and the steps we took to make the UCR data more useful. Section III describes the relationships between the run lengths (i.e., the number of consecutive months of missing data) and the variables year, state, and size and type of agency. The consequences of missing data on county-level data, which were explored previously (Maltz & Targonski, 2002, 2003), are reconsidered in Section IV, using more accurate data. The last section describes steps that remain to be taken to understand the effect of missing data on crime data analyses.

II. Cleaning the UCR Data

The FBI archives its annual UCR data files with the NACJD; it houses UCR data from 1966 to 2003 as of this writing; to augment this, we contacted the FBI directly and were able to obtain data for the years 1960-65 as well.⁴

Our initial task was to disaggregate each annual file into 51 state files and reaggregate them by state, so that all of a state's data would be in one file. We used a combination of SPSS scripts and Excel macros to accomplish this task (Maltz & Weiss, 2006).

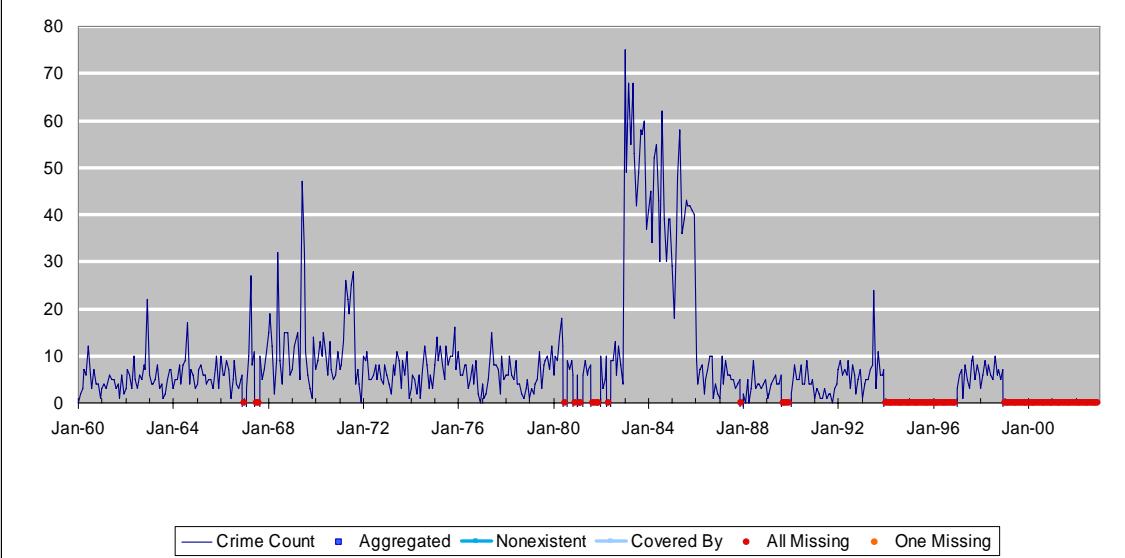
The files may contain indicators of missingness of different types. However, it is not always possible to rely on the accuracy of these indicators.

For example, the files include a variable "Last Update," an indication as to when the data for a given month was received by the FBI. But using this variable is problematic for three reasons. First, in some of the earlier years this variable was not included in the data set; second, sometimes the variable is missing when data for that month is provided; and third, sometimes it is present when data for that month is missing.

Another indicator of missingness is the "Covered By" status, included in the FBI file. This is a problematic indicator, too. It relates to the entire year, when in fact one agency may have begun to cover another agency in the middle of the year. Moreover, there are many instances in which one agency is obviously covering another, but this situation is not recorded. Figure 1 is an example of this situation; the data set indicates that Miles City MT Police Department was covered by the Custer MT Sheriff's Office only for 1985, but the data clearly shows that Custer started covering for Miles City two years earlier.

⁴ There is one notable exception: data for 1962, for state numbers 43-51 (the District of Columbia is considered a state), "were inadvertently erased during an electronic update of the Uniform Crime Reporting (UCR) Program's Master Files" and are not available. The states missing 1962 data are Utah, Vermont, Virginia, Washington, West Virginia, Wisconsin, Wyoming, Alaska, and Hawaii. The last two states are so numbered because they were granted statehood after the FBI assigned numbers to states.

Line 20 of 131, CUSTER, MT00900, NDX



Line 21 of 131, MILES CITY, MT00901, NDX

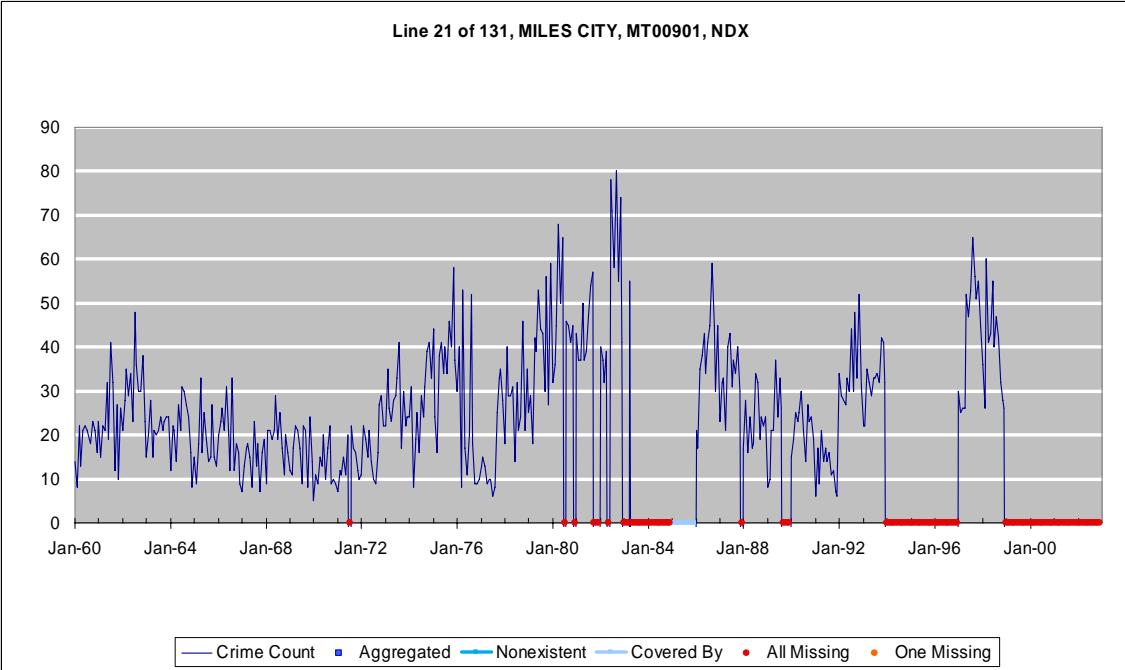


Figure 1. The UCR data set notes that the Miles City MT Police Department was “covered by” the Custer MT Sheriff’s Office in 1985, but the data shows that it started before 1983.

Another type of “missingness” occurs when data are reported quarterly, semiannually, or annually. Cases of this sort are usually indicated by noting that the date recorded in the “Last Update” cell is the same for a number of months in succession – or it may be empty for all months except the month to which the crime counts were aggregated. The data for Jefferson AL indicates reporting only for December from 1993-95, but for every month in 1991, 1992, and

1996-2002. Spreading the data over 12 months allows us to rescale the graph, which provides a clearer picture of crime patterns in Jefferson (Figure 2).

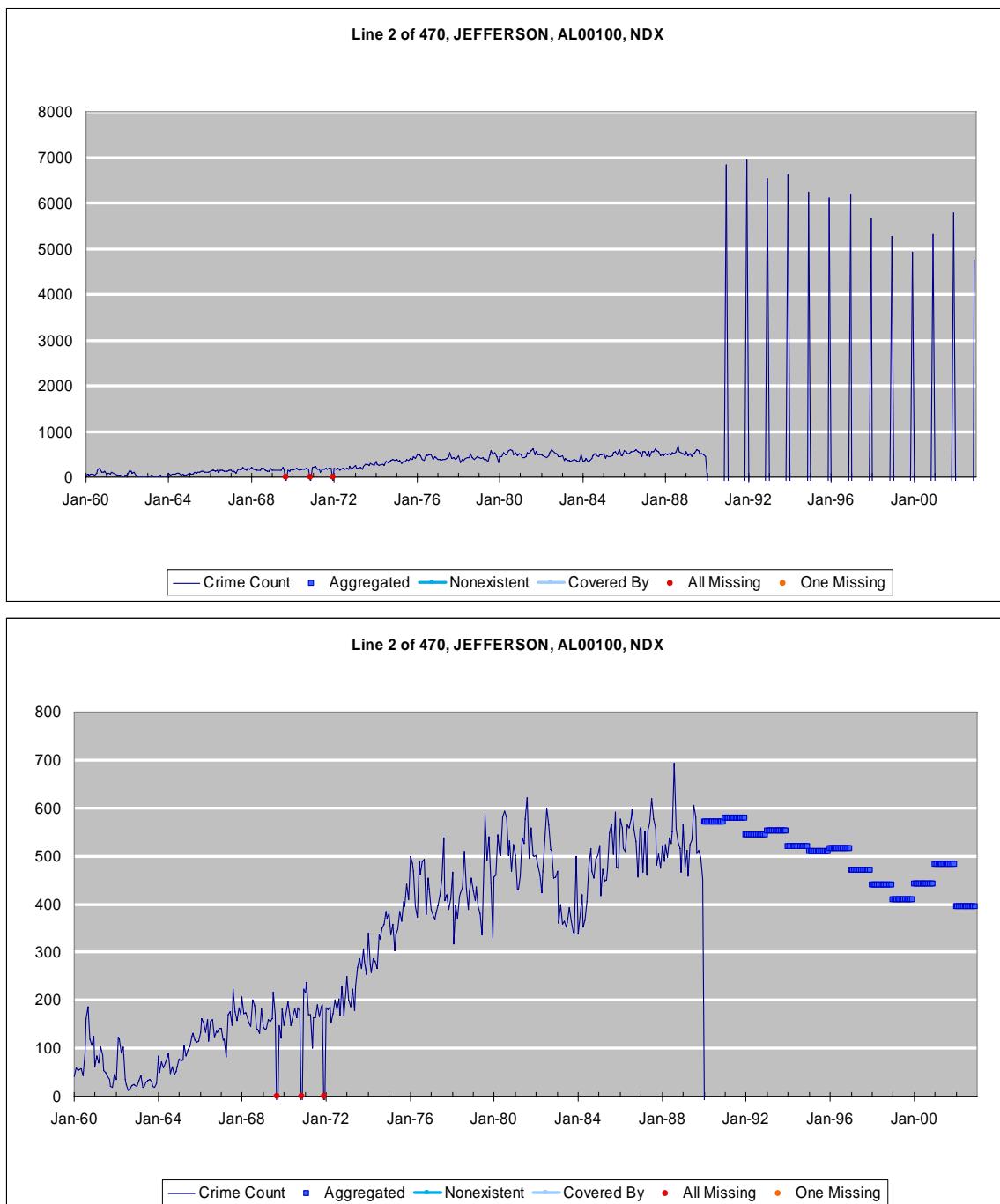


Figure 2. Data spikes represent aggregated data (top panel), and should be spread over the entire year (bottom panel), providing a clearer picture of patterns in the data.

In the end, we decided that the best way to overcome these recording issues was to inspect the data visually. Although there are about 18,000 agencies, the task is made less daunting by automating the process as much as possible.⁵ That is, we used the computer as “power steering” rather than as an “autopilot,” by noting that the computer can be used to select candidates for specifying different types of missingness, but that human judgment is needed to determine whether data is truly missing – and which type of missingness the missing data represents. A description of the process and (the scripts and macros we used) is given in our final report (Maltz & Weiss, 2006).

Of the different types of missing data, the only one that is considered in this paper is true missing data (item 5 in the previous list), when an agency neglects to send data to the FBI.⁶ Sometimes only a single datum is missing (e.g., the rape count is not reported, but all other crimes are reported), but most of the time data for the entire month is omitted.

III. Overall Missingness Pattern

A. Missingness run lengths

Figure 3 shows the overall pattern of missingness for all states, all years. The horizontal axis is scaled logarithmically to highlight the shorter runs. As can be seen, the greatest number of the over 44,000 missingness runs are of length 1, and 70 percent are 10 months or less. The second panel in Figure 3 uses a logarithmic scale for both axes to give a better indication of run length distribution and patterning; although the first 2 peaks are at 1 and 5, respectively, the remaining peaks are in multiples of 12 – i.e., they represent full years of missingness.

⁵ If it takes an average of 60 seconds to review and correct an agency’s trajectory (which is close to the actual value), then it can be accomplished in 300 hours, or 15 20-hour weeks – with time out for eyeglasses!!

⁶ Of course, we were unable to determine if the data sent by the agency was complete; that is, if all of the crimes reported to them for that month were recorded by the agency; see footnote 3.

B. Missingness in different FBI Groups

As shown in Table 1, Groups 1-5 represent cities of different size, but Groups 6-9 represent both cities and other types of jurisdictions, including university, county, and state police agencies, as well as fish and game police, park police, and other law enforcement organizations. Most (but not all) of these “other types” of jurisdictions are called “zero-population” agencies by the FBI, because no population is attributed to them. This is because, for example, counting the population policed by a university police department or by the state police would be tantamount to counting that population twice. So the crime count for these agencies is merely added to the crime count for the other agencies to get the total crime count: for the city, in the case of the university police, and for the state, in the case of the state police.⁷ Many of these agencies are included in Groups 8 and 9.

⁷ Of course, if a university has branches in different cities, each city records the crime count for its respective branch; and if there is a separate state police barracks in each county, then the crime count is allocated by county.

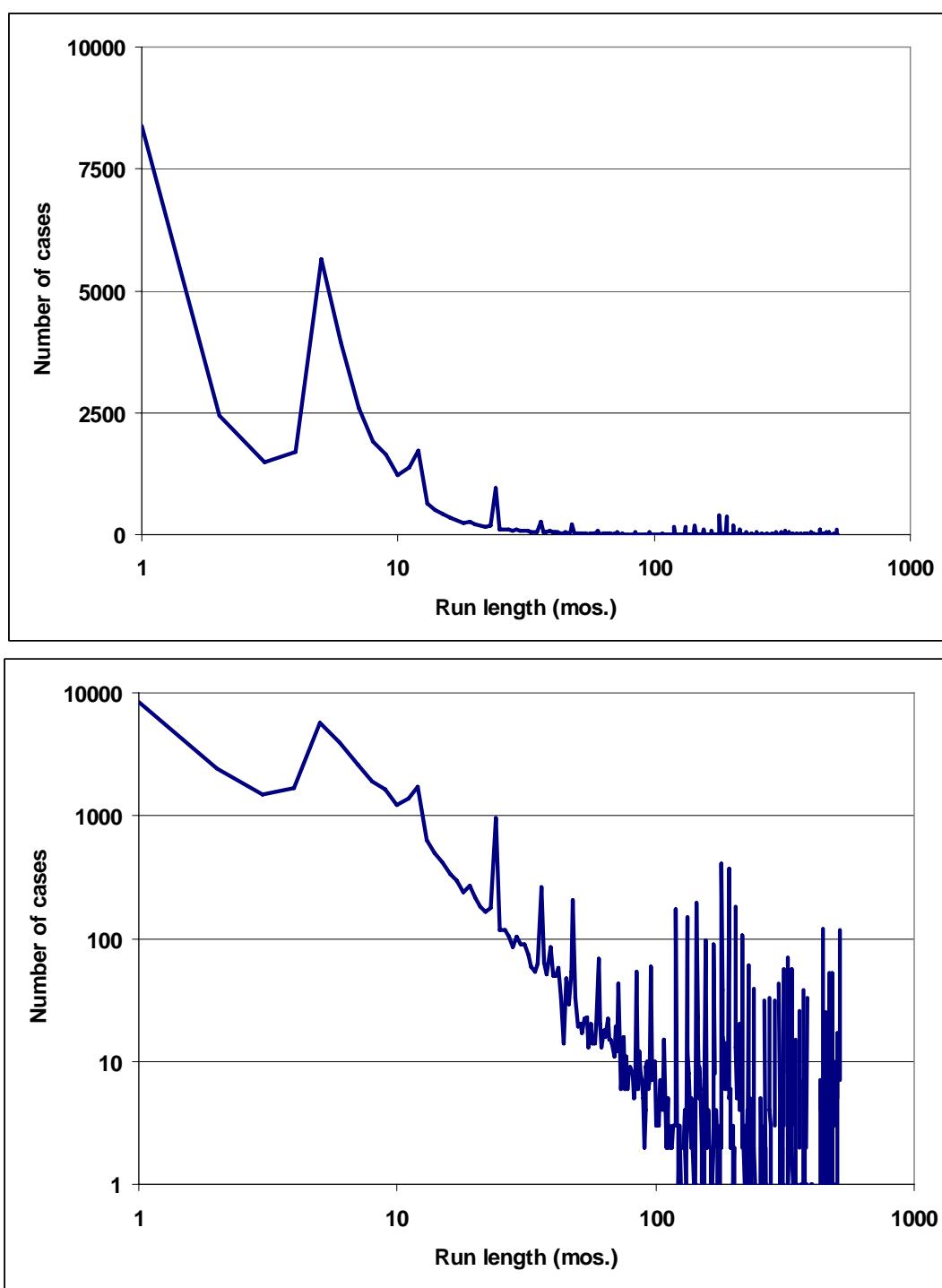


Figure 3. Number of cases of different missingness run lengths, depicted on standard (upper panel) and logarithmic (lower panel) scales. The logarithmic plot clearly shows the periodicity of missingness patterns: most of the run length peaks are multiples of 12, indicating whole years of missing data.

Figure 4 depicts the missingness trends for these different Groups. As can be seen,

Groups 6-9 have the most missingness, concentrated in short run lengths. It also stands to reason

that Group 8 and 9 would have long runs of missingness, since many of the agencies included in these groups (railroad police, fish and game wardens) rarely encounter reportable crimes and are less likely to have an institutionalized crime reporting system.

It also stands to reason that the most populous agencies (Groups 1-5) have the least missingness, and that the larger agencies have the least missingness. First, agencies that have more crime probably have stronger statistical capability; and second, if agencies with populations of 100,000 or more are missing reports, they are contacted by FBI personnel and urged to complete their reports.

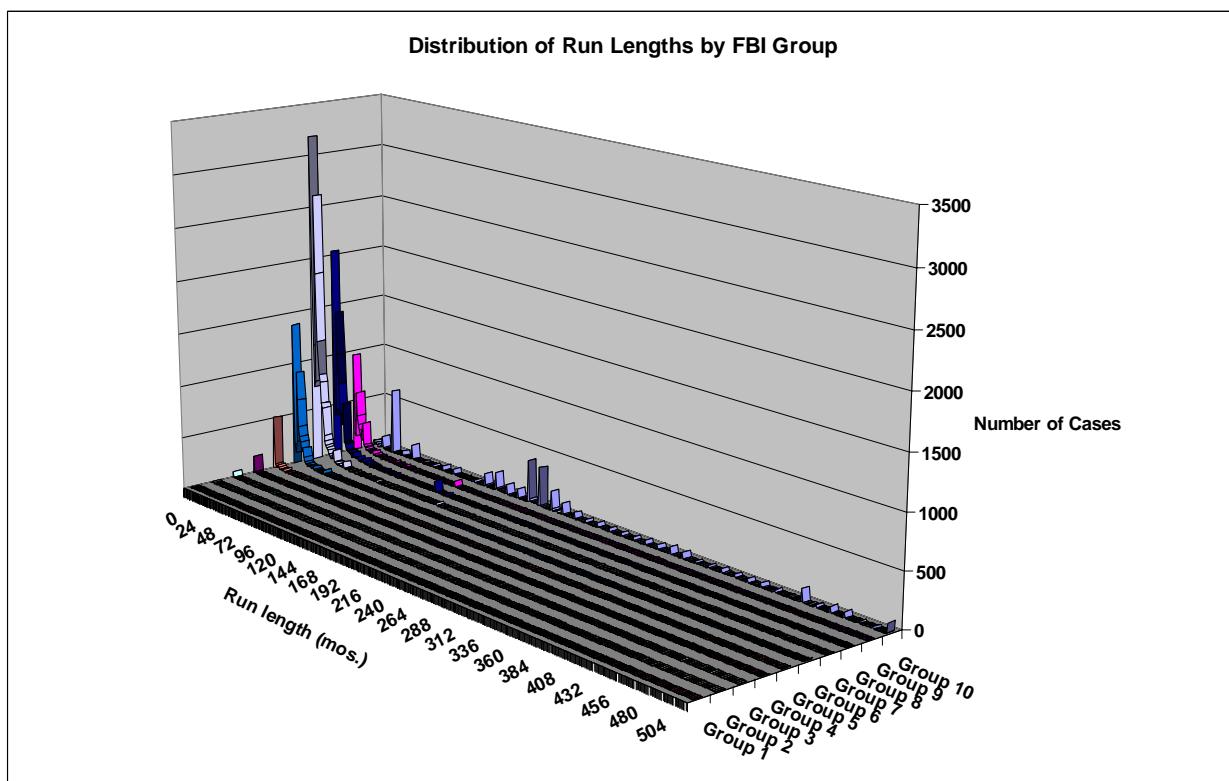


Figure 4. Number of cases of different missingness run lengths for different FBI Groups. The agencies with higher populations (lower Group numbers) have less missing data.

The extent of missingness in Groups 7-9 is substantial, compared to the missingness of Groups 1-6. Group 7, as noted in Table 1, consists of cities with populations under 2,500 and universities and colleges. The trellis plot of this data in Figure 5 (again, with both horizontal and

vertical axes plotted on logarithmic scales) clearly shows that the more populous agencies are considerably less likely to have much missingness (Group 1 agencies, with populations over 250,000, have no missing data).

This figure also shows that the most frequent run length for agencies in Groups 6-9 is 5 months. This peak at 5 appears to indicate that some agencies may be reporting semi-annually, but not indicating that they are doing so. This is not the case, however; a check of the cases with run length 5 does not show most of them to be January-June or July-November runs.

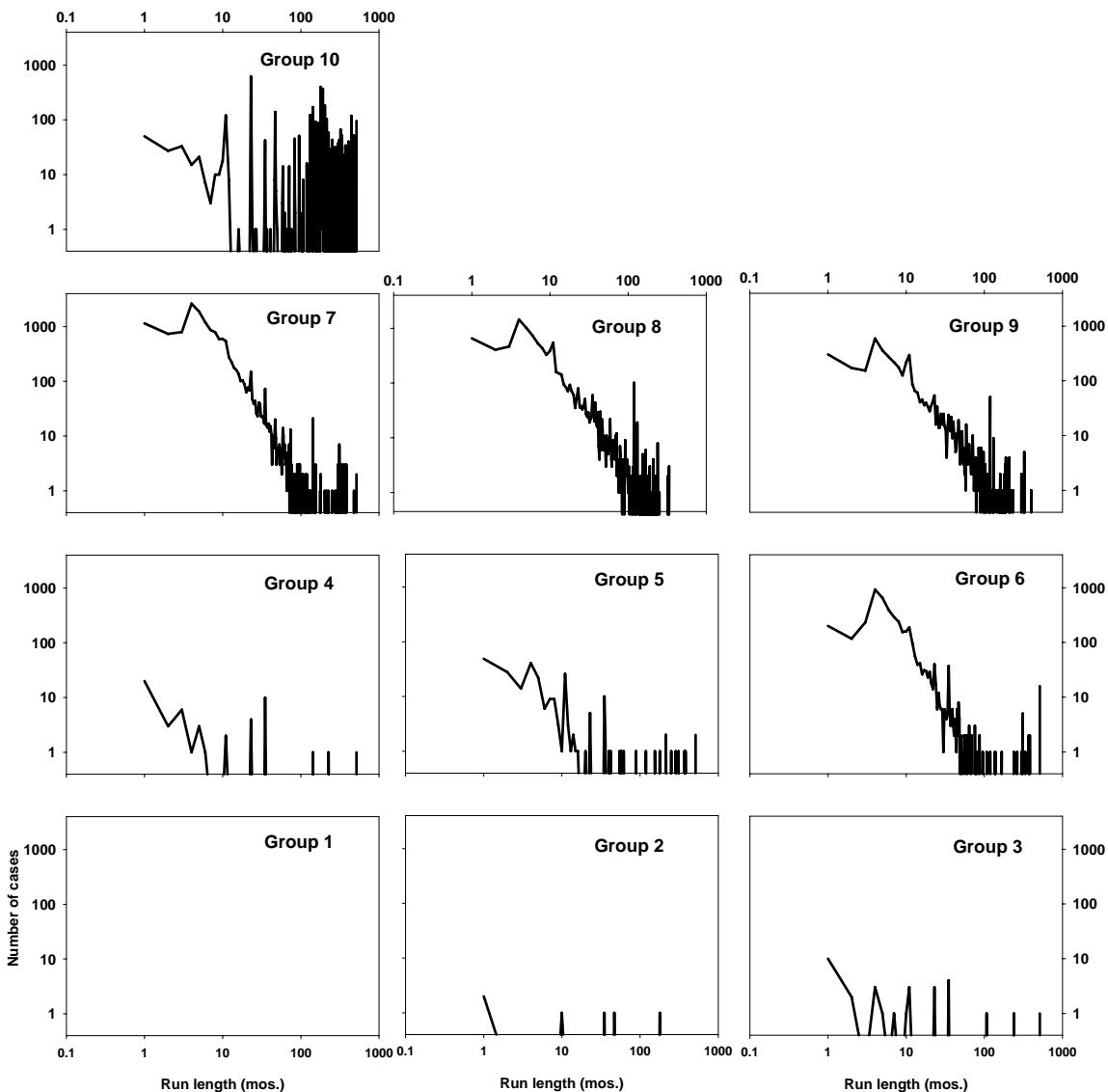


Figure 5. Number of cases of different missingness run lengths for different FBI Groups, plotted on logarithmic scales. The agencies with higher populations have the least missingness.

C. Impact of Missingness on Crime Statistics

An agency was said to have its population (totally) included in the data if it provided at least three monthly reports to the FBI, since if it reported three or more months, its crime count was imputed for the rest of the year based on the reported months – if it only reported four months of data, the crime count for that jurisdiction was estimated to be the total for the four months multiplied by 12/4.

Our analysis prorates the represented population by the number of months reported; that is, if an agency reported four months of data to the FBI, its population representation would be multiplied by 4/12, since that is the actual representation of its population in the (unimputed) data. Figure 6 shows the percent of the population not directly represented in the UCR crime statistics; it ranges between 4 and 16 percent over the 43 years covered.

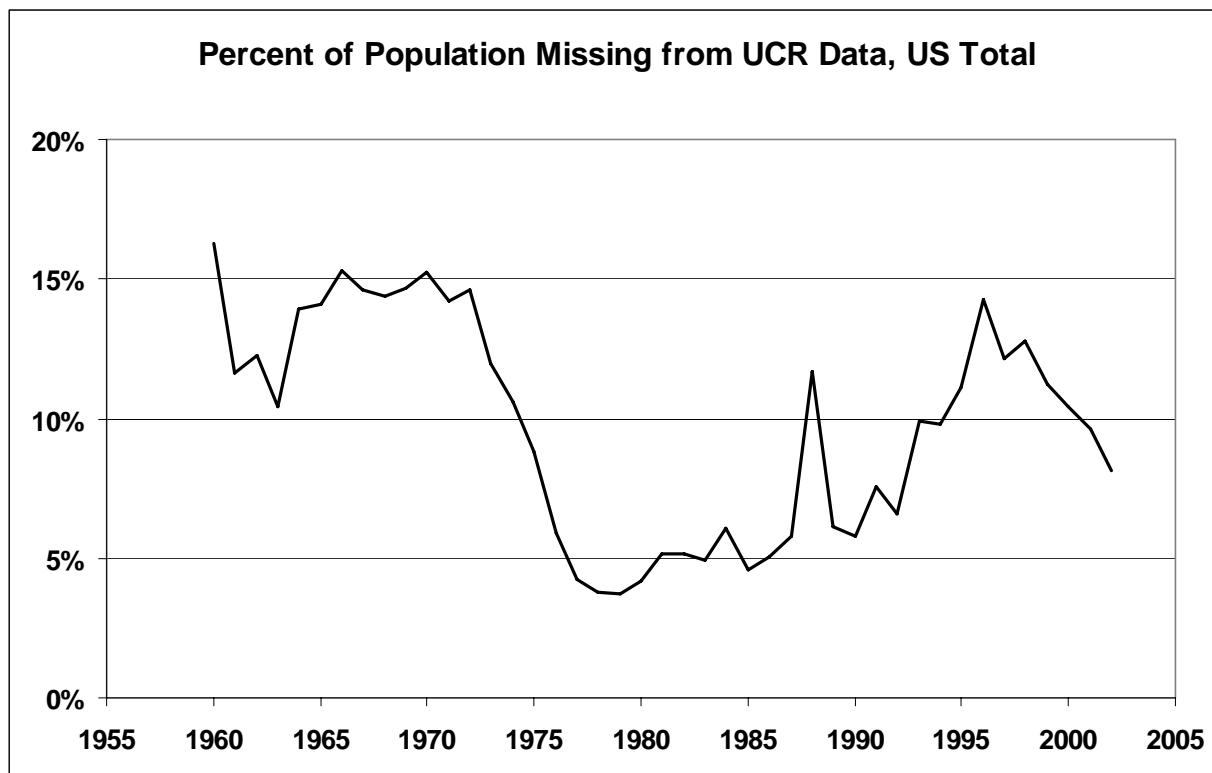


Figure 6. Missing data in national crime statistics Annual variation in the percent of the population not represented in UCR crime statistics

At the state level, however, the reporting is much spottier. Figure 7 shows the nonreporting trajectories of the 50 states on one graph; as can be seen, a number of states have periods when less than half of their population is represented in the crime data.

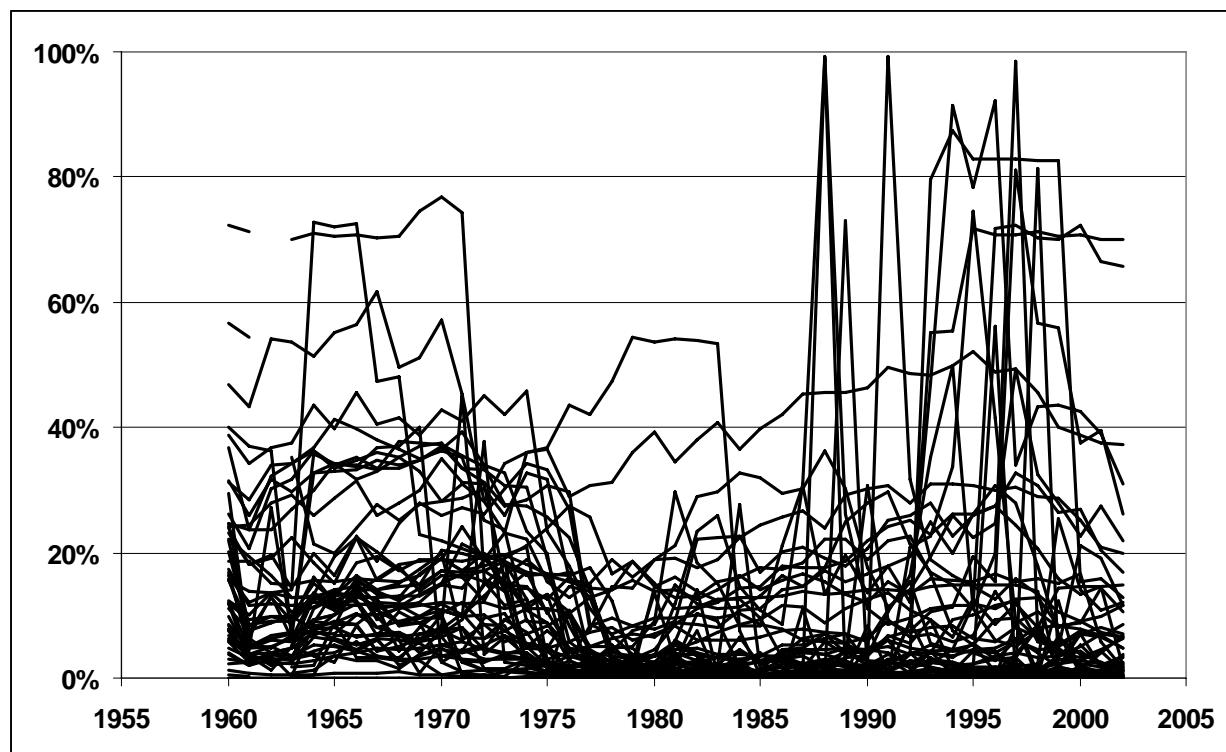


Figure 6. Missing data in state crime statistics Annual variation in the percent of the population not represented in UCR crime statistics.

At the county level the impact of missingness on crime statistics is even greater. To analyze its effect, we calculated the percent of the population missing from each county for each of the 516 (43 x 12) months. As Figure 8 shows, at least 20 percent of the population is missing from over three-quarters of Mississippi's 42312 data points (516 months x 82 counties), and at least 60 percent of the population is missing from over half of the data points. This figure is similar to Figure 6 in Maltz and Targonski (2002, p. 314), which was based on fewer years (1977-92 vs. 1960-2002).

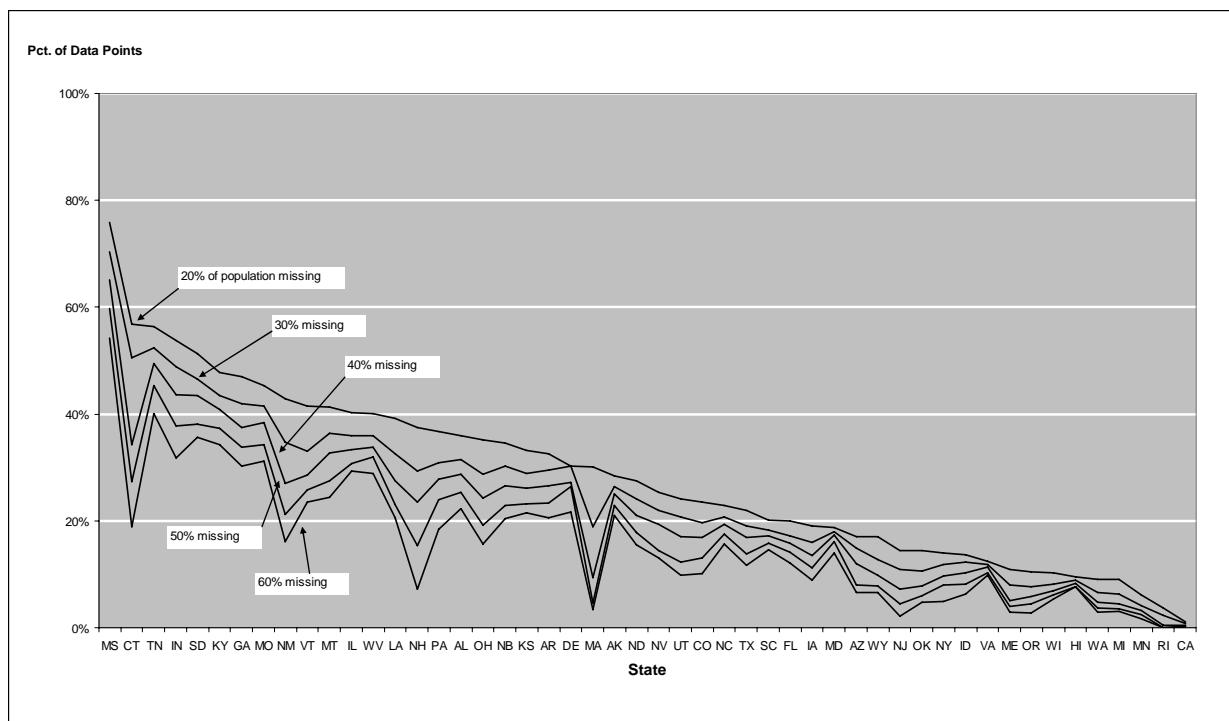


Figure 8. Missing data in county crime statistics. At the top left of the chart, the figure shows that over three-quarters of Mississippi's county-level data points are missing 20 percent of their populations from 1970-2002.

IV. Conclusion

The main lesson to be learned from this look at missing data in the UCR is that the data cannot be used in analyses without regard for gaps in the data. This is especially true for county-level data, as was noted in Maltz and Targonski (2002, 2003). Moreover, even if there were no gaps or inaccuracies in the crime data, calculating rates using the available population data would also cause inaccuracies, as noted in the Appendix. In other words, blithely assuming that the data can be thrown in a canned statistical program and that the only uncertainty in the results are due to “sample size” (put in quotes because the UCR is meant to represent the entire US population, not a sample) is wrong.

This is not to say that the data set cannot be used profitably. The population data can be adjusted, and imputation procedures are being developed to deal with missing data (Ref.). But

attention needs to be paid to the characteristics and missingness in the data to make best use of it, especially in conducting county-level analyses.

V. References

Donohue, J. J., III, and I. Ayres (2003). "Shooting down the 'More Guns, Less Crime' hypothesis," *55 Stanford Law Review* 1193 (2003).

Donohue, J. J., III, and I. Ayres (2003). "The latest misfires in support of the 'More Guns, Less Crime' hypothesis," *55 Stanford Law Review* 1371 (2003).

Goodall, C. R., Kafadar, K., and Tukey, J. W. (1998). Computing and using rural versus urban measures in statistical applications. *American Statistician* , 52: 101–111.

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

Lott, J. R., Jr. (2000). *More Guns, Less Crime*, second edition. University of Chicago Press, Chicago.

Lott, J. R., Jr., and J. Whitley (2003). "Measurement error in county-level data: a response." *Journal of Quantitative Criminology*, 19, 2, 185-198.

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.

Maltz, M. D., and H. Weiss (2006). *Creating a UCR Utility*, Final Report to the National Institute of Justice, Criminal Justice Research Center, The Ohio State University.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, MA: Addison-Wesley Publishing Co.

Appendix; Population Issues in County-Level UCR Analyses

Using population data, the denominator in crime rate calculations, is not straightforward.

During the period in question, additional counties or their equivalent (Census Areas, CAs, or Boroughs in Alaska) were created, requiring that the population data be adjusted to compensate for the changes. Specifically, “The Denali Census Area was created from the Yukon-Koyukuk CA and from a small part of the Southeast Fairbanks CA. Yakutat borough was created from the Skagway-Yakatat-Angoon CA” (email dated 30 March 2006 from Greg Harper, US Census Bureau). Similarly, Broomfield County Colorado first entered the Census data in 2000, containing parts of Adams, Boulder, Jefferson, and Weld Counties.

These are but a few of the boundary and border changes (but the more significant ones) documented by the Census Bureau at

http://www.census.gov/popest/geographic/boundary_changes/. Obviously, they cannot all be included in a study of this sort, but they should not affect the results to any great extent: for the most part we are dealing with population estimates anyway, and these inaccuracies are for the most part minor compared to the errors involved in estimating county populations in intercensal years. To give some indication about the impact of missingness on county-level crime statistics, we need to determine the population of each county and of each agency in the county. This is an effort that can add inaccuracies.

The Census Bureau provides the FBI with agency-level population data. [For agencies that have grown into more than one county, it provides the FBI with their population in each county.] In decennial years the populations are based on the actual census counts, but in intercensal years the estimates are based on Census Bureau projections.

At the next decennial census the earlier estimates are then corrected, but the FBI's population figures are not –the FBI cannot correct all the population figures for the previous nine years, especially since their report, *Crime in the United States*, is distributed widely and new population figures would mean correcting virtually every crime rate they calculate.

To estimate the impact of missingness on county-level data, I used the estimated agency populations and assumed that every missing month translated into 1/12th of the population not included in the crime data for that year. The sum of the missing population for each agency in a county was then compared with the county population for that year – however, the county population I used was not the Census county estimate but rather the sum of the population estimates for the individual agencies within each county.

I compared the Census county population estimates with those calculated in this manner; they usually came to within a few percent of each other. Since the Census figures are estimates themselves, the error introduced by using the FBI's population figures (i.e., Census figures prior to their being revised after a decennial count) is relatively minor.