6-2003

The Application of Missing Data Estimation Models to the Problem of Unknown Victim/Offender Relationships in Homicide Cases.

Wendy C. Regoeczi
Cleveland State University, w.regoeczi@csuohio.edu

Marc Riedel
Southern Illinois University Carbondale

Original Citation

Regoeczi, W. C., & Riedet, M. (2003). The Application of Missing Data Estimation Models to the Problem of Unknown Victim/Offender Relationships in Homicide Cases. Journal Of Quantitative Criminology, 19(2), 155-183.

Repository Citation

THE APPLICATION OF MISSING DATA ESTIMATION MODELS TO THE PROBLEM OF UNKNOWN VICTIM/OFFENDER RELATIONSHIPS IN HOMICIDE CASES

Wendy C. Regoeczi, *Cleveland State University*
Marc Riedel, *Southern Illinois University*

This article was originally published in:

Regoeczi, Wendy C. and Marc Riedel (2003). The Application of Missing Data Estimation Models to the Problem of Unknown Victim/Offender Relationships in Homicide Cases. *Journal of Quantitative Criminology,* 19(2),155-183.

The Application of Missing Data Estimation Models to the Problem of Unknown Victim/Offender Relationships in Homicide Cases

Wendy C. Regoeczi[1,3] and Marc Riedel[2]

Homicide cases suffer from substantial levels of missing data, a problem largely ignored by criminological researchers. The present research seeks to address this problem by imputing values for unknown victim/offender relationships using the EM algorithm. The analysis is carried out first using homicide data from the Los Angeles Police Department (1994–1998), and then compared with imputations using homicide data for Chicago (1991–1995), using a variety of predictor variables to assess the extent to which they influence the assignment of cases to the various relationship categories. The findings indicate that, contrary to popular belief, many of the unknown cases likely involve intimate partners, other family, and friends/acquaintances. However, they disproportionately involve strangers. Yet even after imputations, stranger homicides do not increase more than approximately 5%. The paper addresses the issue of whether data on victim/offender relationships can be considered missing at random (MAR), and the implications of the current findings for both existing and future research on homicide.

KEY WORDS: missing data; victim/offender relationships; homicide; imputation.

1. INTRODUCTION

Our understanding of the nature of crime in society necessarily depends on our ability to collect valid and reliable data describing both the extent of its occurrence and the characteristics of its participants. As criminologists will attest, this is no small feat. And while the development of self-report and victimization surveys have helped to compensate for some of the

[1]Department of Sociology, Cleveland State University, 2121 Euclid Avenue, RT 1724, Cleveland, OH 44115-2214.
[2]Center for the Study of Crime, Delinquency and Corrections, Southern Illinois University, Carbondale, IL 62901.
[3]To whom correspondence should be addressed. E-mail: w.regoeczi@csuohio.edu

limitations inherent in using official statistics, a number of obstacles to obtaining this goal remain. One of the most significant of these, as well as the least addressed, is that of missing data. The purpose of this study is to impute missing values for strangers and other victim/offender relationships using an expectation–maximization algorithm and homicide data for Chicago and Los Angeles. Although the primary focus of this paper is stranger homicide, we consider other victim/offender relationships because imputation of missing values for stranger homicides subsequently affects the proportional distribution of all victim/offender relationship categories.

1.1. Definitions and the Handling of Missing Data

Missing data refers to either unit missing or missing values or both. Unit missing data occur when alternate sources indicate that not all instances of the phenomena have been recorded. For example, in any given year a particular city or state may report a smaller number of homicides to the FBI on the Supplementary Homicide Reports than what is reported in mortality statistics to the National Center for Health Statistics. While some of the variation is due to variations in definition, there are many cases that are not reported by the police, but reported by medical examiners or coroners (Riedel, 1999).

Values are characteristics describing objects, and variables are logical grouping of values. Thus, although we know that a homicide has occurred, the missing value is the lack of information on something like the gender of the victim (Riedel, 2000). While data imputation of values or units in criminology is a relatively unexplored area, we focus on the task of imputing missing values rather than unit missing data.

Rubin and his colleagues (Rubin, 1976; Little and Rubin, 1987, 1989; Madow *et al.*, 1983) have developed a model of missing data that is useful to understand patterns of missingness. Data can be *missing completely at random* (MCAR), *missing at random* (MAR) or *nonignorable*. A great deal of confusion has surrounded the use (or, more typically, misuse) of these terms, particularly the distinction between MCAR and MAR. As such, we will provide definitions of each here.

When data are MCAR, the probability of missing data on a particular variable is unrelated to the value of that variable as well as the values of any other variables in the data set (Allison, 2002). Thus, to meet the assumption that data are missing completely at random, the subset of observations for which there are complete data should constitute a simple random sample of the complete set of cases. In such circumstances, the use of listwise deletion to handle missing data is appropriate. In the context of homicide, for the data to be MCAR, cases which are missing data on victim/offender relationship would not differ significantly with respect to any victim, offender or offense attributes compared with cases with known victim/offender relationships. This assumption can be tested statistically and is highly unlikely to hold.

A somewhat weaker assumption is that the data are MAR. If the data are MAR, "the probability that an observation is missing can depend on the values of observed items but not on the value of the missing item itself" (Heitjan, 1997, p. 549). In other words, the missingness on a particular variable does not depend on variables outside of the dataset being analyzed. With respect to homicide, this would mean that to meet the MAR assumption, missing data on victim/offender relationship could depend on the homicide motive, but within each motive category, the probability of missing the victim/offender relationship must be unrelated to the victim/offender relationship. Unfortunately, unlike the condition of MCAR, there are no statistical tests of the MAR assumption.

When the MAR assumption is not met, the missing data mechanism is nonignorable. Under these conditions, the pattern of missingness would be non-random and not predictable from other variables in the data set. For example, net of other variables in the analysis, cities with high stranger homicide rates might be less likely to report information on victim/offender relationships. The missing data would then be nonignorable. In such circumstances it is typically necessary to model the missing data mechanism to obtain good estimates of the parameters of interest (Allison, 2002; Heitjan, 1997). Missing data mechanisms essentially are an identification of variables explaining why data are missing (Acock, 1997). It has been likened to a logistic regression model which specifies the probability that an item is missing as a function of the values of the data (Heitjan, 1997).

Previous analyses of both of the data sets used in this study have shown the MCAR assumption to be untenable (Regoeczi and Riedel, 1999, 2000). For example, among the 2899 Chicago homicide cases with no missing values, the average victim age is 28.64. This average drops to 25.14 among the 100 cases where the motive and victim/offender relationship variables are missing, and increases to 34.36 among the 83 cases which have missing values for victim/offender relationship, offender gender, offender race, offender age, and total number of offenders (Regoeczi and Riedel, 1999). For Los Angeles homicide cases with complete information ($N = 2071$), the mean age of victims is 29.1. Where offender demographics, motive, and victim/offender relationships are missing, the mean victim age reaches a peak of 33.4. Excluding motive from the preceding pattern drops the mean victim age to 29.3 and it drops further to 27.9 when only offender age, race/ethnicity, and gender are missing.

A further test of the MCAR assumption is through the use of *t*-tests. *T*-tests are based on the hypothesis that, if values are missing completely at random for a given variable, other quantitative variables should have approximately corresponding distributions for cases divided into two groups according to whether or not data are missing (Hair *et al.*, 1995; SPSS Inc., 1997). With respect to the Chicago data, *t*-test results show that the mean age of the victim is *significantly lower* among cases with a value for offender age compared with cases which are missing this value, while *t*-tests for total number of offenders and offender gender indicate that the average number of offenders is *significantly higher* among cases containing a value for offender gender in comparison to cases where this value is missing (Regoeczi and Riedel, 1999). For Los Angeles, mean victim ages were significantly higher when there were missing values for total offenders, victim/offender relationships, and motive. Thus, the removal of cases with missing values on some variables significantly alters the average values of other variables among the complete cases, and hence we can conclude that the data are not MCAR.

The question that remains is whether the data are missing at random (MAR). For the data to be MAR, the probability that a particular variable will be missing values can depend on other observed variables, but not on the variable itself (when controlling for the other observed variables) (Allison, 2000). If the MAR assumption does not hold, the missing data are nonignorable.

While there are no statistical tests of the MAR assumption, there are a variety of different arguments and evidence one can draw on indicating that this assumption holds in the case of data on victim/offender relationships. These include the relationships of other variables in the data set to missingness on victim/offender relationship, prior efforts to recode unknown victim/offender relationships, and research on homicide clearances.

For the present purposes, for the data to be MAR the missingness should be able to be predicted by other variables in the data set, such as victim characteristics, the location, weapon, and circumstances surrounding the offense, and the clearance status of the offense. So, for example, if gang-related homicides are more likely to involve unknown victim/offender relationships than are non-gang homicides, we can adjust for missing data on victim/offender relationship using a variable indicating whether it was a gang-related homicide.

Previous analyses of the Los Angeles data show that in a breakdown of the motive classification where offender demographics are missing, the majority of homicides are gang and organized crime homicides, followed by robbery and other felonies (Regoeczi and Riedel, 2000). For cases missing values on age, race/ethnicity, and gender of offender, there was little evidence to suggest that these cases were predominantly stranger homicides. Rather, it seemed that gang killings made a larger contribution, a conclusion consistent with results found by Pampel and Williams (2000).

Even if stranger homicides are more likely to be missing data on victim/offender relationship than other types of homicides, victim/offender relationship can still be MAR if other variables in the data set can be used to predict this difference. *T*-tests conducted previously on the data used in this study, for example, indicate significant differences in the number of victims and victim age when comparing cases in which the victim/offender relationship is known compared with those where it is missing (see Regoeczi and Riedel, 1999), suggesting that victim characteristics could be used in an imputation model to explain why victim/offender relationship is missing (Acock, 1997).

Note that a causal relationship is not implied by the notion of prediction; in other words we can use information about the victim and offense to predict the victim/offender relationship without having to argue that these variables are a cause of it (see King *et al.*, 2001). Even in situations where the MAR assumption is suspect, procedures based on this assumption often perform well, particularly in multivariate situations, and efforts directed toward improving the modeling of the data structure as opposed to a shift to non-ignorable modeling have been advocated elsewhere (David *et al.*, 1986; Schafer, 1997). The plausibility of the ignorability assumption in any given situation is closely related to the richness of the observed data (Schafer, 1997).

The results of prior efforts to recode unknown victim/offender relationships also suggest the data are MAR. For example, Decker's (1993), careful recoding of all unknown victim/offender relationships for St. Louis homicides indicated that the unknown category consisted of all types of victim/offender relationships, suggesting that the missingness of that variable was not systematically related to the value of that variable. Other research indicates that homicides with unknown victim/offender relationships share several characteristics that are similar to acquaintance homicides (Petee *et al.*, 2001). In fact, at present time there is no concrete evidence on which to base the argument that those homicides in which the victim/offender relationship is unknown are stranger homicides. Rather, as will be discussed shortly, much of the existing evidence suggests quite the contrary.

Trends in clearance rates also support the notion that the data are MAR. In particular, while the clearance rate has declined, the percent of police classified stranger homicides, although they may be underestimated, has remained relatively stable (Riedel, 1998). If missingness depends upon the character of stranger homicides, it would seem reasonable that they should covary with the percent of uncleared homicides, which they do not. Furthermore, a recent multistate study on factors affecting homicide clearance rates found a variety of law enforcement and community characteristics that affected homicide clearances. For example, ‘‘a case was more likely to be solved when witnesses were at the crime scene and provided valuable information, including the circumstances of death, the motivation for the homicide, an identification of the offender, an identification of the victim, or the location of the offender. When a neighborhood survey of the crime scene provided valuable information or the neighbors of the victim were interviewed, the crime was more likely to be solved. However, when friends of the victim were interviewed, the case was less likely to be solved.’’ (Wellford and Cronin, 1999, p. iii) It appears from this study that victim/offender relationships did not play a prominent role in arrest clearances. Combined, this evidence supports the notion that missing values on victim/offender relationships are a byproduct of cases remaining uncleared, but that the lack of clearance is not indicative of a stranger relationship.

This study is premised on the argument that the probability that data are missing on victim/offender relationship is predictable from other characteristics about the homicide event, rather than being due specifically to the nature of the victim/offender relationship. Unfortunately it is not possible to empirically verify this assumption. In this paper we explore how various sets of predictors can be used to impute values for unknown victim/offender relationships in homicide cases.

1.2. Missing Data in Homicide Research

Missing values for offender variables and victim/offender relationships are a consequence of the significant decline in arrest clearances. In 1960, of all murders and nonnegligent manslaughters reported in the United States, 92.3% were cleared by arrest; by 1999, 69.1% were cleared by arrest (Federal Bureau of Investigation, 1999). If about one-third of homicide cases in the U.S. are uncleared, information is missing for offender-related variables. In addition, researchers are missing information where offender input is essential, such as prior relationships between victims and offenders.

Comparisons of cleared and uncleared homicides have focused on two problems. First, research in Canada and the United States indicates that uncleared homicides predominantly involve homicides with concomitant felonies such as robberies or rapes (Cardarelli and Cavanagh, 1992; Regoeczi *et al.*, 2000; Riedel and Rinehart, 1996; Rinehart, 1994; Silverman and Kennedy, 1997).

Second, because most felony homicides are believed to involve strangers, claims have been made by law enforcement officials that uncleared homicides are predominantly stranger homicides. However, such an assumption may be problematic. Riedel (1987), for example, argues that not all felony homicides involve strangers and not all stranger homicides are felony homicides. He further suggests that while missing data may pose a problem, existing research indicates that a minimum of one-third of stranger homicides are not felony-related and only around 20% of robbery killings occur among strangers. A number of other studies have also found that equating felony homicides with stranger killings does not hold across a considerable number of cases (see, for example, Decker, 1993; Flewelling and Williams, 1999; Williams and Flewelling, 1988). It is clear, then, that the level of stranger homicides cannot be merely inferred from the amount of felony-related homicides.

Further complicating reporting problems are recording difficulties in the statistical systems of local, state, and the FBI's Uniform Crime Reporting program. None of these systems have a uniform system of quality control in which the validity and reliability of reports are checked by independent agencies (Biderman and Lynch, 1991). For example, complete information is not available on homicides reported on the Supplementary Homicide Reports that are part of the FBI's Uniform Crime Reporting program (Pampel and Williams, 2000; Riedel 1999; Williams and Pampel, 1998). Comparisons of Supplementary Homicide Reports (SHR) and medical examiner data at the county level suggest that it is less reliably reported than once thought (Wiersema *et al.*, 2000).

A fundamental approach to the problem of missing data is to learn the relationship between missing and nonmissing data and to use that information to impute what the missing values are likely to be. Because missing data are a consequence of a decline in arrest clearances, successful and consistent imputation points to important policy implications for arrest clearances. Increasing arrest clearances is important because regardless of the goals of criminal justice (incapacitation, deterrence, rehabilitation, or retribution), without arrests, there is neither further processing of offenders nor reduction of crime.

2. ESTIMATING STRANGER HOMICIDES

There are two questions with respect to estimating stranger homicides. First, why is it important to estimate stranger homicides as well as other victim/offender relationships? The second question is what approaches have been and can be used for estimation, given the number of stranger homicides cannot be inferred from the amount of felony-related homicides?

2.1. The Importance of Estimating Stranger Homicides

First, homicides involving strangers is one of the most fear provoking crimes faced by an urban dweller. Indeed, Conklin (1975) and McIntyre

(1967) have argued that the fear of crime is, at bottom, the fear of stranger violence. The fear is generated because urban dwellers are often in the presence of strangers who may launch an indiscriminate attack (Riedel, 1993; Silberman, 1978).

Second, from a social constructionist perspective, unreliable and biased data are an opportunity for a variety of claims makers to promote their version of social problems (Spector and Kitsuse, 1987). There is a substantial literature on how statistics are used to shape and promote crime-related social issues (Best, 1988, 1999; Gilbert, 1991; Hotaling and Finkelhor, 1990; Jenkins, 1994; Reuter, 1984). For example, Riedel (1998) has shown that police recorded stranger homicides increased only slightly, from 13.4% in 1977 to 15.1% in 1995. For the same period, unknown relationship percentages increased from 27.0 to 39.4%. The combined percentages for stranger and unknown show an increase from 40.4% in 1977 to 54.5% in 1995—mostly accounted for by an increase in unknowns. By combining unknown relationships with police-recorded stranger homicides, claims were made by the FBI and reinforced by the media that stranger homicides had increased to approximately 53% of all homicides. Such a feat was accomplished by implying that "unknown" referred to strangers rather than a police classification that victim/offender relationships were not known. Because of its fear provoking capability, successful claims that over half of the homicides involved strangers sets the stage for greater claims on criminal justice resources.

Third, in the face of missing data problems, the tendency of researchers has been to use listwise deletion or drop those cases most seriously plagued by missing data (Riedel, 1987). But the assumption that the distribution of characteristics for the events where they are known and for those where they are missing is the same may not be accurate (Williams and Flewelling, 1987). It has been hypothesized, for example, that events where the victim/offender relationship is unknown disproportionately involve strangers (see, for example, Maxfield, 1989; Riedel, 1987). But whether and the degree to which this is so is uncertain.

Fourth, ignoring missing values in the calculation of family, acquaintance, and stranger homicide rates when there is a correlation between the level of missing values and any of the independent variables used in a comparative analysis may lead to erroneous estimates of the effects of these variables (Williams and Flewelling, 1987). Even where such a correlation does not exist, the exclusion of cases from homicide calculations on the basis of missing information may increase random error, which can in turn reduce the model goodness-of-fit, the efficiency of estimates, etc. Both kinds of problems can impinge on the goal of achieving an accurate and sound understanding of variation in homicide rates and its causes (Williams and

Pampel, 1998). Adjustment for missing data is particularly important in the case of longitudinal analysis, since the percent missing varies from year to year (Williams and Flewelling, 1987).

Finally, our focus on estimating stranger homicides and victim/offender relationships occurs because this small body of literature is one of the few research efforts at imputation in criminology. In addition to reviewing this research in the following section, we indicate what models of imputation have been used, the results, and why the current data set and imputation model are more useful for imputation.

2.2. Imputing Stranger Homicides

There are three general perspectives on the content of "unknown" victim/offender relationships. The first of these perspectives, described above, suggests that all unknown relationships involve stranger homicides. The authors have found no empirical support for this position.

A second perspective takes the view that estimating stranger homicides can be done by nothing more than careful coding of available material. Careful coding indicates that stranger homicides may be distributed among the unknown cases in the same proportions as they are among known victim/offender relationships (Decker, 1993).

A third perspective, using different imputation methods, was done by Williams and his colleagues (Williams and Flewelling, 1987; Williams and Pampel, 1998) and by Messner *et al.* (2002). These studies show a higher proportion of stranger homicides.

2.2.1. Careful Coding

Decker (1993) used all available paper records from the St. Louis Police Department from 1985 through 1989, an expanded classification system, and recoded 777 cases. Because of intensive data classification and reliability checks among three coders, only 4% of the victim/offender relationships remained unknown.

Decker recalculated the percentages of victim/offender relationships omitting the category of unknowns (31%). The results are given in Table I.

He found remarkable agreement between St. Louis data and national adjusted scores. For example, Table I shows that 18% of the former and 19% of the latter were stranger homicides. Both the St. Louis and national adjusted scores showed the same percent of homicides involving acquaintances. He found that stranger homicides do not account for the majority of homicides classified as unknown relationships; indeed, they may be

Table I. Comparison of St. Louis and National Victim/Offender Relationships[a]

Nature of Relationship	National data 1985–1989 (%)	National data 1985–1989 adjusted (%)	St. Louis 1985–1989 (%)
Strangers	13	19	18
Acquaintance	32	46	46
Friends	5	7	12
Other relative	8	12	8
Romantic link	11	16	12
Unknown	31		4

[a]Taken from Decker (1993, p. 597).

distributed among uncleared cases in the same proportions as they are among cleared homicide cases. Decker (1993, p. 608) concluded:

> Our ability to classify a large proportion of homicides resulted in a distribution across categories of victim/offender relationships that corresponded closely to national data. This finding suggests that stranger homicides may not account for the bulk of those events which remain unclassified, and that missing data from unsolved homicide cases may not distort the distribution of cases across victim/offender relationships.

2.2.2. Weighting, Adjusting, and Imputing Stranger Homicides

The most extensive research on estimating stranger homicides has been done by Kirk Williams and his colleagues. In a 1987 article, Williams and Flewelling (1987) introduced a weighting and adjustment procedure using SHR data from 1980 through 1984. In a recent article, Pampel and Williams (2000) added an imputation method, compared it to other methods, and compared 1980 and 1990 city data. Both studies used single victim and single offender cases where that information was available.

1987 Research

For the 1980 through 1984 SHR data, Williams and Flewelling (1987) calculated a weighted unadjusted rate and a weighted, within cities adjusted rate, called the circumstances adjusted rate.

For the 1980 through 1984 SHR data, Williams and Flewelling (1987) calculated a *weighted unadjusted* rate by dividing the number of victims reported in the *Crime in the United States* by the number of victims reported in the SHR. The number reported in the former document includes FBI estimations for nonreporting agencies. The unadjusted counts are then multiplied by this weighting factor. Weights were computed for cities over

100,000 and all states in 1980. The effect of this procedure is to compensate for the numbers of unreported events.

The circumstances adjusted rate in the 1987 study consisted of using a variable that is (a) more frequently reported than victim/offender relationships and (b) is correlated with recorded values of the latter. Williams and Flewelling (1987) used felony involvement to adjust stranger homicides because it is a correlate and better reported than stranger homicides. Comparing nationwide SHR data to the circumstances adjusted procedures, the percentages of family homicides declined from 26 to 23%, from 54 to 52% for acquaintances, and increased from 20 to 25% for strangers. The adjustment procedure resulted in small differences between adjusted and reported percentages for family, and acquaintance homicides, but a larger increase in stranger homicides.

2000 Research

Pampel and Williams (2000) compared the unweighted unadjusted rates, which exclude missing information, to rates calculated using two adjustment and one imputation procedure. The weighted unadjusted method is identical with the method used in the 1987 research. The weighted, within-city adjusted method was similar to the circumstances adjusted method in the 1987 study. The same classification of circumstances was used, but victim/offender relationships were divided into family, intimate nonfamily, acquaintance, and stranger homicides. The procedure is described as follows:

> To illustrate the adjustment of family homicide rates, the procedure finds for each city the proportion of all felony homicides that involve family members. It then multiplies that proportion by the number of felony homicides with an unknown offender. When added to the original number of family homicides in the felony category, the product gives an adjusted number of family homicides. It then repeats the calculation for family homicides in each of the other four circumstances. It finally sums the family homicides across the five circumstance categories to obtain an adjusted number of homicides involving family members. Dividing by the population and multiplying by 100,000 turns the number of family homicides into an adjusted rate. The procedure is the same for calculation of intimate nonfamily, acquaintance, and stranger homicides. (Pampel and Williams, 2000, p. 666.)

Finally, there is a weighted, between-city method[4] which weights and imputes missing values for victim/offender relationships. The independent variables used for imputing were gender, race, and age of the victim, the

[4]It is called a weighted between-cities method because imputation is done using data from all cities combined in order to estimate the average relationship between victim and circumstance characteristics for all homicide incidents. Thus, the results yield parameter estimates which are not city-specific. However, since the homicides in each individual city have their own unique distribution in terms of victim and offense characteristics, the imputed values based on the characteristics of the homicide will be unique to each city.

homicide circumstances, weapon type, and size and location of the city. This method begins by computing a multinomial logistic regression using the four types of family relationships mentioned above. The regression saves the predicted probabilities for each category of the dependent variable, both for cases used and those not used because of missing values. Since there are probabilities for each type of victim/offender relationship, the category with the highest probability is assigned a value of "1" while all the other categories are given "0".

Table II gives the percentage distribution by victim/offender relationship by estimation method for 1980 and 1990.

While there is no difference between the weighted unadjusted percentages and the unweighted unadjusted in 1980 and 1990, both the weighted within- and between-city methods show significant changes. In the case of weighted within-city percentages in 1980, acquaintance homicides showed a significant decrease while stranger homicides showed a significant increase when compared with the unweighted unadjusted percentages; this was not true for 1990 within-city adjusted percentages.

In the 1990 data, family, intimate, and stranger homicides show a significant decline for the weighted between-city method while acquaintance homicides show a significant increase in comparison to the unweighted unadjusted percentages. The decline in family and intimate homicides has been documented in other studies (Browne and Williams, 1989, 1993; Browne *et al.*, 1999; Dugan *et al.*, 1999). What is relatively new is the marked increase in acquaintance homicides which may be due to more homicides because of gangs and drugs (Blumstein, 1995).

Table II. Percentage Distribution of Homicide Relationship Type by Measurement Method: 1980–1990 Cities ($N = 91$)[a]

Relationship type	1980				1990			
	UU[b]	WU[c]	WW[d]	WB[e]	UU[b]	WU[c]	WW[d]	WB[e]
% Family	19	19	18	13**	14	14	14	8**
% Intimate	6	6	6	3**	6	6	6	3**
% Acquaintance	51	51	46**	46*	54	54	51	58*
% Stranger	24	24	30**	38**	27	27	29	30*
N	5868	6086	9998	9998	5959	6523	11,587	11,587

[a]Table taken from Pampel and Williams (2000, p. 670).
[b]Unweighted, unadjusted.
[c]Weighted, unadjusted.
[d]Weighted, within-city adjusted.
[e]Weighted, between-city adjusted.
*$0.01 < p < 0.05$; **$p < 0.01$ for t-test of difference between unweighted, unadjusted percentage and each of the other percentage measures.

2.2.3. Log-Multiplicative Association Models

Messner *et al.* (2002) have developed a very different approach to imputing missing values for unknown victim/offender relationships which is based on a log-multiplicative model known as the heterogeneous column RC(L) model. In this model the category of unknown victim/offender relationships is "scaled" relative to those categories in which the victim/offender relationship is known based on associations with other variables. The scale scores are then used to allocate cases with unknown victim/offender relationships. Using this technique, they impute values for unknown victim/offender relationships in SHR data separately for the years 1996 and 1997 based on the association between victim/offender relationships and circumstances (felony; other felony; non-felony; other non-felony; undetermined). Their imputation method results in a greater proportion of unknown victim/offender relationships being allocated to the stranger category (which increased from 17 to 24%) than the methods used by Williams and his colleagues, while the proportion of cases in all other categories declined after imputation.

3. THE PRESENT RESEARCH

The current research improves upon past research in four fundamental ways. First, Williams and his colleagues, as well as Messner *et al.*, relied on data available from the SHR. Although widely used as a data source, it has limitations that are avoided by using data from a city police department. Thus, the present analyses take advantage of the superior quality of homicide data available from the Los Angeles and Chicago police. This allows us to both circumvent some of the limitations of using SHR data, as well as to examine whether imputations differ across data sets drawn from different cities, where the nature of homicide may also vary.

Second, we improve upon past research with respect to the categorization of victim/offender relationships. Williams and Flewelling (1987) do not distinguish between spousal and other family relationships. Moreover, Maxfield (1989) argues that the highly aggregated categories used by Williams and Flewelling in their development of adjusted homicide rates lead to the loss of important distinctions between event types. The categorization of victim/offender relationships used by Williams and Pampel (1998) does expand beyond that used in Williams and Flewelling (1987) to include family, intimate non-family, acquaintance, and stranger. However, there may be problems with their inclusion of spousal homicides together with other family homicides, while boyfriends, girlfriends, ex-wives, ex-husbands, and homosexual couples constitute a separate category. Thus, there remains a need for analyses using more precise victim/offender categories. The current research tests several categorizations of victim/offender relationships to see whether the number and coding of victim/offender relationship categories has an impact on the degree to which unknown categories will be assigned to the stranger category.

Third, the current research also seeks to examine whether and how imputations are affected by the set of predictors from which the parameters are estimated. This is accomplished by varying the types of variables used. Our imputation procedure seeks to take advantage of as much information as possible about the characteristics of victims, offenders, and the offense in allocating unknown cases to victim/offender relationship categories. We use far more variables than any of the existing research, much of which relies on a single variable—the circumstances surrounding the offense—which also suffers from high levels of missing values (for Los Angeles, 11% of cases are missing information on circumstances; for Chicago, 25.5% of cases are missing information on circumstances), making it a dubious predictor of other unknown variables.

The final issue has to do with the method of imputation itself. Among the available methods for handling missing data are listwise deletion, pairwise deletion, mean substitution, hot-deck procedures, regression, and expectation–maximization (EM). Only the latter three methods impute a *value* from the predictive distribution. While hot-deck procedures have been shown to reduce bias associated with nonresponse, these gains are offset by corresponding increases in the variance of estimates (Cox and Folsom, 1978). Although regression models are widely used, they are not appropriate here because the dependent variable and many of the predictor variables are categorical data. In addition, Acock (1997) notes that regression based techniques result in overprediction because there is a lack of adjustment for errors in prediction. Hence, missing values, now replaced with predicted values, will be perfectly predicted where the same independent variables are being used for explanation. The EM algorithm adds residual error terms which correct for the underestimation of variances that typically befalls more conventional imputation methods (Allison, 2002). Furthermore, the EM algorithm is able to accommodate missing data on the predictors as well as the variable to be imputed and uses the full set of variables as predictors in the imputation process. This is highly significant, since, as noted by King *et al.* (2001), the MAR assumption can be made to fit the data by including more variables in the imputation process (see also Acock, 1997; David *et al.*, 1986). Acock (1997) reviewed a large number of missing data estimation models and concluded that expectation–maximization (EM) is the best general solution to missing data problems.

As a method of imputation, EM has very good properties when the data are MAR. Other methods of imputation, such as multiple imputation, have many of the same optimal properties (see Allison, 2002). However, the EM algorithm may be preferable for two reasons: (1) unlike multiple imputation, it produces a determinant result; and (2) multiple imputation has optimal statistical properties only when producing an infinite number of data sets. Thus, our research uses the EM algorithm as the basis for imputation.

4. METHODOLOGY

4.1. Data

The data for this study were derived from two sources. To minimize intercity variation we analyze separately two cities: Los Angeles and Chicago. The Los Angeles homicide data are taken from the California Homicides Data File and consist of all homicides occurring in the jurisdiction of the Los Angeles Police Department and reported to the California Criminal Justice Statistics Center. The content of the Criminal Homicides Data File is similar to data found in the SHR with some additions. For example, the Criminal Homicides Data File contains the month, day, and year of the homicide as well as the month and year it was reported to the UCR. To minimize the effect of trends, data used in this study consist of 3,380 wilful homicides from 1994 through 1998. Excluded were 60 justifiable homicides by private citizens, 10 manslaughters, and 71 justifiable homicides by police officers. It was felt that analyzing homicides over a five year period would reduce year-to-year aberrations in reporting.

The Chicago homicide data were derived from the Homicides in Chicago Data File (Block and Block, 1997). This file contains information collected on all homicides included in the murder analysis files of the Chicago Police Department for the years 1965 through 1995. Justifiable homicides and manslaughters are excluded. Since using the full set of cases contained in the data file would be too cumbersome given the nature of the research, the current analysis uses only those cases for the 5-year period between 1991 and 1995, a total of 4459 cases.

4.2. Measures

Given the comparative nature of the research, the predictor variables in the two data sets were coded to be as similar as possible. The frequencies and levels of missingness for all variables are displayed in Table III.

Table III. Frequencies and Levels of Missingness for All Variables

Variable	Chicago observed	Chicago % missing	LA observed	LA % missing
Victim/offender relationship	$N = 3225$	27.7	$N = 2736$	19.1
Intimate partner	9.6%		4.2%	
Other family	7.4%		4.1%	
Friend/acquaintance	67.3%		51.3%	
Stranger	15.7%		40.4%	
Victim gender	$N = 4459$	0	$N = 3380$	0
Male	83.9%		87.2%	
Female	16.1%		12.8%	
Victim age	$N = 4448$	0.2	$N = 3363$	0.5
Victim race/ethnicity	$N = 4459$	0	$N = 3375$	0.1
White victim	8.3%		9.9%	
Black victim	76.0%		34.5%	
Latino victim	15.0%		52.2%	
Other victim	0.7%		3.3%	
Total number of victims	$N = 4459$	0	$N = 3380$	0
Offender gender	$N = 3181$	28.7	$N = 2197$	35.0
Male	90.6%		95.2%	
Female	9.4%		4.8%	
Offender age	$N = 3103$	30.4	$N = 2121$	37.2
Offender race/ethnicity	$N = 3130$	29.8	$N = 2190$	35.2
White offender	6.3%		7.0%	
Black offender	77.8%		38.4%	
Latino offender	15.3%		52.4%	
Other offender	0.7%		2.2%	
Total number of offenders	$N = 3504$	21.4	$N = 2740$	18.9
Location	$N = 4459$	0	$N = 3380$	0
Private indoor	29.9%		22.1%	
Public indoor	6.8%		3.7%	
Public outdoor	51.0%		65.6%	
Vehicle	12.4%		8.6%	
Circumstances surrounding the offense	$N = 3320$	25.5	$N = 3014$	10.8
Domestic altercation	18.2%		6.3%	
Other altercation	38.0%		27.8%	
Felony-related	16.7%		20.7%	
Gang/organized crime	25.9%		43.8%	
Other motive	1.3%		1.4%	
Weapon	$N = 4459$	0	$N = 3373$	0.2
Handgun	24.5%		71.7%	
Longgun	10.9%		7.9%	
Knives	13.4%		9.5%	
Other weapons	14.6%		10.8%	
Semi-/fully-automatic	36.6%			

Victim/Offender Relationships. Victim/offender relationships were coded into four dummy variables: intimate partners (reference category), other family, friends/acquaintances, and strangers.[5]

Gender. Gender was coded separately for victims and offenders with males coded as "1" and females coded as "2."

Age. Age was treated as a continuous variable for both victims and offenders. Offender age was logged in both data sets, and victim age was logged in the Los Angeles data, to correct for skewed distributions of these variables.

Race/Ethnicity. For both victims and offenders, race/ethnicity was coded into a set of four dummy variables: white (reference category), Black, Latino, and other.

Total Number of Victims and Offenders. Both of these variables were treated as continuous. The total number of offenders was logged in both data sets to correct for skewed distributions.

Location. Locations were grouped into four dummy variables: private indoor location (reference category); public indoor location; public outdoor location; vehicle.

Circumstances Surrounding the Offense. The circumstances surrounding the offense were categorized into domestic altercation (reference category), other altercation, felony-related, gang/organized crime, and other motive.[6]

[5]Coding of victim/offender relationship for the Los Angeles data was as follows: (1) intimate partners (husband—legal or common-law; wife—legal or common-law; ex-husband; ex-wife; boyfriend; girlfriend; homosexual relationship), (2) other family (mother; father; son; daughter; brother; sister; in-law; stepfather; stepmother; stepson; stepdaughter; other family), (3) friends/acquaintances (neighbor; acquaintance; employee; employer; friend; gang member; other known to victim), (4) strangers. Coding of victim/offender relationship for the Chicago data was as follows: (1) intimate partners (husband—legal or common-law; wife—legal or common-law; ex-husband; ex-wife; ex-common-law husband; ex-common-law wife; boyfriend; girlfriend; ex-boyfriend; ex-girlfriend; homosexual couple), (2) other family (which includes categories like father; mother; son; daughter; half-brother; half-sister; uncle; aunt; stepfather; stepmother; foster father; foster mother; father-in-law; mother-in-law), (3) friends/acquaintances (which includes categories like landlord; roomer/roommate; business partners; employer; neighbor; acquaintances; gang member; sexual rivals; cell mate/inmate; informant of crime; restaurant/bar staff; drug pusher), (4) strangers.

[6]The circumstances surrounding the offense were coded as follows: (1) domestic altercation (which includes categories like lovers' triangle; altercation over children; general domestic altercation; sexual altercation; altercation over desertion/termination of relationship), (2) other altercation (which includes categories like altercation over gambling; argument over money or property; altercation over politics; racial/hate altercation; altercation over (alleged) theft), (3) felony-related (which includes categories like burglary; armed robbery; rape; unlawful use of a weapon; victim is a narcotics dealer; victim is a prostitute; arson; attempted theft/shoplifting; blackmail; deceptive practice; ransom), (4) gang/organized crime (gangland killing; drive-by shooting; organized crime; contract killing; contract arson; sniper attack), (5) other motive (medical treatment; escape; insurance fraud; mental disorder; mercy killing; suicide pact).

Weapons. In the Los Angeles data, weapons were categorized as handguns, longguns, knives, and other weapons. Because the Chicago data contained an additional category of semi-/fully-automatic weapons from which it could not be deciphered whether the gun was a handgun or a longgun, this category was retained as a separate category in the analyses of the Chicago data. The remaining weapons were classified into the categories of handgun, longgun, knives, and other weapons. Knives functioned as the reference category.

4.3. Analysis

Imputation of missing values was carried about by way of the Expectation–Maximization (EM) algorithm. EM is a computational method for obtaining maximum likelihood estimates in situations where there are missing data (Allison, 2002; Dempster *et al*., 1977). EM is a technique particularly well suited for imputing missing values where there are few continuous variables as is the case with much criminal justice data. EM is carried out in two steps; the first step, the E step, finds the conditional expectation of the missing values given observed values and current estimates of parameters. The second step, the M step, consists of finding maximum likelihood parameters as though the missing values were filled in. The process is repeated although the second cycle now has missing value estimates from the first cycle.

After each EM step, a covariance matrix is computed; when the values of the covariance matrix do not change or change by trivial amounts, the process comes to a halt (Acock, 1997). Modified formulas are used for the variances and covariances of terms involving missing data, which entails the addition of terms corresponding to residual variances and residual covariances. The inclusion of these residual terms provides a correction for the problem of underestimated variances that befalls conventional imputation schemes (Allison, 2002).

We use a SAS macro developed by Paul Allison called *Macro Miss* for carrying out the EM procedure.[7] This macro is available from Allison upon request.[8] It uses the EM algorithm to carry out maximum likelihood estimation of the mean and covariance matrix of the multivariate normal dis-

[7]As a general approach to missing data imputation, EM will produce estimates of the standard error that are too low, and consequently overestimate the correlations, because it treats the imputed data as if they were real values (Allison, 2002). However, some software programs correct for this problem by making random draws from the residual distribution of each imputed variable, which are then added to the imputed values (Allison, 2002). Allison's macro uses this method as a means of incorporating additional random variation due to uncertainty surrounding the parameter estimates.

[8]www.ssc.upenn.edu/~allison

tribution for incomplete data. The algorithms in the macro are modeled on those discussed by Schafer (1997).

Models were run separately for the Los Angeles and Chicago data. Each model was analyzed twice using different sets of predictor variables. In the first model, only victim characteristics were included in the model as predictors, since these predictors contained far fewer missing values than offender-related variables. Included in this model were the following: victim gender; victim age; victim race/ethnicity; homicide location; motive; weapon; and number of victims. Reference categories were selected on the basis of the assumption that they contained cases with fewer missing values than for other categories. Once values were imputed for this model, it was rerun adding the offender-related variables of offender gender, age and race/ethnicity (with white offender as the reference category), and total number of offenders. This permitted a comparison of results to assess the possibility that additional information could be gleaned from the extra predictor variables. Although frequently cases which are missing information on victim/offender relationship are also missing data on other offender-related variables, this is by no means a universal pattern. For example, there are 191 cases in the Chicago data which are missing on victim/offender relationship but contain data on offender gender and an additional 167 cases containing data on offender race. Because EM always begins with a full covariance matrix, this allows regression estimates to be obtained for any set of predictors, regardless of how few or many cases exist within any particular pattern of missing data (Allison, 2002). Since the assumption underlying this imputation technique is that missing data on victim/offender relationship can be explained by the observed data, we test different models as an attempt to capitalize on the maximum amount of information for predicting the missing values.

When values are imputed for categorical variables such as victim/offender relationship, the initial values will be relatively meaningless. That is, a value of 0.53 for a variable which can take only one of two values (0 or 1), is relatively meaningless if left in this form. As a result, it is necessary to assign values of 0 and 1 by applying a basic set of rules. In the case of a dichotomous variable, values of 0 and 1 are assigned on the basis of which is closer to the imputed value. For a four-category variable such as victim/offender relationship, the variable will be represented by three dummy variables. After imputation is complete, the following must be determined. If the imputed values for the three dummy variables can hypothetically be thought of as X_1, X_2, and X_3, then all three should be set to zero if $1-X_1-X_2-X_3$ is greater than either X_1, X_2, or X_3. Otherwise, if X_1 is greater than X_2 and X_3, X_1 should be assigned a value of 1 and X_2 and X_3 should be assigned values of 0, and so on (Allison, 2002).

5. RESULTS

Using *Macro Miss* and assigning values on the basis of the rules delineated above, the following results were obtained. Table IV shows the distribution of victim/offender relationships for Los Angeles homicides both before and after imputations first using victim and offense variables as predictors and then with the addition of offender-related variables. What is most striking about these results is how little the distribution changes once the "unknown" victim/offender relationships are assigned to one of the four "known" categories. For both models (with and without offender-related variables), the percentage of cases in each victim/offender relationship category changes by less than 1% after missing values are assigned.

The proportion of homicide cases involving strangers is unusually high in Los Angeles—around 40%. To examine the extent to which the imputation results may be due to the unique nature of homicide in this city, missing values on victim/offender relationships were imputed for Chicago. As with the Los Angeles data, the model was run twice, first without and then including offender-related variables. The results are displayed in Table V.

The distribution of victim/offender relationships among homicides committed in Chicago differs considerably from that in Los Angeles. In particular, there are proportionately more cases involving intimates, other family, and friends/acquaintances, and consequently fewer stranger homicides in Chicago than Los Angeles. The proportion of "unknown" victim/offender relationships also differs between the two cities. In Chicago, 27.7% of cases involve unknown victim/offender relationships, compared with 19.1% in Los Angeles. However, in terms of the extent to which the distribution of victim/offender relationships changes once missing values are

Table IV. Distribution of Homicide Victim/Offender Relationships for Los Angeles Before and After Missing Value Imputation, 1994–1998

	Before imputation		After imputation (victim and offense predictors)			After imputation (victim, offender and predictors)		
Victim/offender relationship	# of cases	% of cases	# of cases	% of cases	Difference (%)	# of cases	% of cases	Difference (%)
Intimate partner	115	4.2	125	3.7	−0.5	121	3.6	−0.6
Other family	112	4.1	118	3.5	−0.6	123	3.6	−0.5
Friend/ acquaintance	1404	51.3	1752	51.8	+0.5	1748	51.7	+0.4
Stranger	1105	40.4	1385	41.0	+0.6	1388	41.1	+0.7
Total	2736	100	3380	100	0	3380	100	0

Table V. Distribution of Homicide Victim/Offender Relationships for Chicago Before and After Missing Value Imputation, 1991–1995

Victim/offender relationship	Before imputation # of cases	Before imputation % of cases	After imputation (victim and offense predictors) # of cases	% of cases	Difference (%)	After imputation (victim, offender and predictors) # of cases	% of cases	Difference (%)
Intimate partner	309	9.6	378	8.5	−1.1	385	8.6	−1.0
Other family	239	7.4	284	6.4	−1.0	288	6.5	−0.9
Friend/acquaintance	2172	67.3	3054	68.5	+1.2	3022	67.8	+0.5
Stranger	505	15.7	743	16.7	+1.0	764	17.1	+1.4
Total	3225	100	4459	100.1	0	4459	100	0

imputed for the unknown cases, the results are very consistent across the two data sets. Regardless of whether offender-related variables are included as predictors, the percentage change in the distribution of intimate partner, other family, friend/acquaintance, and stranger homicides in Chicago is minimal. The greatest change occurs for stranger homicides in the model containing victim, offender, and offense variables as predictors (last three columns of Table V), but even there the increase is only 1.4%.

Admittedly, these results are quite unexpected, and they differ from the imputation results of Williams and his colleagues and Messner *et al.* They are, however, quite consistent with the work of Decker (1993), whose recoding of St. Louis data revealed that the distribution of unknown victim/offender relationships was the same as among those where the relationship is known.

There are, however, two factors which may be affecting the results. The first concerns the categorization of victim/offender relationships. Since our ability to predict unknown victim/offender relationships is influenced by the richness of the observed data and the complexity of the data model (Schafer, 1997), we expand the number of victim/offender relationship categories to try and capitalize on this more detailed information available in the data set. We ran a second set of imputations using a 6-category classification of victim/offender relationships which drew on the distinction between blood-related and non-blood related variables, as emphasized in the work of Daly and Wilson (1988). Victim/offender relationships were reclassified into the following categories: intimate partners, primary-blood, primary-other, secondary relationships, crime-related relationships, and strangers. Values were imputed for the Los Angeles data, first using victim- and offense-related variables and then adding offender-related variables. Due to the similarity of the results, only the former are shown (Table VI).

The findings for this set of imputations reveal that refining the classification of victim/offender relationships to include more categories does little to change the percentage distribution of cases after missing values are imputed for the unknowns. The biggest change occurs in the crime-related category, but the increase is only 1.6%. The similarities of the distributions pre-and post-imputation are really brought home by the finding that there is no change at all in the percentage of cases involving strangers. Very similar results were obtained for Chicago using the 6-category victim/offender relationship classification (Table VII).

Examining Table VII, we see that although the initial distribution of known cases among the victim/offender relationship categories differs somewhat from the Los Angeles data, the imputation results are consistent with those found for Los Angeles. The distribution of cases does not change substantially once the unknown cases are assigned to one of the existing categories based on the imputation results. Thus, it appears that a more refined classification of victim/offender relationship categories does not have a significant impact on the extent to which imputed values change the distribution of cases across these categories. To this point, then, we are forced to agree with Decker (1993) that the distribution of unknown cases mirrors very closely the distribution of cases for which the victim/offender relationship is known.

The other possibility that must be considered as influencing the results concerns the predictor variables. Thus far we have used a set of variables which describe the characteristics of the victim and offense, and sometimes also the offender. Conspicuously absent from this list is a variable that is

Table VI. Distribution of Homicide Victim/Offender Relationships Using an Alternative Classification for Los Angeles Before and After Missing Value Imputation Using Victim and Offense Characteristics, 1994–1998

	Before imputation		After imputation		
Victim/offender relationship	Number of cases	% of cases	Number of cases	% of cases	Difference (%)
Intimate partner	115	4.2	121	3.6	−0.6
Primary blood	87	3.2	92	2.7	−0.5
Primary other	25	0.9	25	0.7	−0.2
Secondary	562	20.5	681	20.2	−0.3
Crime-related	842	30.8	1094	32.4	+ 1.6
Stranger	1105	40.4	1367	40.4	0
Total	2736	100	3380	100	0

Table VII. Distribution of Homicide Victim/Offender Relationships Using an Alternative Classification for Chicago Before and After Missing Value Imputation Using Victim and Offense Characteristics, 1991–1995

	Before imputation		After imputation		
Victim/offender relationship	Number of cases	% of cases	Number of cases	% of cases	Difference (%)
Intimate partner	309	9.6	375	8.4	−1.2
Primary blood	178	5.5	203	4.6	−0.9
Primary other	244	7.6	286	6.4	−1.2
Secondary	850	26.3	1220	27.4	+1.1
Crime-related	1139	35.3	1615	36.2	+0.9
Stranger	505	15.7	760	17.0	+1.3
Total	3225	100	4459	100	0

likely a strong predictor of missing data on victim/offender relationships: clearance status. The reason for its exclusion thus far concerns its unavailability in the Los Angeles data set. However, there is information for Chicago homicide cases concerning whether or not the offense has been cleared. Thus we decided to run one last set of imputations for Chicago, this time adding a dichotomous variable for clearance status (with cleared coded as "1" and uncleared coded as "2"). Since the results were very similar for models with and without offender-related variables, only the latter are shown (Table VIII).

The inclusion of clearance status in the model has a discernible impact on the results. In particular, there is a notable increase in the percentage of cases involving strangers from 15.7 to 21.2% after imputation, and from 16.7 to 21.2% between the model without and with the clearance status variable. Thus, the addition of clearance status to the model leads to a 27% change in the proportion of stranger homicides, which seems substantially significant. The percentage of cases in the remaining categories all drop once the unknown cases have been assigned, with the biggest drop occurring in the friend/acquaintance category (3.1%).

Calculating percentages within categories of victim/offender relationships provides an alternative view. Doing so indicates that while approximately 84% of both intimate partner and other family homicides are classified as such by the police (thus not requiring an imputed value) and would be present in analyses which included only cases with complete information, only 53.5% of stranger homicide cases and 75.8% of friend/acquaintance homicide cases contain information about the victim/offender relationship without imputation. Clearly, then, those analyses which drop cases for which offender-related variables such as victim/offender relation-

Table VIII. Distribution of Homicide Victim/Offender Relationships for Chicago Before and After Missing Value Imputation Using Victim and Offense Characteristics and Clearance Status, 1991–1995

	Before imputation		After imputation		
Victim/offender relationship	Number of cases	% of cases	Number of cases	% of cases	Difference (%)
Intimate partner	309	9.6	367	8.2	−1.4
Other family	239	7.4	283	6.4	−1.0
Friend/acquaintance	2172	67.3	2865	64.2	−3.1
Stranger	505	15.7	944	21.2	+5.5
Total	3225	100	4459	100.1	0

ship are missing are losing more stranger and friend/acquaintance cases than other types, therefore biasing the results.

These new results for Chicago now reveal a degree of intercity variability with respect to imputations. In Chicago, the changes ranged from a 3.1% decline in friends/acquaintance homicides to a 5.5% increase in stranger homicides. By contrast, in Los Angeles, the largest change was a 0.7% increase in stranger homicides. In short, victim/offender relationships in Los Angeles distributed themselves after taking account of missing values in about the same percentages as was found before imputation. This is what Decker (1993) found in his study of St. Louis homicides. The same cannot be said for homicides in Chicago.

What accounts for the difference in the results for the two cities? There are several factors to consider. First, there are a larger number of missing values in the Chicago data in comparison to Los Angeles. In Chicago, 27.7% of cases are missing victim/offender relationship while in Los Angeles, 19.0% of cases are missing victim/offender relationship. Also, a substantially higher percent of stranger homicides are identified and reported in Los Angeles (40.4%) than in Chicago (15.7%).

Missing values in Chicago more frequently indicate uncleared homicides than in Los Angeles. The fact that offender information is available in Los Angeles does not mean an offender has been arrested; it means police officers have identified suspects, but have not necessarily taken the legal step of arrest. Indeed clearance percentages in Los Angeles for 1994 through 1998 are substantially lower (58.2%) than in Chicago for 1991 through 1995 (71.8%). Without the minimum legal standards that are required of an arrest, we are left with police officer judgement as a criteria for the validity and reliability of victim/offender classification. If we assume that police are entering victim/offender relationships on the basis of what they believe is the case, rather than establishing it after arresting an offender, there is a less rigorous selection factor. If that is the case, then imputation will make less of a difference in the resulting classification.

Second, one of the better predictors of missing values in the Chicago model was clearance status. Clearances were not used in the Los Angeles models for two reasons: they are not linked to individual cases as is true in Chicago and they are aggregated on a monthly rather than a victim basis. The absence of such a variable from the Los Angeles model may produce a less than adequate imputation model.

6. CONCLUSIONS

There are a number of conclusions to be drawn from this study. First, this study confirms once more that unknown victim/offender relationships are not composed primarily of homicides involving strangers. This research suggests that while imputed values on the victim/offender relationship variables comprise an increase in the *number* of intimate partner, other family, and friends/acquaintance homicides cases, *proportionately*, the notable increase is among stranger homicides.

Second, without the introduction of clearances as a predictor variable our results agree with Decker's (1993) view that unknown cases are distributed in the same fashion as known cases. As noted, with clearances as a predictor variable, stranger homicides increased by 5.5%.

Unlike Decker's study of the city of St. Louis and the present study of Chicago and LA, Messner *et al.* (2002) and Pampel and Williams' (2000) research relied on aggregated SHR data. Although this makes comparisons difficult, we generally find a smaller increase in the percentage of stranger homicides than has been the case to date, and a much smaller change in the percentage of intimate partner and other family homicides than do Pampel and Williams.

In short, our results suggest that the two existing diametrically opposed claims about missing values are both overdrawn. Our findings do not support the argument that missing data resulting from offenders not being arrested makes very little difference. They also do not support the view that there are substantial numbers of stranger homicides represented by missing values. As is so often the case, the reality is likely to fall somewhere in between. The current analyses indicate that the "unknown" category contains intimate partner, other family, friend/acquaintance, and stranger homicides. However, proportionately more stranger homicides are classified as "unknown" than the other three categories.

Third, the results for Chicago raise an important issue. It appears the assignment of cases with unknown victim/offender relationships to known categories on the basis of missing value imputation is influenced by the types of variables available to be used as predictors; the availability of a clearance status variable being particularly important in this instance. A tremendous amount of homicide research involves secondary data analysis where there is little or no information about clearances or other predictors that may be important. Thus, it is possible that in some cases imputation may lead to assigned values that suggest that missing values have no effect when the result is due to the absence of significant predictors. In this regard we advocate examining the sensitivity of imputations to differing sets of predictor variables. In certain situations, listwise deletion may actually prove to be the best method for handling missing data (see Allison, 2002, for a detailed discussion of this issue).

6.1. What Can Be Done?

There are two approaches that can be used. First, it is important for researchers to determine the pattern of missing values. Imputations aside, analyzing the pattern of missing values should encourage researchers to temper their conclusions with caution. Second, given the frequency with which criminologists use listwise deletion, a better understanding of missing data models may be useful. We offer the logic of the present study as a means of better understanding the approach.

How missing data should be handled depends to a great extent on whether the data are missing completely at random, missing at random, or nonignorable. It is therefore surprising how little attention has been paid to this issue, even in the few studies which seek to address missing data among homicide cases.[9] It should come as no surprise that information about homicides such as victim, offender, and offense characteristics are not missing completely at random. Yet the tendency of researchers to deal with missing data through the use of listwise or pairwise deletion, as is the common practice these days, seriously brings into question the findings of analyses based on what are almost certainly not a random subset of the full range of cases. While that might be considered the bad news, the good news is that it is likely the data are missing at random, providing access to a wide range of imputation methods which would not be available if the data were nonignorable. The EM algorithm is one such imputation method. At minimum, researchers should analyze their data both without and with missing values imputed to determine the robustness of their findings.

[9]One of the few exceptions is the work by Messner *et al.* (2002).

ACKNOWLEDGMENTS

We gratefully acknowledge the helpful comments and suggestions of the Editor and reviewers of *JQC*.

REFERENCES

Acock, A. C. (1997). Working with missing data. *Fam. Sci. Rev.* 10: 76–102.

Allison, P. (2000). Multiple imputation for missing data: A cautionary tale. *Sociol. Methods Res.* 28: 301–309.

Allison, P. (2002). *Missing Data*, Sage, Thousand Oaks, CA.

Best, J. (1988). Missing children: Misleading statistics. *Public Interest* 92: 84–92.

Best, J. (1999). *Random Violence: How We Talk About New Crimes and New Victims*, University of California Press, Berkeley.

Biderman, A. D., and Lynch, J. P. (1991). *Understanding Crime Incidence Statistics: Why the UCR Diverges from the NCS*, Springer-Verlag, New York.

Block, C. R., and Block, R. (1997). *Homicides in Chicago, 1965–1995*. Vol. ICPSR 6399. Inter-University Consortium for Political and Social Research, Ann Arbor.

Blumstein, A. (1995). Youth violence, guns, and the illicit drug industry. *J. Crim. Law Criminol.* 86: 10–36.

Browne, A., and Williams, K. R. (1989). Exploring the effect of resource availability and the likelihood of female-perpetrated homicides. *Law Soc. Rev.* 23: 75–94.

Browne, A., and Williams, K. R. (1993). Gender, intimacy, and lethal violence. *Gend. Soc.* 7: 78–98.

Browne, A., Williams, K. R., and Dutton, D. G. (1999). Homicide between intimate partners: A 20-year review. In Smith, M. D. and Zahn, M. A. (eds.), *Homicide: A Sourcebook of Social Research*, Sage, Thousand Oaks, CA, pp. 149–164.

Cardarelli, A., and Cavanagh, D. (1992). Uncleared homicides in the United States: An exploratory study of trends and patterns. Paper presented at the annual meeting of the American Society of Criminology, New Orleans.

Conklin, J. E. (1975). *The Impact of Crime*, Macmillan, New York.

Cox, B. G., and Folsom, R. E. (1978). An empirical investigation of alternative item nonresponse adjustments. *Am. Statistical Association. Proceedings of the Section on Survey Research Methods* 219–221.

Daly, M., and Wilson, M. (1988). *Homicide*, Aldine de Gruyter, New York.

David, M., Little, R. J. A., Samuhel, M. E., and Triest, R. K. (1986). Alternative methods of CPS income imputation. *J. Am. Stat. Assoc.* 81: 29–41.

Decker, S. H. (1993). Exploring victim-offender relationships in homicide: The role of individual and event characteristics. *Justice Q.* 10: 585–612.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood estimation of incomplete data via the EM algorithm. *J. Roy. Stat. Soc. Series B.* 39: 1–38.

Dugan, L., Nagin, D. S., and Rosenfeld, R. (1999). Explaining the decline in intimate partner homicide: The effects of changing domesticity, women's status, and domestic violence resources. *Homicide Studies* 3: 187–214.

Federal Bureau of Investigation (1999). *Crime in the United States: Uniform Crime Reports*, U.S. Government Printing Office, Washington.

Flewelling, R. L., and Williams, K. R. (1999). Categorizing homicides: The use of disaggregated data in homicide research. In Smith, M. D. and Zahn, M. A. (eds.), *Homicide: A Sourcebook of Social Research*, Sage, Thousand Oaks, CA, pp. 96–106.

Gilbert, N. (1991). The phantom epidemic of sexual assault. *The Public Interest* 103: 54–65.
Hair, J. F., Jr., Anderson, R. E., Tatham, R. L., and Black, W. C. (1995). *Multivariate Data Analysis,* Prentice Hall, Englewood Cliffs, N.J.
Heitjan, D. F. (1997). Annotation: What can be done about missing data? Approaches to imputation. *Am. J. Public Health* 87: 548–550.
Hotaling, G. T., and Finkelhor, D. (1990). Estimating the number of stranger-abduction homicides of children: A review of available evidence. *J. Crim. Justice* 18: 385–399.
Jenkins, P. (1994). *Using Murder: The Social Construction of Serial Homicide*, Aldine de Gruyter, New York.
King, G., Honaker, J., Joseph, A., and Scheve, K. (2001). Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *Am. Polit. Sci. Rev.* 95: 49–69.
Little, R. J. A., and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*, John Wiley, New York.
Little, R. J. A., and Rubin, D. B. (1989). The analysis of social science data with missing values. *Sociol. Methods Res.* 18: 292–326.
Madow, W. G., Olkin, I., and Rubin, D. B. (eds.) (1983). *Incomplete Data in Sample Surveys*, 3 vols., Academic Press, New York.
Maxfield, M. (1989). Circumstances in supplementary homicide reports: Variety and validity. *Criminology* 27: 671–695.
McIntyre, J. (1967). Public attitudes toward crime and law enforcement. *Annals of the Am. Acad. Polit. Soc. Sci.* 374: 34–46.
Messner, S. F., Deane, G., and Beaulieu, M. (2002). A log-multiplicative association model for allocating homicides with unknown victim-offender relationships. *Criminology* 40: 457–480.
Pampel, F. C., and Williams, K. R. (2000). Intimacy and homicide: Compensating for missing data in the SHR. *Criminology* 38(2): 661–680.
Petee, T. A., Weaver, G. S., Corzine, J., Huff-Corzine, L., and Wittekind, J. (2001). Victim-offender relationships and the situational context of homicide. In Blackman, P. H., Leggett, V. L., and Jarvis, J. P. (eds.), *The Diversity of Homicide: Proceedings of the 2000 Annual Meeting of the Homicide Research Working Group*, Federal Bureau of Investigation, Washington, DC.
Regoeczi, W. C., Kennedy, L. W., and Silverman, R. A. (2000). Uncleared homicides: A Canada/United States comparison. *Homicide Studies* 4: 135–161.
Regoeczi, W. C., and Riedel, M. (1999). Criminal homicides and missing data. Paper presented at the annual meeting of the American Society of Criminology, Toronto.
Regoeczi, W. C., and Riedel, M. (2000). Estimating stranger homicides: Los Angeles and Chicago. Paper presented at the annual meeting of the American Society of Criminology, San Francisco.
Reuter, P. (1984). The (continued) vitality of mythical numbers. *The Public Interest* 75: 135–147.
Riedel, M. (1987). Stranger violence: Perspectives, issues, and problems. *J. Crim. Law Criminol.* 78: 223–258.
Riedel, M. (1993). *Stranger Violence: A Theoretical Inquiry*, Garland Publishing Co., New York.
Riedel, M. (1998). Counting stranger homicides: A case study of statistical prestidigitation. *Homicide Studies* 2: 206–219.
Riedel, M. (1999). Sources of homicide data: A review and comparison. In Smith, M. D. and Zahn, M. A. (eds.), *Homicide: A Sourcebook of Social Research, Sage*, Thousand Oaks, CA, pp. 75–95.
Riedel, M. (2000). *Research Strategies for Secondary Data: A Perspective for Criminology and Criminal Justice*, Sage, Thousand Oaks, CA.

Riedel, M., and Rinehart, T. A. (1996). Murder clearances and missing data. *J. Crime and Justice* 19: 83–102.

Rinehart, T. A. (1994). *An Analysis of Murder Clearances in Chicago: 1981–1991*. Unpublished master's thesis. Southern Illinois University, Carbondale.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* 63: 581–592.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*, Chapman & Hall, Boca Raton.

Silberman, C. E. (1978). *Criminal Violence, Criminal Justice*, Random House, New York.

Silverman, R. A., and Kennedy, L. W. (1997). Uncleared homicides in Canada and the United States. In Riedel, M., and Boulahanis, J. (eds.), *Lethal Violence: Proceedings of the 1995 Meeting of the Homicide Research Working Group*, U.S. Government Printing Office, Washington.

Spector, M., and Kitsuse, J. I. (1987). *Constructing Social Problems*, Aldine de Gruyter, New York.

SPSS Inc. (1997). *Missing Value Analysis 7.5*. Chicago.

Wellford, C., and Cronin, J. (1999). An analysis of variables affecting the clearance of homicides: A multistate study. Justice Research and Statistics Association, Washington, DC.

Wiersema, B., Loftin, C., and McDowall, D. (2000). A comparison of Supplementary Homicide Reports and national vital statistics system homicides estimates for U.S. counties. *Homicide Studies* 4: 317–340.

Williams, K., and Flewelling, R. L. (1987). Family, acquaintance, and stranger homicide: Alternative procedures for rate calculations. *Criminology* 25: 543–560.

Williams, K., and Flewelling, R. L. (1988). The social production of criminal homicide: A comparative study of disaggregated rates in American cities. *Am. Sociol. Rev.* 53: 421–431.

Williams, K. R., and Pampel, F. (1998). Intimacy and homicide: Compensating for missing information in the Supplementary Homicide Reports. Paper presented at annual meeting of the American Society of Criminology, Washington, DC.