

## THE MATCHING OF CENSUS AND PROBATION DEPARTMENT RECORDS SYSTEMS\*

Jon E. Simpson and Maurice D. Van Arsdol, Jr.

University of Southern California

Records linkage as a new methodological technique has generally been neglected by social scientists. Also, with a few notable exceptions, little attention has been given to the linkage of Federal and community data sources on an individual case basis to maximize the efficiency of data collection and analysis.<sup>1</sup>

The present research, known as the Southern California Records Matching Project, is concerned with the matching of two disparate systems - a metropolitan probation department and the United States Bureau of the Census - which are not designed to obtain comparable information, and for which there has been little a priori concern with possible records linkage. By merging the two systems, it is possible to obtain a better quality of data for delinquents on a broader range of variables than is usually provided by local data systems. Furthermore, in the past, delinquency rate analyses have been inhibited by differences in the recording of information between adjudicating agencies and the Bureau of the Census, the usual general population data source on which rates are constructed.<sup>2</sup> More explicitly, police and probation departments provide information for the numerators of delinquency rates (usually consisting of population and housing characteristics of the delinquents and their families), while the U.S. Censuses of Population and Housing furnish data for delinquency rate denominators (i.e., information on the presumed corresponding characteristics of population aggregates). If, as is often the case, census data pertain to populations or variables that differ in an unknown way from those defined by delinquency adjudicating agencies, the data may not be comparable and the rates will be in error. Subsequent analyses of delinquent and nondelinquent populations may be invalid.

An alternative to correct for these limitations is records linkage. Direct delinquency rates and comparisons can be derived by matching, on an individual case basis, the probation department records for juveniles (which form the basis of the rate numerators) with comparable individual 1960 census returns (which provide

data for rate denominators) for these juveniles and their households. With this technique, numerators and denominators of delinquency rates as well as comparisons between subsamples of the delinquent and general populations refer to the same phenomena - census variables for which comparable data are available to pursue theory building endeavors, to enhance flexible statistical analysis, as well as to evaluate some of the contemporary conceptions of delinquency. In addition to internal comparisons involving subcategories of the delinquent population, the delinquents and/or their households can be examined in terms of: (1) the total population delimited by the same gross criteria which pertain to the delinquent, e.g., in this study, only those households in which there are one or more children between the ages of 10 and 17; (2) the "nondelinquent" siblings of the delinquent; and (3) computer "matched" nondelinquent households within the same or other communities selected on the basis of the presence of a youth with comparable characteristics. Finally, it is possible to analyze the type and incidence of delinquency from one neighborhood (enumeration districts) to the next, contrasting similar as well as dissimilar areas.

This report discusses both the techniques and results of the records matching approach as applied in the present study. Also consideration is given to the attendant problems which have special implications for the tabulation program.

The Records Matching Procedure

In order to obtain a universe of "delinquent" cases, summary information on case identification, name, age, sex, race, and offense was collected for the 23,543 juvenile cases referred to the Probation Department from July 1, 1959 through December 31, 1960 - an 18 month period centered on April 1, 1960. The study population was limited to 13,351 cases comprised of juveniles age 10 to 17 inclusive, who were defined as official cases through the filing of formal juvenile court petitions, as a result of alleged delinquent acts.<sup>3</sup>

Procedures were instituted for the collection of intake data from the case

folders, and an editing system was devised to furnish standardized identification information for the Bureau of the Census. Thus, information on age, sex, race, and offense was coded, a specification was made of the address of the adult members of the household with whom the juvenile would most likely be matched, and the cases were allocated to census tracts. To achieve maximum efficiency in retrieving the required information from the case folders, only the intake "face sheets" or "book sheets" were examined.

Work by the Bureau of the Census. Rules had been previously established for determining whether a match was obtained between the Probation Department and Bureau of the Census records. The key criteria for establishing match status were relationship of juvenile to head, and age, sex, and race of juvenile. Actually two different matching procedures were implemented by the Bureau of the Census.

In the "feasibility" phase, designed to evaluate the practicality of merging the two records systems, the face sheets for a systematic random sample of 2,316 cases comprising one-sixth of the total projected study sample were matched by hand (visually) with the enumeration schedules in storage at Jeffersonville, Indiana. The Bureau allocated the tracted cases to enumeration districts and searched the district files for the records of the juveniles and/or adults in the juvenile's household. Each match failure on the first (primary) address was reallocated to new enumeration districts and/or a search and reallocation of second and third addresses, when available, was completed. Institutional lists were reviewed for cases not located in housing units. The cases for which a complete match - juvenile and head of household - was not effected were returned to the project staff for estimates of new addresses which could be rematched with the census returns. Although the matching rates for this stage will be discussed in the following section on results, it should be noted that the manual procedures located both the juvenile and adult in 84.0 percent of the cases.

Subsequent to the feasibility study and before arrangements could be completed to forward the remaining five samples of the study universe for the matching cycle, the enumeration schedules were destroyed. The relevant matching information was transferred to microfilm reels which could not be utilized without prohibitive increases in financial and time commitments. An

alternative method combining the Census listing books and the computer tapes which contain the 25 percent sample information proved almost as effective as the first method with a substantial reduction in cost.<sup>4</sup> The face sheets were matched with the listing books which provide for address, sample status, surname of the head of household, number of household persons, Fosdic page number, ED (enumeration district) number and administrative material. For the cases in the 25 percent sample, the ED and Fosdic page designations as well as the information pertaining to relationship to head of household, age, race, and sex of the juvenile from the face sheets were punched on cards. The cards were matched with the 25 percent sample tapes for Los Angeles County which were specifically constructed for this project from the Los Angeles City and California files. The match failures were processed by techniques similar to those described for the manual method with one additional step. The 25 percent sample microfilm records were used as a final verification source for unmatched and "marginal" cases.

As each case was located, all the population and housing characteristics of the household were added to a tape file for the delinquent population. The general population tape file, which contains the data for the delinquency rate denominators and which is comprised of all families and housing units with one or more children 10 to 17 years of age, was derived from the Los Angeles County tape file constructed for matching purposes.

Finally, comparative information on the relative effectiveness of the two methods suggests that the computer alternative produces substantially similar matching rates and allows the same degree of confidence in the accuracy of the match as the visual procedures. The determining advantages in the choice of alternatives are the lower costs and time requirements to match the 25 percent sample cases by the computer method as contrasted with the availability of complete count data in the microfilm record or enumeration schedules for matching all cases.

Matching results. Table I summarizes the Bureau of the Census search results for the two basic matching operations which have been described in this report.<sup>5</sup> The outcome of the more recent and complete study universe search is reported under the headings: "Listing Book Search" and "Computer Matching Rates." The "Visual Matching Rates" section presents the findings

for the earlier feasibility study.

the 2,919 sample cases are shown in

TABLE I  
SUMMARY OF BUREAU OF CENSUS SEARCH OF LOS ANGELES COUNTY PROBATION DEPARTMENT  
JUVENILE CASES

Outcome	Number of Cases	Percent
Listing Book Search		
(1) Total Cases	13,351	100.0
(2) Total Cases Within Scope	12,597	94.4
(3) Case out of 25 Percent Sample	9,678	76.8
(4) Case in 25 Percent Sample	2,919	23.2
(5) Total Cases out of Scope	754	5.6
Case in Group Quarters	321	42.6
Case out of Los Angeles County	364	48.3
Duplicate	22	2.9
Status Undetermined	47	6.2
Computer Matching Rates		
(6) Total 25 Percent Sample Cases	2,919	100.0
(7) Juvenile and Adult Found	2,270	77.8
(8) Juvenile Found - Adult Not Found	59	2.0
(9) Juvenile Not Found - Adult Found	196	6.7
(10) Juvenile and Adult Not Found	394	13.5
Visual Matching Rates		
(11) Total Cases (Feasibility Study)	2,125	100.0
(12) Juvenile and Adult Found	1,785	84.0
(13) Juvenile Found - Adult Not Found	57	2.7
(14) Juvenile Not Found - Adult Found	80	3.8
(15) Juvenile and Adult Not Found	203	9.5

A total of 13,351 cases (Row 1), which comprise the study universe, were submitted to the Bureau of the Census. As may be seen in Rows 2-4, a total of 12,597 cases were within the scope of the study design, of which 23.2 percent had been allocated to the 25 percent sample enumeration. The listing book and collateral search techniques established that 754 juveniles (Row 5) were out of scope because of residence in group quarters or out of Los Angeles County, duplication, or lack of information.

The computer matching results for

Rows 6-10. Both the juvenile and adult were found in 77.8 percent of the cases; the juvenile, but not the presumed adult, was located in an additional 2.0 percent of the cases; the adult only was found in 6.7 percent of the referrals; and neither the juvenile nor the adult was matched in the remaining 13.5 percent of the cases.<sup>6</sup>

The outcome of the earlier feasibility study to test the practicality of merging Probation Department and Census records is reported in Rows 11-15. The somewhat higher rates are probably a function of the more elabo-

rate efforts to obtain a match and the less conservative requirements for accepting a match. Thus, in order to acquire new addresses for rematching, an extensive field follow up of unmatched cases involving a review of school, public assistance, vital statistics, Youth Authority, and Juvenile Index records, and a detailed study of the Probation Department case folders was completed for the feasibility sample, but only the examination of probation files was undertaken for the unmatched cases of the entire study population. Furthermore, location of the juvenile and appropriate adult in the enumeration schedules did not assure that the 25 percent sample information would be obtained without case attrition as occurred, by definition, with the computer matches.

Matching rates by date of referral and social categories. Although the results for the total study population are not yet available, the findings for the feasibility sample provide data for tentative conclusions. An analysis of the matching percentages and numbers of cases located for all possible combinations of months evenly bracketing the census date (i.e., two to eighteen) indicated no substantial decrease in matching rates as a result of using the full eighteen month period, rather than the usual maximum of two to four months. To illustrate the matching rates for the 18 month range of referrals, the highest percentage of juveniles matched for any single month was 94.3 for December 1959, and the lowest percentage matched, 73.5, occurred for July 1959. However, only three months were characterized by rates less than 80 percent. These figures may be compared with the total matching percentage of 86.7 (Rows 12 and 13) for the entire 18 month period.

The findings by sex, race, and offense classifications suggest that the matched cases are representative of the Probation Department universe of "official" cases. With offense controlled no statistically significant sex or race variations were found in comparisons of the matched and unmatched cases. Although differences by offense were not statistically significant, there was a tendency to achieve more success in locating the juveniles brought to the attention of the court for auto theft, major traffic, and property violations than for those youth processed for sex delinquencies and offenses against the persons (e.g., robbery and forcible rape).

To determine whether the matching criteria were systematically relaxed with

respect to the matched 25 percent sample cases and at the same time to assess the degree of relationship between the Probation Department and the Bureau of the Census recording of key information, the correspondence between the respective entries for relationship to head, sex, race, and age was reviewed. The two systems were in one to one agreement on more than 96 percent of the relationship to head, sex, and race comparisons and for over 92 percent of the juveniles by chronological age  $\pm$  one year to account for the 18 month entry span in the probation referrals bracketing the April 1, 1960 census date.

In summary of the matching results, it is clear that the "delinquent" cases in Los Angeles County can be traced and identified in the returns of the 1960 U.S. Census of Population. This objective can apparently be accomplished for a time span of 18 months without major attrition of cases by month and without eliminating subcategories of the universe based on sex, race, or offense.

#### Implications for the Tabulation Program

Any matching study must consider the possible biases introduced as a consequence of the limitations associated with the data from the systems that are merged and with the matching process. This final section will focus on selected matching problems currently under examination in connection with the tabulation program; a description of which follows for purposes of orientation. The Bureau of the Census has tabulated group data on an individual case basis for over 100 population and housing variables. In addition, two variable cross-tabulations for 27 key variables and specially constructed indices were completed. The tabulation format provides for 36 sort groups and all possible combinations based on 2 family, 2 sex, 3 age, and 3 ethnic categorizations of the delinquent population. The printouts for the delinquent and general populations include rates and reflect the weighting program used in preparing the 1960 Census sample results.

Underenumeration and sampling errors may be introduced by both the Probation and Census records systems. Certainly, the probation cases are not representative of all youths who have committed delinquent acts or even the "official" delinquents known to adjudicating agencies. However, the total universe of court referrals meeting the project specifications was included and less than one percent was deleted due to administrative difficulties. The underenumeration of census returns for

the Probation Department base population can only be estimated with considerable subjective judgment, since there are no survey results that are directly applicable to this specific base population. Nevertheless, an evaluation of this issue resulted in gross underenumeration estimates of approximately 4 and 6 percent as the low and high parameters, respectively. As noted earlier, 23.2 percent of the "in scope" cases were also within the 25 percent sample. At present, estimates have not been developed to indicate the extent to which underenumeration and sampling errors effected the ratio actually obtained.

Match failures and mismatches represent the two basic sources of potential bias resulting from the search process. Since our investigation was not initiated until three years after the 1960 census enumeration, special procedures, such as a sample survey, to acquire information that would permit an analysis of the characteristics of the unmatched cases were not possible. The representativeness of the matched cases can only be tested on the basis of Probation Department data on age, sex, race, date of referral, and offense, as described for the feasibility study. The findings reported on the close correspondence of the Probation and Census entries for the matched cases lends support to the assumption that few cases were mismatched.

In conclusion, it must be noted that the matching approach facilitates the analysis of possible sources of bias to a degree not possible, or usually not explored, with other types of data collection methods. Records linkage is a technique that, with continued refinement, should stimulate the type of interest that has been manifested in the procedural matters pertaining to censuses, surveys, and sampling. Although its utility is being explored with reference to juvenile delinquency in this project, applications of the method should be consistent with a wide variety of substantive interests.

#### Footnotes

1. Illustrative exceptions include: David L. Kaplan, Elizabeth Parkhurst, and Pascal K. Whelpton, "The Comparability of Reports on Occupation from Vital Records and the 1950 Census," Vital Statistics Special Reports, LIII (June 1961); Lillian Guralnick and Charles B. Nam, "Census NOVS Study of Death Certificates Matched to Census Records," The Milbank Memorial Fund

Quarterly, XXXVII (April 1951), pp. 144-153; Monroe G. Sirken, "Research Uses of Vital Records in Vital Statistics Survey," The Milbank Memorial Fund Quarterly, XLI (July 1963), pp. 309-316; Evelyn M. Kitagawa and Philip M. Hauser, "Methods Used in a Current Study of Social and Economic Differentials in Mortality," in Emerging Techniques in Population Research (New York: Milbank Memorial Fund, 1963), pp. 250-266; and the NIMH Matching Study under the direction of Earl S. Pollack, which is matching census records with mental patients' records to obtain tabulations of patients' characteristics by family characteristics.

2. See, for example, the studies summarized by: Terrence Morris, The Criminal Area (London: Kegan Paul, 1958); Bernard Lander, Toward an Understanding of Juvenile Delinquency (New York: Columbia University Press, 1954); Henry D. McKay, Rate of Delinquents by Communities in Chicago, 1953-1957 (Chicago: Institute for Juvenile Research, 1959), mimeographed; Clifford R. Shaw, et al, Delinquency Areas (Chicago: The University of Chicago Press, 1929); and Clifford R. Shaw and Henry D. McKay, Juvenile Delinquency and Urban Areas (Chicago: The University of Chicago Press, 1942).

3. A total of 9,378 cases were eliminated as a consequence of the following criteria which were implemented in the order listed: under age 10 (3,427 cases - primarily unofficial and/or dependency cases), unofficial (4,089 cases - petition not filed, closed at intake or non-court investigation), dependency (1,842 cases - guardianship, parental neglect, transient, and court consents), and over 18 years of age (20 cases). The final universe of 13,351 cases reflects the additional loss of 814 cases because of duplication or residence outside of Los Angeles County at the time of census enumeration based on the records reviewed by the Population Laboratory.

Males represented about four-fifths of the universe. Anglo-Americans constitute approximately two-thirds of the cases, and Negroes and Spanish surname cases each comprised about one-sixth of the sample. Individual offense categories accounted for 8.5 to 26.0 percent of the total cases with the following percentage distribution by offense type: auto theft (15.2), offense against property (26.0), offenses against persons (8.5), sex delinquencies (9.9), major traffic offenses (14.1), narcotics (4.1), and delinquent tendencies (22.2). Neglect and other dependency referrals and minor traffic

violations were previously eliminated from the universe.

4. The 25 percent sample is the most effective source of information from the point of view of the study objectives since complete count data is available for only a few variables.

5. Results described in this report relating to the comparison of Bureau of the Census records with other records are based on work done by the Bureau of the Census and transmitted to the authors in a manner such that the identification of individuals was not possible and the confidentiality of reports to the Bureau of the Census was maintained.

6. The data tabulations which do not involve the youth characteristics include the 6.7 percent of the cases (row 9) in which the adult only was found as well as the 77.7 percent of the cases (row 7) in which both the juvenile and adult were found for a total of 84.5 percent. The former cases were retained for the analyses in view of the very definite evidence that these households were correctly matched even though the youth was not enumerated.

7. The elaborate and detailed tabulations are deemed appropriate in view of the exploratory objectives of the study and the time delays that would occur with a series of runs by the Bureau of the Census.

\*This investigation was supported in whole by Public Health Service Research Grant 06729 from the National Institute of Mental Health. The data are drawn from the Southern California Records Matching Project. Agencies conducting the research are the University of Southern California Department of Sociology and Anthropology Population Research Laboratory and the Youth Studies Center. Patricia A. Nelson, Muriel Schad, and James Taggart assumed the primary responsibility for the collection of data from the Probation Department files. John C. Beresford, Staff Assistant, Population Division, United States Bureau of the Census has coordinated the work of the Bureau of the Census in connection with this work. Dr. Georges Sabagh participated in the formulation of the original project design and assisted in the development of the tabulation program. The authors are indebted to Dr. Calvin Goldscheider for a critical reading of this paper.