Chapter 9

Using Microsoft Access to Perform Exact Record Linkages

Scott Meyer, Statistics Canada

Abstract

The author describes how using Microsoft Access to perform record linkage may be a viable alternative to specially designed record linkage software for certain applications. Access was pursued since it is fairly easy to use, flexible, interactive, and reasonably fast when performing simple queries. The major drawbacks and minor difficulties will also be discussed. Examples will be drawn from a project which involved linking court records to police records for selected Canadian cities.

A description of the project to link the data from the court and police surveys will be given. The motivation for beginning the linkage and the long term goals will be discussed. The history of the project will be briefly reviewed. The author will then focus on Access, and how it is considered to be an effective alternative to methods previously used for this project. The advantages and disadvantages will be presented.

The strengths of Access include: flexibility -- the criteria which must be met for the records to be considered matches are fully controlled and easily altered by the user, plus it is simple to select subsets of large files which can then be easily explored; availability -- Access, being part of Microsoft Office Suite, is available to many users; speed -- for our application the queries took very little time to run, making the session highly interactive; ease of use -- Access is easy to learn, and even fairly complicated queries can be done with only "point and click" actions with no knowledge of how to program in SQL required.

The primary disadvantage is that there is no probabilistic matching based on the theory of Fellegi and Sunter (1969). This is a significant drawback; however, for many applications, exact matching is enough to meet the project's goals.

Lastly, some results of linking court and police records using Access will be presented.

Introduction

The goal of this project is to combine court data from the Adult Criminal Court Survey (ACCS) which collects provincial court data and the Revised Uniform Crime Reporting Survey (UCR2) which collects police data on criminal incidents. The populations of the two surveys overlap a great deal. By linking the two files, we can map an offender's movement through the criminal justice system from the point of arrest to the point of sentencing. The ACCS provides data about the decision (guilty or not guilty) plus full sentencing information. The UCR2 survey provides details about the criminal incident, the accused, and, for violent offences, the victim(s). It is anticipated that linked data will provide answers to some interesting questions asked in the justice community. For example, is there a difference in the types and lengths of sentences for accused charged in spousal versus non-spousal assaults? Does the location of a break and enter or act of vandalism affect the severity of the sentence?

This report will show that using Access has the potential to be an effective method to combine data. A detailed explanation of how the linked data sets were created using Access queries is not included here, instead linkage results and discussion of some of the advantages and disadvantages of using Access are presented. The first two sections provide a description of the preprocessing of the survey data. This is followed by some statistics obtained for the city of Regina and some explanation of possible reasons for nonmatches in this study. The disadvantages and advantages of using Access are then presented, and a short summary and some conclusions are given in the final section.

Data Sources

Police and court data from the city of Regina was used in this study. Specifically, ACCS charges which had a date of offence between July 1, 1993 and December 31, 1993, were loaded into an Access table. Similarly, UCR2 records from the Regina police department which had a report date in the same six months were loaded into an Access table. Most of these police records also had a date of offence between July 1, 1993 and December 31, 1993, but there were incidents where the date of offence was many years prior. Although these UCR2 records will likely not match, they do not distort the linkage rate since this calculation is based on the percentage of ACCS records for which a match to a police record is found. Regina was chosen since the coverage of the two surveys currently overlaps only in Quebec and Saskatchewan. All previous record linkage studies have been done using only Quebec data.

Preprocessing of the ACCS and UCR2 Data

Before the linkage could be performed, some preprocessing of both raw data files was done. The goal was that each charge on the court file would link to its corresponding violation on the police file. This is a significant change from prior linkage attempts where many charges were linked internally or "rolled-up" into one ACCS record (Brown, 1995; Cooley, 1996).

The raw ACCS data comes from every courthouse in the city and there is one record for every appearance. Since the unit of count used in ACCS published tables is the charge, a program was available which converted the raw data into its one record per charge equivalent. This file with one record per charge was loaded into Access.

The raw UCR2 data is reported by the municipal police department of each city and consists of three files, the Incident, Accused, and Victim files. All three files contain two variables which uniquely identify any incident, the *respondent code* and *incident file number*. These codes allow the files to be easily joined. The Incident file contains information about the criminal incident including the location, date and time of offence, value of property stolen, etc. The Accused file contains a record for each accused that has been identified. Variables such as date of birth, sex, and ethnic status (aboriginal or non-aboriginal) appear on the Accused record. Similarly, the Victim file contains information about each victim of a violent offence. Variables appearing on the Victim file include level of injury, age, and relationship to the accused. An incident may involve more than one violation (up to four are captured on a single UCR2 Incident record). For example, one UCR2 incident could involve both theft and mischief violations. Also, there may be more than one accused involved in a single incident, and for violent offences, a single incident could involve multiple victims.

In order to make the UCR2 data more compatible with the ACCS data structure, a violation file was created. This file had one record for each accused/violation. For example, if an incident involved two accused and three violations, there would be six violation records created, one for every accused/violation combination. This UCR2 preprocessing made the linkage between ACCS charges and UCR2 violations more straightforward than in previous attempts. The UCR2 violation file was loaded into Access along with

the Victim file. The three Access tables (the ACCS charge table, the UCR2 violation table, and the UCR2 victim table) together form an Access database.

In addition to creating ACCS charge and UCR2 violation and victim files, derived fields were added. The most important of these fields was the Common Offence Classification (COC). The COC consists of 28 codes which represent broad categories of crime. The three digit ACCS offence code and the four digit UCR2 violation code were each mapped to their corresponding COC codes. For example, a COC of "1" represents homicide and related offences; a COC of "2" represents attempted murder; and "3" represents robbery. Previously, the offence and violation codes were not used in the linking strategy. By incorporating the COC into the linkage procedure, there is higher confidence that the court record and police record relate to the same event. This minimizes the incidence of the "false-positive" matches encountered in earlier work.

Within Access, the variables available on both files which were used for linkage are Soundex (encrypted name), date of birth, date of offence, sex and COC. The premise of Soundex coding is that names which sound alike (regardless of spelling) are assigned the same code consisting of one letter followed by three numbers. The Soundex codes are created when the survey records are extracted from the local databases, therefore, the full names of the accused are not available from either survey. Records which link exactly on all five of these variables are deemed to represent the same criminal violation and subsequent court charge.

Results for Regina

For Regina, the overall match rate based on an exact match for all five variables was 58%. This is calculated from 3105 of the 5360 charge records from the court data linking to violation records from police data. As a second step, constraints could have been relaxed and another link using only unmatched records from the first step could have been done. However, the goal was not to create one large linked file but rather to produce Access tables which could be used to link records for specific research questions.

The following example illustrates the use of Access to examine one specific problem concerning assaults. There is a potential linking problem because common assault and major assault have different COC codes. Table 1 shows that when using the restriction that the COC codes must agree exactly, 62% of the 442 assault records on the ACCS table were linked. By allowing major assaults to be linked to common assaults, and vice versa, an additional 10% of the records were successfully linked. These are likely to be true matches and the match rate for Regina assaults was increased to 72%. Further steps were then taken to expand the linked file by allowing other constraints placed on the linking variables to vary. For instance, allowing some range for the date of offence, rather than matching exactly, brought together records that, in all likelihood, refer to the same UCR2 violation and ACCS charge.

Table 1 reveals that a match rate of around 85% was achieved while still maintaining high confidence in the quality of the links. Confidence in the quality of the matches declines as more differences among the linking variables are allowed. Judgement of the analyst is important in deciding whether to increase the size of the linked Access table at the risk of allowing "false positive" errors to be made. Table 1 shows the results of following one particular linking strategy for Regina data. Other equally effective strategies may be used, and an analyst is free to reorder the steps, add or omit steps, or decide how much relaxation of matching constraints is appropriate within any particular step (e.g., using 10 days instead of 35 for the date of offence range). The method used to create the analytical Access table will depend on the application and on the input data being used. For example, if some linking variables from certain jurisdictions are known to have data quality problems, then constraints on these variables can be relaxed at an early stage.

Table 1. -- Linkage Rates Using Various Strategies -- Regina Assault Charges (N = 442)

Link #	Soundex	Date of Birth	Date of Offence	COC	Sex	# of new matches	Cumulative # of matches	Cumulative match rate
1	exact	exact	exact	exact	exact	276	276	62%
2	exact	exact	exact	same family[1]	exact	42	318	72%
3	exact	exact	within 35 days	same family	exact	32	350	79%
4	exact	close[2]	exact	same family	exact	11	361	82%
5	exact	close	within 35 days	same family	exact	3	364	82%
6	close[3]	exact	exact	same family	exact	11	375	85%
7	exact	exact	exact	same family	disagree	2	377	85%

[1]Same family = there was no distinction between major and common assaults. Links to UCR2 violations of sexual assault were also permitted, but there were only two matches of this type.
[2]Close for Date of Birth = agreed on two of year, month, and day, or had a month/day reversal.
[3]Close for Soundex = agreed on first letter and first digit of Soundex code.

Possible Reasons for Unlinked ACCS Records

Court and police records may not link for two distinct reasons. There may be no corresponding police record for the charge or there exists a record, but it can not be matched to the charge based on the linking variables and strategy.

There are several possible reasons why no corresponding police record exists. First, there are problems of geographical (jurisdictional) coverage. For example, persons who were charged by the Royal Canadian Mounted Police (RCMP) will have no UCR2 record because the RCMP does not currently report to the UCR2 survey. It is estimated from UCR aggregate data that the proportion of charges laid by the RCMP for Regina is around 5%.

Another aspect of coverage difficulties is charges pertaining to offences which are court related and may not involve the police at all, for example, offences against the administration of justice. Charges for these offences often have no corresponding police record.

There is also the possibility that the police record would be located in another city or province. Since the database includes only the records for one city, the ACCS charge record would remain unmatched. In future, databases which include records from larger areas, such as provinces or regions, could be produced and these would provide the opportunity to link records for an individual who offends in one city and is tried in another.

A fourth reason is that a UCR2 record exists, but due to the restrictions put on the report date when

preparing the Access table, it was excluded. This will be investigated further, and some adjustments to the table preparation method may be required.

The reasons for failing to find true matches when both records exist are harder to describe. Name changes, keying errors on Soundex (incorrect first letter), and missing data are examples of data quality problems on the source files that can result in nonmatches.

Microsoft Access as a Record Linkage Tool

While this report shows that Access can be an effective tool, there are, as with any software, some problems or difficulties. Obstacles and drawbacks encountered in this study will be discussed first, followed by a summary of some advantages of using Access.

These are some difficulties with using Access:

- Although Access does allow for some inexact matches, there is no real probabilistic matching based on weighting. It is possible to use weights when doing exact matching, however, assigning weights in Access is difficult, and this is a major drawback. Probabilistic matching based on the theory of Fellegi and Sunter (1969) is possible with Statistics Canada's GRLS system (Felx, 1995). Further, GRLS and other record linkage software allow the use of sophisticated comparison rules (e.g., string comparator metrics), which would be very difficult, if not impossible, to imitate using Access.

- When the linkage is performed, duplication can occur. If two UCR2 violation records have exactly the same values for all matching variables, they would both be linked to the same charge record. In a sense, the charge record has been duplicated. This duplication is not a problem when simply counting the number of successful matches, or when the UCR2 records are very similar. The difficulty occurs when the UCR2 records differ in important ways. For instance, an analyst is interested in comparing sentencing for break and enters (B & E) committed against businesses to sentencing for break and enters committed against personal homes. If one break and enter charge links to two different UCR2 violations, and one violation is against a business and the other is against a home, then the charge is difficult to classify. Should the sentence length be used in the mean sentence length calculations for business B & E, residential B & E, both, or neither? The analyst must be aware of this possibility, and, when doing analysis, these ambiguous records may have to be excluded or handled in some other fashion.

- Access does have some mathematical functions (sum, average, max/min value, etc.), but to do more sophisticated statistical analysis with the linked data set, it would have to be exported to a statistical software package.

- Though Access is easy to use, careful attention to detail is required. Queries with seemingly small differences can produce vastly different results. Careful design of queries is needed to ensure that the final result is what was intended. Novice users, not knowing what kind of output to expect, may not immediately recognize flawed queries. Also, depending on the linking strategy used, a relatively long sequence of steps may be involved. Though each step is fairly easy to perform, the entire procedure can become quite complicated.

- The study used Access 2.0 for Windows and there are some important technical limitations. The speed, and hence the convenience, of using Access is affected by the power of the PC that it is running on. Some important Access limitations are listed here. The maximum database size is 1 Gigabyte; the maximum number of tables plus queries in the database is 32,768; the maximum number of fields per record/table is 255; the maximum number of tables used in a query is 32; the

maximum number of sorted fields per query is 10. In the Regina study these maximum capabilities were not generally restrictive. One problem encountered was by continually using the output from one query as the input to the next, after several layers of depth, the error message "query is too complex" would appear. This is avoided by saving the output from an intermediate step as a table, then using this newly created table, rather than the output from the query, as the input to subsequent queries. MS Access Version 7.0, which is now available, may have greater capacities. For large applications, MS SQL server could be adopted as the underlying relational database management system, while the user interface would still be MS Access.

These obstacles are not terribly severe. The advantages of Access, which are listed below, outweigh the problems or difficulties.

- The greatest advantage of using Access is the flexibility. As mentioned, the criteria which must be met for the records to be considered matches is fully controlled and easily altered by the analyst.

- Also, there is flexibility when creating the linked analytical file with respect to which variables are included. Since only the selected variables will be written to the linked table, the analyst is able to work with an uncluttered data set. In addition, the analysis of nonmatched records from any Access table is very easy. A simple built-in query wizard will provide the analyst with the unmatched records. Patterns among the unmatched records may be discovered by reviewing them visually or via subsequent queries on the unmatched data set. For instance, match rates may, for some reason, be lower for certain courtrooms within the city. If a situation like this is discovered, it can be further investigated.

- Another asset of Access is its availability. In particular, it is available to analysts in the Canadian Centre for Justice Statistics (CCJS), and generally, it is a component of the ubiquitous MS Office Suite.

- Another benefit of using Access is its speed. How quickly a query runs depends on the computer's hardware, the size of the tables which are being queried, and the complexity of the query. Using a pentium computer, queries on the Regina database took only a few seconds to complete. The result is a highly interactive session where one can quickly learn about the data while creating the linked table.

- Since it runs in the PC environment, Access is inexpensive to use. The only cost is an up-front cost of purchasing the software/software licences.

- Another asset of Access is its ability to use data from and provide data to a number of sources (e.g., spreadsheets, other database software, flat file, etc.). Since the preprocessing was done using SAS, and further analysis requiring sophisticated statistical procedures may be done, it is important that Access be able to import and export the files. Indeed, the import and export capabilities of Access are quite good, thus lending compatibility with other packages.

- Lastly, Access is easy to learn and use and requires no special programming skills to use effectively. Table 1 shows the results of a sequence of queries. This was done to show how relaxing the constraints can increase the number of matches. In practice, the analyst would not usually run several follow-up queries. It is more likely that a single complex query which achieves much the same result would be run. The drawback of a single secondary query which follows the exact match is that for the added records, it is not immediately obvious why they failed to match on the first attempt. For example, records which did not match exactly on the COC and records which did not match exactly on date of offence would be added at the same stage, and it would not be obvious

how many records were in each group. A complicated query which would allow inexact matching on any one of date of birth, date of offence, COC, sex, or Soundex needs to be prepared only once, after which it can be re-used (if some consistent table naming convention is used). In this way, the analysts who are new to using Access and not confident about preparing their own queries can still perform an effective record linkage using these pre-written queries.

In summary, there are both positive and negative aspects to using MS Access to link ACCS and UCR2 records. Weighing these various considerations, Access appears to be a viable and practical way to link records, and meets the goals of this project.

Conclusions

The preliminary work using Access to perform the record linkage is very encouraging. This report focuses on one application, linking adult criminal court records to police records, but Access could also be used for other CCJS record linkage projects. Some possibilities are: youth court to youth corrections (YCS-YCCS), youth court to police (YCS-UCR2), and the marriage of these, police to youth court to youth corrections (UCR2-YCS-YCCS). The match rates achieved for the UCR2-ACCS linkage in Regina were similar to previous studies, but the interpretation of the resulting file is easier. This is an advancement in the record linkage work done in the past three years, since for the first time meaningful analysis of the linked file seems possible.

References

Brown, C. (1995). *Record Linkage Feasibility Study: Uniform Crime Reporting Survey/Adult Criminal Court Survey*, Internal Statistics Canada Report.

Cooley, D. (1996). *Record Linkage Feasibility Study: UCR2/ACCS - Part II*, Internal Statistics Canada Report.

Fellegi, I.P. and Sunter, A.B. (1969). A Theory for Record Linkage, *Journal of the American Statistical Association*, 64, 1183-1210.

Felx, P. (1995). *Feasibility of Using CANLINK for Linkage: An Application in the Canadian Centre for Justice Statistics*, Internal Statistics Canada Report.