

## Article

Symposium 2008:  
Data Collection: Challenges, Achievements and New Directions

### Collecting Crime Data in Canada

by Brad Fetter

2009



## Collecting Crime Data in Canada

Brad Fetter<sup>1</sup>

### Abstract

Justice surveys represent a unique type of survey undertaken by Statistics Canada. While they all essentially use administrative data, Statistics Canada has had considerable input into the type of data that is collected as well as quality assurance methods guiding the collection of this data. This is true in the areas of policing, courts and corrections. The main crime survey, the Uniform Crime Reporting Survey (UCR), is the focus of this paper and was designed to measure the incidence of crime in Canadian society and its characteristics. The data is collected by the policing community in Canada and transmitted electronically to Statistics Canada. This paper will begin by providing an overview of the survey and its distinctive properties, such as the use of intermediaries (software vendors) that convert data from the police's information systems into the UCR survey format, following nationally defined data requirements. This level of consistency is uncommon for an administrative survey and permits a variety of opportunities for improving the overall data quality and capabilities of the survey. Various methods such as quality indicators and feedback reports are used on a regular basis and frequent two-way communication takes place with the respondents to correct existing data problems and to prevent future ones. We will discuss recent improvements to both the data itself and our collection methods that have enhanced the usability of the survey. Finally, future development of the survey will be discussed including some of the challenges that currently exist as well as those still to come.

Key Words: Administrative data, Quality indicators, Crime reporting.

### 1. Introduction

This paper describes the methods and challenges related to the data collection and quality of the Uniform Crime Reporting Survey (UCR). The paper begins with a few key definitions and concepts. The subsequent section presents some of the data collection distinctions and challenges. The paper then describes some of the quality indicators and techniques that are applied to the UCR. Following that, there will be some examples of measurements of crime. The paper concludes with a discussion of the methodological approach to the survey and remarks on some possible issues for consideration.

### 2. What is UCR?

#### 2.1 Definitions

The UCR Survey collects police-reported crime statistics in co-operation with the policing community. Data submissions are sent on a monthly basis from the municipal, provincial and federal police services. The UCR Survey was designed to measure the incidence of crime in Canadian society and its characteristics. The survey is very similar to its counterparts in other countries such as the USA (United States Department of Justice website, 2008).

The target population for the survey consists of any criminal incident that has taken place in Canada and has come to the attention of the police. Although the survey role of the police forces is to collect the data, in terms of the UCR they are considered respondents. This varies from a more typical survey in which the survey population and the respondents are the same entity (person, household, business or organization, etc.). Response to the UCR is difficult to measure due to the fact that the number of criminal incidents reported in a jurisdiction varies randomly

---

<sup>1</sup>Brad Fetter, Household Survey Methods Division, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa Ontario, Canada, K1A 0T6, [Bradley.fetter@statcan.gc.ca](mailto:Bradley.fetter@statcan.gc.ca)

from month to month. It is impossible for Statistics Canada to know if the police forces have in fact sent in all of the relevant data. If Statistics Canada has received any amount data from a police force, they have effectively responded even if the data is incomplete.

- An incident is defined as a set of connected events that are in contravention of the Criminal Code of Canada or other provincial statutes.
- An accused is a person who has been identified as an offender in an incident.
- A victim is a person who is the target of a violent/aggressive act.

Master data refers to a database of all criminal incidents in their current status. This data is stored and maintained at Statistics Canada, although police forces usually possess their own respective databases. The master data at Statistics Canada is constantly changing because of new incidents that are being introduced as well as existing incidents that are being updated. As an example, suppose a criminal incident has been submitted. If the person who committed the crime is found at a later date through an investigation, the database would be updated with this new information.

Raw data refers to data files that are submitted by each respondent on a monthly basis. It serves as the input to the master database.

## **2.2 Aggregate and microdata**

Aggregate data are monthly counts of criminal incidents and accused persons. This is the type of data that has been provided by the police forces in UCR version 1 since its origin in 1962 (Dauvergne, 2008).

In the context of the UCR, microdata refers to data that is collected at the incident level. That is to say that the data records include the characteristics for each individual incident and any associated victims or accused persons where possible. Microdata has been provided from certain respondents since the creation of UCR version 2.0 in 1988. UCR versions 2.0, 2.1 and 2.2 all collect microdata in a relatively compatible format. For UCR version 2.0 up to the current version, data is organized into 3 connected files for incidents, victims, and accused persons. For each incident, there may be any number of accused persons and if there are any violent or aggressive violations in the incident there would be victims.

## **3. Challenges in data collection**

There are several characteristics of the UCR that make it unique among the surveys at Statistics Canada.

### **3.1 Coverage**

First, the survey is technically a census. Statistics Canada will receive a record for every reported criminal incident in the country. It is very rare for the coverage of a survey to be as complete as this. This coverage is exceptionally high with the limitation that it can only collect data on reported crime. Information on non-reported crime is collected through the Victimization cycle of the General Social Survey (GSS) which will be mentioned later in the document.

### **3.2 Response rate**

Second, the UCR response rate is 100%. It is mandatory for police departments to participate in the survey. This is the highest coverage from any survey at Statistics Canada.

The high response rate does however impose an expectation for the data from all police forces to be 100% present before analysis can be done. National and provincial crime statistics can be produced once data from all law

enforcement agencies has been submitted, processed and reviewed. The release and processing of data can be significantly delayed if even one respondent is late with their submissions.

### **3.3 Data collection process**

Third, the collection and data entry are performed by the respondent. Beginning in 1962, the monthly summary counts were originally sent to Statistics Canada in the paper questionnaire format of UCR version 1.0. Now the data is sent electronically as an ASCII text file with a standard record layout. Police departments will generally maintain their data through their own database or software that they have either developed themselves or through a software co-operative vendor. In fact, there are software companies such as Versaterm and Niche that are in the business of making software for police forces, which will have the UCR collection application built right into it.

The microdata acquisition procedure is comprised of the following steps:

- A new criminal incident is entered into a police force's data system, or an existing incident is modified or updated (such as when an accused person has been named after a lengthy police investigation).
- Data is extracted into monthly production files by the police force and sent to Statistics Canada.
- Data records are subject to an edit & imputation program which checks for logic and consistency of field values.
- Data records are integrated into the master data file which is stored at Statistics Canada.
- A production report is produced which includes a list of incidents that contain errors. A report is sent to each police force for feedback and possible rectification of errors.

### **3.4 Survey versions**

Unlike other surveys, the UCR has multiple versions that are being used simultaneously. To date, four different versions of the survey have been developed throughout the years and all are still being used. Each new version of the survey is able to collect more detailed information than its predecessor. For example, the newest version of the survey allows details to be collected on cyber crime for the first time. It would be very much advantageous if all of the police forces in Canada were using the same version of the survey, however the version that is being used by a police force is beyond the control of Statistics Canada. For a new version of the survey to be implemented, a police force would have to upgrade their software and be trained on the new version of the survey. When this could be done would depend on the time and resources that a police department has available. Therefore, Statistics Canada cannot control this.

In order to produce crime statistics at the national level, it is necessary for the data to be in one standard format. In the case of the UCR, the only common way to express the data is to count it at the aggregate offence level. Microdata can be converted to aggregate counts easily enough, but the reverse would be more difficult.

Even though the UCR started collecting microdata from certain respondents in 1988, the main output of the survey is still aggregate offence data. The aggregate reports for the UCR in the United States of America are described as follows: "These reports provide little information about the characteristics of crime, victims, offenders or arrests. Use of these data in research and policy analysis is severely limited." (United States Department of Justice website, 2008).

To produce the national aggregate counts, the following data categories are streamed together:

- Aggregate counts from respondents that are using UCR version 1. This data is already in the aggregate format and therefore does not require conversion.
- Microdata from respondents that are using UCR version 2.0, which are converted to aggregate counts through an automated process.
- Microdata from respondents that are using UCR version 2.1 or 2.2, which are converted to aggregate counts through an automated process that is designed to work for both versions of the survey. UCR 2.2 is the newest version of the survey, and was designed to be as compatible as possible with the previous version.

UCR 2.0 is still active at time of writing because of one last respondent that is still using it. Once this respondent has upgraded, version 2.0 can be retired (UCR 1 respondents are expected to convert to the newest survey version available when they are ready to upgrade). For now, the microdata output of the survey is converted into aggregate offence data using a process called Derived Aggregate Offence (DAO). The output of that process is compatible with the UCR 1 output, and can then be used to create the national counts.

The UCR versions 2.1 and 2.2 also have to go through this process. However, because these two versions were designed to be more compatible, they are able to use the same conversion program to create the offence output.

### **3.4 Survey amendments**

The UCR can be amended due to changes in legislation at any time. Otherwise, amendments can be made to field values, definitions or reporting practices as specified by Statistics Canada with input from the policing community. For example, a new violation had been created in 2005 as a result of a new law dealing with the trafficking of humans. Such a change would require alterations to the reporting manuals and logic edits. This information would need to be passed on to the police forces and to their respective software vendors in order to implement the changes into their respective databases and police reporting software. Ideally, any changes to the survey methodology would be made simultaneously by all parties at the same time since it is an ongoing survey. However, this is not always feasible. The inability to make changes at once could lead to inconsistent or erroneous data if proper measures are not put in place. Those measures will be discussed in the next section.

## **4. Data quality methods and indicators**

### **4.1 Data quality in monthly production**

Data that is supplied by the police forces must be processed by Statistics Canada before adding new data to the master file. During that process, various quality tools are used to verify the data.

#### **4.1.1 Edit & imputation**

As mentioned earlier, the administrative records submitted by the police will be subject to logical, consistency and validity edits. These edits are created with the collaboration of the subject matter experts. This process cleans the data to a certain extent, and also overwrites fields with logical replacements where possible. At the end of the process, a status value (accepted without warnings, accepted with warnings, rejected) is assigned to each record and the data is incorporated into the master files.

Many police forces use software with logical edits built in to the reporting applications. Provided that they are consistent with those at Statistics Canada, these edits can reduce the number of potential errors before the data is sent to Statistics Canada.

#### **4.1.2 Production reports**

One notable distinction of the UCR data collection is that it is an interactive process. Rather than simply receiving the data and working with it as best as possible, Statistics Canada asks the respondents to repair errors in the data. It is uncommon for the respondent to assume responsibility for data quality, yet that is in the nature of the UCR. The main tool for this feedback is the monthly production report that is sent to each respondent following the processing of their data submissions.

The production report is divided up into the following sections:

- Part A: This contains the amount of records submitted, and the amount of records output from the process categorized by the outcome (accepted without errors, rejected, etc.). By checking these record counts,

potential problems can be acknowledged right away. For example, a big drop in the number of incidents submitted in a month may indicate a problem with the reporting system or one of the data files.

- Part B: This displays the year-to-date monthly counts of accepted and rejected records. This list can help to identify significant changes in overall counts and the number of errors and warnings. Although criminal acts are considered to be random events, the volume of incidents is expected to be reasonably similar from one month to the next.
- Part B-2: This is a newer component that notifies the respondent of all data submissions from the past 15 production periods (months) that are in a state of error. This reminds respondents to repair records on the database with critical errors. This supplies the respondent with information needed to correct the data: the incident identification number, the error code, and a brief error message that explains the error itself. In some cases, this message can even show the individual field values that triggered the error.
- Part C: This component provides a list of the errors, imputations and warnings for incidents submitted or updated in the current production month. Like report B-2, this list also provides the respondent with information needed for correction.
- Part D: Finally, this section provides a frequency count of the error types that were found during the current month of production. An error type that is significantly high in frequency could identify a systematic problem that needs to be addressed.

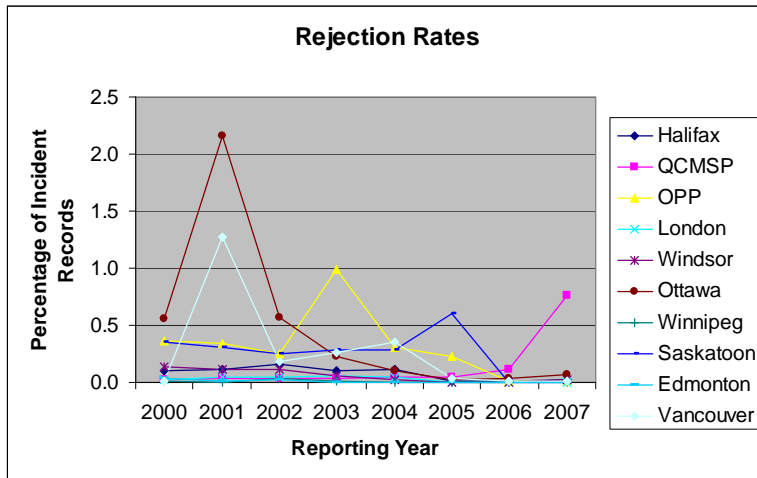
## **4.2 Annual rejection rates**

Rejected records are due to critical errors that are found during the edit & imputation process when there are invalid or illogical data. If the impact is substantial and there is not enough information in the data to derive a logical value for a field, the incident must be rejected. For example, an incident that involves a homicide violation should also include at least one victim record or else the incident is rejected. Many crime studies are done based on data from an entire year's worth of data, or multiple years for time studies. Any incidents that are rejected in production are generally excluded from microdata analysis. The capability of a police force to repair data that has been flagged as erroneous depends on factors such as police resources, the record management system, and the extent of incident information available in their own records.

To monitor the quality of the survey, rejected incident records are calculated as a percentage of overall incidents for a year's worth of data. This rate is usually calculated at the national level and by individual police forces level. By presenting these annual rates on a chart, patterns and potential problems can be identified. For most years, the annual rejection rate is less than 0.5%.

The simple graph shown in Figure 4.2-1 has been used as a data quality tool. When the chart was presented to a respondent with a higher rejection rate, it began to take steps towards improving their data quality by applying for more resources, funding, and a software upgrade. In general, the rejection rates show that respondents are driven to maintain a low rate. Due to the interactiveness of the collection process, the feedback and reports that result from the collection process allow the respondents to improve the quality of their data. The percentage of rejected incidents has generally decreased or remained low over time.

**Figure 4.2-1**  
**UCR Rejection Rates by Year for Selected Respondents**



Warning errors can also occur when there is an inconsistency that is less serious in nature. A field can be replaced by something that is logically derived, or else a general 'unknown' value can be derived. This is still a slight compromise in the data. The percentage of records with a warning status can also be calculated and used for quality improvement purposes. Although the rejection rate is very low (less than 0.5% of the incidents on file), it is also important to keep the warning rate low. That rate represents incidents where data has been subject to imputation, which is say that data fields may have been replaced by logical substitutes. Sometimes that may involve a slight compromise of the data. These compromises are necessary in order to maintain a reasonable level of response burden for the police forces. Past investigations into warning rates have shown that respondents tend to generate more warnings when they upgrade to a newer version of the survey. As the police forces become more familiar with the survey version, their warning rate tends to decrease.

### 4.3 Re-processing master file data

At different points in time, various changes have been made to the survey. For example, the violation (type of crime) coding structure has changed, the edit rules have been altered, and field values have been created or discontinued. These changes have led to inconsistencies in the master files. In order to have master data files that are accurate and reliable, it was necessary to perform a more rigorous data quality action. It would impose too much burden on the police forces to go back historically and modify all of their records, therefore it was necessary to find another alternative.

Normally it is the raw data that is subject to edit and imputation in monthly production. However, by applying the current edit and imputation process to the master file data, many of the historical differences and inconsistencies could be resolved. The program required modifications to avoid excessively rejecting records, but the overall result was that the master data files are much more functional for analysis.

### 4.4 Error correction rates

Although the annual rejection rate is generally very low (section 4.2), there are still opportunities to improve the data quality. Another new data quality initiative is being developed for the UCR. Error correction rates will be created to measure the respondents' success in repairing erroneous incidents in the database. This will be used as another form of quality feedback for the respondents.

The number of corrections will be calculated by determining the number of rejected incidents that have been successfully corrected by updates from the police force over a given period of time. The number of corrections as a

percentage of the total number of rejected incidents for a respondent in a given time period will create the error correction rate.

#### **4.5 Data field value frequency tables**

Not all data quality issues in the UCR involve acceptance or rejection. When a full year of microdata becomes available, another technique is used to investigate the data. Annual data field value frequency tables are created from the master files. These tables examine the percentage distribution of responses by police force for a particular variable in one year. The variation from the national average for each possible field value is calculated and then summed to produce an outlier score. The respondents are then ranked by their outlier score value in order to display the field value distributions that deviate the most from the average.

These tables help to identify differences in reporting practices between police forces. The tables could uncover any biases in the characteristics that are reported in the survey and used for analysis.

### **5. Measuring crime in Canada**

#### **5.1 Crime rates**

The crime rates are well-known products of the UCR. Because crime is a social issue, comparisons of crime between places or years are normally performed on some sort of population basis. The crime rates are expressed as the number of criminal incidents per 100,000 persons in a given time period. To create a crime rate by geographical region (city, province or nation), the number of criminal incidents is divided by the number of people in the area as per the Statistics Canada population census. Statistics Canada publishes the national and provincial crime rates and the crime rates for census metropolitan areas (CMA). Although some municipal police forces produce their own rate for local purposes, all official crime rates in Canada come from Statistics Canada publications.

The crime rate is a useful statistic for many purposes, such as evaluating the effectiveness of crime prevention measures or the relative safety of a particular city or neighbourhood.

While the crime rate is very practical and simple to interpret, it is not without flaws. The main disadvantage is that each type of violation has more or less the same impact on the overall crime rate.

#### **5.2 Crime severity index**

Another new method for measuring crime is through the creation of a crime severity index (CSI). This involves assigning an index weight to each type of crime so that the weighted total number of crimes would better reflect their seriousness or impact. The preliminary methodology for creating the index is to calculate the average sentence given to an offender for each type of crime. The average sentence is calculated based on data from the Integrated Criminal Courts Survey (ICCS), another justice survey at Statistics Canada that uses administrative data and is independent of the UCR. Such an index is in development currently at Statistics Canada and is scheduled for introduction in the summer of 2009.

#### **5.3 Victimization**

The survey of victimisation is conducted as part of the General Social Survey (GSS) at Statistics Canada. This survey identifies victims of crime and their experiences. Unlike the UCR, this survey collects data on both reported and unreported crime. The drawback of this type of crime survey is that only a sample is collected, therefore rare crimes are difficult and sometimes impossible to collect.

One way in which victimisation surveys are useful is that they show some types of crime are well reported to law enforcement officials, while other types of crime are under reported. For example, according to self-reported

victimization data, about half (51%) of all violent incidents committed against seniors were reported to police (Ogrodnik, 2006). The UCR does not collect victim information on non-violent crimes such as fraud or theft, therefore the victimization survey would be needed to produce statistics on victims of these types of crimes.

These surveys also give insights as to why crime is reported, or not. The surveys show that the need to make an insurance claim, seek medical assistance, and the seriousness of an offence tend to increase the level of reporting, while the inconvenience of reporting, the involvement of intimate partners and the nature of the offending tend to decrease reporting. This allows degrees of confidence to be assigned to various crime statistics. For example: Motor vehicle thefts are generally well reported because the victim may need to make the report for an insurance claim, while domestic violence, domestic child abuse and sexual offences are significantly under-reported because of the intimate relationships involved, embarrassment and other factors that make it difficult for the victim to make a report.

## 6. Concluding remarks

In the media and in society there is a certain fascination with crime data. The UCR survey has been active for a long time in several formats and will continue with a bright future. As new versions of the survey have been introduced, the importance of making the updates more compatible and consistent has grown. With development, the UCR has a greater ability to adapt to changes with minimal impact on the overall integrity and use of the data. Quality indicators and feedback have not only protected the integrity of the data, but also strengthened the communication and working relationship with the policing community. Monthly submissions ensure that emerging collection/ data quality issues are monitored and can be addressed quickly.

By developing new measures and products, Statistics Canada can provide the public with even more useful and practical statistics. The UCR is a very distinctive survey due to its interactive data collection process. While it does present certain challenges, it allows great opportunities for further quality improvement of both the data and the collection process. Feedback that is provided to the respondents helps to facilitate collection and correct errors in advance. The end result will be better data, which provides a better understanding of crime in Canada.

## Acknowledgements

The author would like to thank Colin Babyak of Statistics Canada for his valuable input and guidance, to Lenka Mach for the review and helpful comments, and to the staff that have worked with the justice surveys at Statistics Canada over the years.

## References

- Catalano, S.M. (2006). *The Measurement of Crime: Victim Reporting and Police Recording*. New York, LFB Scholarly Pub.
- Dauvergne, M. (2008). Crime Statistics in Canada, 2007, *Juristat*, 28(7). Statistics Canada: Ottawa, Canada.
- Gannon, M. (2006). Crime Statistics in Canada, 2005, *Juristat*, 26(4). Statistics Canada: Ottawa, Canada.
- Ogrodnik, L. (2006). *Seniors as Victims of Crime*, Statistics Canada: Ottawa, Canada.
- Ravidra, D. (2006). Use of Historical and Administrative Data in Edit and Imputation, *SIMP II – Economic Statistics Document Series*.
- United States Department of Justice (2008). About Incident-based Statistics and the National Incident-Based Recording System (NIBRS), <http://www.ojp.usdoj.gov/bjs/ibrs.htm>.