

# Data File Structure

The State Data System SAS® data files are organized such that information about each crash in the file is contained in one of three separate files:

- **Crash file:** Information describing general crash characteristics, e.g., the environmental and roadway conditions at the time of the crash.
- **Vehicle file:** Information describing the vehicles involved in the crash.
- **Person file:** Information describing the drivers, passengers, pedalcyclists, pedestrians and other non-motorists involved in the crash.

This common data file structure for the state crash data files makes them easier to use. Appendix A contains a list of the available state years in the State Data System and identifies years that also include VIN information. More information concerning how to use each data file, including methods of combining crash, vehicle, and person data, can be found in Appendix D. Researchers with knowledge of the SAS® software should be able to reproduce the results herein by applying the methodology in Appendix D.

Each state has its own police accident report format, incorporating different data elements. Consequently, the SAS® data files for each state in the State Data System do not contain the same variables. In an effort to standardize the SAS® data files, NHTSA has adopted common variable names where appropriate (e.g., crash severity, vehicle damage severity, and injury severity). Also, note that those variables provided by the state that are not relevant to NHTSA analyses (e.g., police officer's badge number and the individual participants' names and addresses) have been eliminated.

Although common variable names are used in the SAS® data files, the variable definitions present in the files differ among states. This is further complicated by the fact that different police jurisdictions within a state often interpret common data elements in different ways. This lack of uniformity should be carefully noted, and researchers are advised to be wary when drawing conclusions from data obtained from different states. When analyses involve two or more states in the State Data System, results need to be examined to ensure that any differences in the data being collected and coded by each state are taken into account.

The differences from state to state in data collection and reporting limit the utility of the state crash data files for most nationwide analyses. For example, the KABCO injury severity scale is used by most (but not all) states, yet the percentage of injuries in each of the four severity categories of the scale varies widely across states, indicating different interpretations or different crash reporting thresholds. Similar problems occur with the reporting of crash severity, where different and generally subjective scales are used in different states.

## DATA FILE STRUCTURE

---

Additionally, states differ in their criteria for reporting traffic crashes. Since time pressures and limited resources prevent police from collecting all of the information desired by the many users of the data, reporting thresholds have been implemented to focus the data collection on information perceived as most important. Some states have a minimum dollar value for reporting property damage only crashes, while other states use a towaway criterion (see Appendix B).

In general, every state will report all of their fatal and injury crashes. However, uninjured vehicle occupants may or may not be reported (see Appendix C). One additional factor complicating data analysis is that states may change their reporting criteria, their policy on coding uninjured occupants, the interpretation of their data variable codes or, indeed, even their entire data structure. Any changes that have come to the NCSA's attention are duly noted in this report.

Data for some crashes in the state crash data files are incomplete. For this reason, a set of special codes is used to indicate why a given element may not be available for a particular crash, or for all crashes from a particular state. These codes are:

- **Unknown:** Information was coded as "Unknown" in the original police accident report.
- **Missing:** Information is normally available from the state, but was either coded as "Not Stated", "Not Coded", "Uncoded" or other similar designation. Missing information may also be indicated by the simple omission of any code.

Note that for a given state and variable, unknown and missing entries from the PARs may be grouped under a single code.

In the situations described in the preceding paragraphs and also for situations in which a data element is miscoded (e.g., valid data elements are [0, 1, 2] and a "3" is entered), the data elements are computationally treated as "unknowns" and are not included in the valid reporting categories.

For example, given the hypothetical frequency distribution of raw data for the variable "SEX":

0 = Not Stated	0.2%
1 = Male	47.3%
2 = Female	47.7%
3 = Unknown	2.2%
4,5 = Miscoded	0.1%
Not Coded	2.5% (entries are blank for these observations)

This would be computationally treated as:

1 = Male	47.3%
2 = Female	47.7%
3 = Unknown	5.0%

The state crash data files contain large amounts of information that are in a relatively raw form. Edit procedures are performed by states at the point of data entry in an effort to correct coding problems. In certain circumstances, minor additional editing of the data files are performed for consistency purposes. This editing typically involves recoding inconsistent dates, times, and ages to unknown. In addition, quality control mechanisms are applied to the files received from the states by NHTSA. NHTSA has implemented these quality control efforts to ensure the SAS® conversion programs work correctly. If discrepancies in the data are detected, they are referred to the state agency providing the data file, but are not corrected in the State Data System data files. However, in tabulating the descriptive statistics contained in this report, a small number of crashes were removed from consideration. These crashes fall into two categories:

- **Crashes with missing elements.** These are crashes without associated vehicle information, or vehicles without associated crash information.
- **Crashes with ambiguous information.** Some crashes may have multiple vehicles with the same vehicle number (making it difficult to place vehicle occupants in their correct vehicles). In rare cases, multiple crashes may have the same case number (making it difficult to associate vehicles with their proper crashes).

## File Characteristics

### GENERAL NOTES

1. Unless otherwise noted, the data structure for a given state is uniform from 1990-1999. If the data structure has changed, a given variable's codes may or may not have changed. Consult the specific queries in Appendix D for more information.
2. Specific alcohol variables may only be coded for certain person types. If a given person type is associated with an alcohol indicator, it should be assumed that at least one of the alcohol variables in the Person file is coded for that person type. Again, consult the specific queries in Appendix D for more information.
3. Unless otherwise noted, parked vehicles and hit and run vehicles are included in the Vehicle file. Phantom vehicles, vehicles that cause but are not involved in the crash, are not included in the Vehicle file.

### STATE-SPECIFIC NOTES

#### California

- The Number of Vehicles (num\_veh) variable includes pedestrians and pedalcyclists.
- Alcohol indicators are coded for drivers, pedestrians, and pedalcyclists.

# DATA FILE STRUCTURE

---

## Florida

- Two data structures are present, 1990-1992 and 1993-1999.
- Phantom vehicles are included in the Vehicle file.
- The Number of Vehicles (num\_veh) variable includes pedalcyclists.
- Alcohol indicators are coded for drivers, pedestrians, and pedalcyclists.
- Florida does not code gender information for passengers.

## Georgia

- Due to problems arising from re-engineering its database, Georgia was unable to provide crash data for 1999.
- Three data structures are present, 1990-1993, 1994-1997, and 1998.
- The total number of crashes for 1998 is approximately 7% less than previous years due to the deletion of inconsistent cases.
- Four months of 1998 data are considered unreliable due to a change of administrative system and are currently being re-tabulated.
- The Number of Vehicles (num\_veh) variable includes pedalcyclists.
- Alcohol indicators are coded for drivers, pedestrians, and pedalcyclists.

## Illinois

- Illinois has undergone several significant changes over the reporting period. While many of the pertinent data variables and codes are unchanged over the ten-year period, their usage has changed. The data structure is identical for 1990-1995, but the interpretations used by police officers changed with the introduction of a new PAR in mid-1993. Another PAR was introduced in 1996. The data structure for 1996-1999 retains some of the variables from the 1990-1995 period, modifies or adds codes for other variables, and also introduces new variables not used before 1996. Researchers should consult the individual queries used in Appendix D to determine if a variable in question has changed. Further complicating matters is the fact that the 1996-1999 raw data files received from Illinois are incomplete when compared to the 1990-1995 raw data files. The 1996 file is approximately half the normal size, consisting only of state routes. The 1997 raw data file is missing approximately one-third of the non-fatal crashes in Chicago. The 1998 and 1999 raw data files are missing all non-fatal crashes in Chicago. Due to the above circumstances, the periods 1990-1992, 1994-1995, and 1997-1999 should be viewed as three distinct intervals, and the years 1993 and 1996 are transition years.

- Pedestrian and pedalcyclist injuries for 1996 are considerably lower than expected (as a proportion of total injuries), possibly due to the exclusion of crashes not occurring on state routes. Pedestrian and pedalcyclist fatality totals for 1996 are also somewhat lower than expected.
- The Vehicle Type (veh\_type) variable categories of 'Not Stated/Other/Unknown' comprise approximately 4-5% of the observations for 1990-1993, 37-39% for 1994-1995, and 2-6% for 1996-1999. For this reason, fairly sharp decreases may be observed for vehicle-specific tables when comparing 1993 and 1994 data.
- The Restraint Device (rest1) variable 'Unknown/Not Stated/Missing' categories comprise approximately 93% of the observations in 1990, 79% in 1991, and 28-37% in the period 1992-1995. Hence, the safety equipment data for 1990-1995 demonstrates distinct variability for this period. As noted above, the 1996-1999 data is incomplete, and comparisons with the 1990-1995 period should be made with care.
- Neither pedestrians nor pedalcyclists are included in the Number of Vehicles (num\_veh) variable.
- Alcohol indicators are coded for drivers.
- Speeding-related crash totals dramatically increase from 1996 through 1999 due to the introduction of a new Person-level variable, Driver Contributing Circumstance (con\_cir1), with a speeding indicator of 'Too fast for conditions'.

### Indiana

- Neither pedestrians nor pedalcyclists are included in the Number of Vehicles (num\_veh) variable.
- Alcohol indicators are coded for drivers, pedestrians, and pedalcyclists. Passengers may also be coded for alcohol indicators at the officer's discretion.

### Kansas

- Neither pedestrians nor pedalcyclists are included in the Number of Vehicles (num\_veh) variable.
- Alcohol indicators are coded for drivers, pedestrians, and pedalcyclists.
- Rollover totals decrease by approximately one-third beginning in 1997. Kansas believes this is due to a change in the coding of the First Harmful Event (event1) variable.

### Maryland

- Two data structures are present, 1990-1992 and 1993-1999. Due to the transition in 1993, Maryland believes that the data for 1993 and 1994 are unreliable, especially the vehicle data.
- Neither pedestrians nor pedalcyclists are included in the Number of Vehicles (num\_veh) variable.
- Alcohol indicators are coded for drivers, pedestrians, and pedalcyclists.

## Michigan

- There are three distinct sets of data for Michigan representing the periods 1990-1991, 1992, and 1993-1999. The data structure changed dramatically in 1992. Additional coding changes were made in 1993. In general, the 1992 Michigan crash statistics are believed to be artificially low by an estimated 6-10% due to a number of the PARs not being processed. Physical damage to some PARs and technical problems with data recorded on other PARs are reported as the primary causes.<sup>1</sup>
- Michigan totals exclude non-traffic crashes. These crashes can be identified using the Accident file Crash Indicator (crash) variable. Case numbers with crash='1' were excluded from the Accident, Vehicle, and Person files prior to generating the Michigan totals reported herein.
- Parked vehicles are not included in the 1990-1991 Vehicle files; they are included in the 1992-1999 Vehicle files. Hit and run vehicles are included, and phantom vehicles are excluded, in the 1990-1999 Vehicle files.
- Neither pedestrians nor pedalcyclists are included in the Number of Vehicles (num\_veh) variable for 1990-1991. Pedestrians and pedalcyclists are included in the Number of Vehicles variable for 1992-1999. There is a drop in the percentage of single vehicle crashes from approximately 40% (1990-91) to approximately 30% (1992-99). This drop is partly attributable to the inclusion of parked vehicles, partly due to the inclusion of pedestrians/pedalcyclists as traffic units, and partly due to data entry errors (1992-99).
- Alcohol indicators are coded for drivers, pedestrians, and pedalcyclists.

## Missouri

- Two data structures are present, 1990-1992 and 1993-1999. The following variables used in this report also changed in 1995: Contributing Circumstance 1-5 (confac1-5), Pedestrian Contributing Circumstance 1-4 (cont\_cir1-4). These variables were used in alcohol- and speeding-related crash determinations.
- The Number of Vehicles (num\_veh) variable includes pedalcyclists.
- Alcohol indicators are coded for drivers, pedestrians, and pedalcyclists.
- The classification of light trucks and large trucks for 1990-1992 is approximate due to Missouri's use of a generic truck descriptor "Single Truck" in this time period. The Truck Licensed Weight (gvwr) variable was used to make an approximate distinction. Trucks with a gvwr of 12,000 lbs. or less (Missouri's low weight classification is <12,000 lbs. as opposed to NHTSA's classification of <10,000 lbs.), and those single trucks with an unknown gvwr were designated light trucks. This method gives general agreement with the frequencies of light and large trucks for 1993-1999.

<sup>1</sup>Source: 1993 Michigan Traffic Crash Facts, Michigan Department of State Police.

### New Mexico

- Parked vehicles are included in the Vehicle files. Hit and run vehicles and phantom vehicles are not.
- New Mexico does not have a Number of Vehicles variable. However, the corresponding derived variable used in this report does not include pedestrians or pedalcyclists.
- Alcohol indicators are coded for drivers, pedestrians, and pedalcyclists. Passengers who are fatally injured also are coded for alcohol indicators. Passengers not fatally injured may also be coded for alcohol indicators at the officer's discretion.
- New Mexico changed their police accident report in 1993. However, there were no variables used in this report that were affected.

### North Carolina

- Data from North Carolina were not incorporated into the State Data System until 1992, and as a result this report does not contain crash data from North Carolina for 1990-1991.
- The Number of Units (numunit) variable includes pedestrians and pedalcyclists.
- Alcohol indicators are coded for drivers, pedestrians, and pedalcyclists.

### Ohio

- The Number of Vehicles (num\_veh) variable includes pedestrians and pedalcyclists.
- Alcohol indicators are coded for drivers, pedestrians, and pedalcyclists.
- Private property crashes, included in the 1990-1992 data files and identified by an Accident Severity (severity) variable value of '4', have been excluded from the report analysis.

### Pennsylvania

- Illegally parked vehicles and hit and run vehicles are included in the Vehicle files. Legally parked vehicles and phantom vehicles are not.
- The Number of Vehicles (num\_veh) variable includes pedalcyclists.
- Alcohol indicators are coded for drivers, pedestrians, and pedalcyclists.
- The contributing factor codes used in speeding crash determinations changed in 1993.

### Texas

- Parked vehicles, hit and run vehicles, and phantom vehicles are excluded from the Vehicle files.
- Neither pedestrians nor pedalcyclists are included in the Number of Vehicles (num\_veh) variable. Parked vehicles, hit and run vehicles, and phantom vehicles, while not included in the Vehicle file, are included in the Number of Vehicles total. Also, vehicle types described as “machinery” are not included in the Number of Vehicles total. For these reasons, approximately 2-3% of the crashes reported by Texas will have a Number of Vehicles total greater than the number of vehicle records in the Vehicle file; a negligible number of crashes will have Number of Vehicles totals less than the number of vehicles in the Vehicle file.
- Alcohol indicators are coded for drivers, pedestrians, and pedalcyclists.
- Texas raised its minimum assessed property damage reporting standard for property damage only (PDO) crashes in 1995. PDO crash totals markedly decrease in 1995 and 1996.

### Utah

- The Number of Vehicles (num\_veh) variable does not include pedestrians or pedalcyclists.
- Alcohol indicators are coded for drivers, pedestrians, and pedalcyclists.

### Virginia

- The Number of Vehicles (num\_veh) variable includes pedalcyclists.
- Alcohol indicators are coded for drivers, pedestrians, and pedalcyclists. Fatal passengers also are coded for alcohol indicators.
- Virginia does not code motorcyclist helmet information.

### Washington

- Due to problems arising from re-engineering its database, Washington was unable to provide crash data for 1997-1999.
- Washington does not have a Number of Vehicles variable. However, the corresponding derived variable used in this report does not include pedestrians or pedalcyclists.
- Alcohol indicators are coded for drivers, pedestrians, and pedalcyclists.