4. ANALYSIS OF DATA SETS

4.1 INTRODUCTION

We concentrated on finding data sources with information on population, income, driver/driving distance, physical/functional limitations, and crashes/fatalities. Because of the need to project results based on age groups, gender, and region, for each five-year time period, we needed specific data elements which could be matched among the various data sets.

Although population projections were available from the U.S. Census Bureau, no other projections were available at the level of detail needed for the other categories. Therefore, we searched for historical data sets in the above categories to establish trends and relationships on which projections could be based. The perfect data source would contain records which could be identified by age group, gender, and region; in addition, it would span a study that had continued for at least 25 years (since that was the time span of the projections).

No single historical data set contained all needed information. For example, the Nationwide Personal Transportation Surveys (NPTSs) contained demographics, driving status, and the availability of other drivers in the household but did not contain a functional/health indicator; the National Health Interview Surveys (NHISs) contained demographics and a functional/health indicator but did not contain data on driving status or the existence of another driver in the household. Neither of these data sets contained crash data.

After a review of the available data sets, we concluded that finding sufficient historical data on which to base projections is difficult. Especially difficult is finding data to describe cohort behavior and conditions for the years on which the model is based (i.e., 1977, 1983, 1990, 1995).

4.2 ASSESSMENT OF DATA SOURCES

We collected and evaluated data from many sources before determining the most useful and reasonable data sets for use in the model. Data sets were examined for their availability, completeness, quality, and appropriateness for the purposes of the project. Each data set is described briefly below under one or more of the categories of information needed for the projection model. Reasons for inclusion or exclusion of data sets are provided.

4.2.1 Population Data and Projections (U.S. Census Bureau)

Population data from the U.S. Census Bureau (http://www.census.gov), with historical data dating back to 1790 and projections through 2100, were the easiest to obtain and most complete data sets available. Appropriate as a measure of the rapid increase in the numbers of elderly after 2010, population projections also included considerations of migration, immigration, emigration, birth/death rates, and other population-changing influences. Population data could be obtained for every permutation needed by the model. Additional information on how population data were used in the model is given in Section 5.3 and Appendix B.1.

4.2.2 Income Data (Bureau of Economic Analysis, U.S. Census Bureau, and National Health Interview Survey)

The Bureau of Economic Analysis (BEA) provides historical personal income data from 1969, and makes projections on personal income by states through 2045. The projections are for all age groups. Because of confidentiality and disaggregation issues, however, we were unable to use the BEA income projections.

Census data sources contain historical long-term income data by various “types” of households (i.e., female head of household with no husband present, etc.), and by age groups. There are also historical data sets of median income by states for a “typical” family. Unfortunately, income data from the U.S. Census Bureau contain limited disaggregation capabilities.

The NHIS is a large (about 40,000 households), well-designed, cross-sectional study, conducted annually, of non-institutionalized individuals. We considered using NHIS to derive income projections using either

- Direct projection of elderly cohort-specific income, or
- Separate projections of (a) non-elderly cohort-specific working income and (b) same cohort-specific decline in income during elderly years.

Income data from the NHIS, however, had limited usage within the model because NHIS is not a longitudinal study, and income trends were difficult to project. Because we felt that projections of income were a critical component in the model, ORNL issued a subcontract to a reputable firm (Standard & Poor’s DRI) having experience in using population and income files from the U.S. Census Bureau to obtain income projections in the format and disaggregations desired. The DRI estimates and forecasts are explained in further detail in Appendix B.2.1.1.

4.2.3 Drivers and Driving Distance Information (*Highway Statistics* and Nationwide Personal Transportation Survey)

The U.S. DOT's FHWA produces annual reports containing highway statistics for the entire United States. These reports provide a dependable, reliable, consistent, long-term (many statistics going back to 1949) data source. Relevant information includes numbers of driver's licenses by age, gender, and state; total VMT; miles of highways of various types; and other potentially useful information (e.g., the transportation budget). There are limitations in this data source, however. For example, *Highway Statistics* does not contain any projections and does not supply annual VMT per person. In addition, if data from the NPTS are used in the model, it is important to obtain data for 1977 and 1983. *Highway Statistics* Table DL-22, which provides the number of driver's licenses by age-gender-state, does not exist for odd-numbered years prior to 1990, and it also combines all ages above age 70. Although these constraints could be overcome through interpolation and disaggregation, we determined fairly early in the project that driver totals from the NPTS would be used in the model rather than numbers of drivers licenses as reported in *Highway Statistics.* We concluded that the driver totals from NPTS more closely approximated the numbers of people who actually drove, as opposed to people who just held a license, particularly among the elderly.

The NPTS, a large (42,000 households in 1995) transportation survey taken in 1977, 1983, 1990, and 1995, contains historical data on personal travel habits (driver/non-driver status, VMT), availability of transit (except for 1983), number of drivers in a household, age, income, and other relevant information. NPTS does not contain projections, is not available for every year, is a survey rather than a census, and consists of self-reported estimates. In addition, the 1983 NPTS does not contain a regional identifier. Because, however, the NPTS has such a wealth of individual travel information, we decided to base projections of driving status and VMT on the NPTS data from the four surveys noted above. Additional information on how the data from NPTS were used in the model is provided in Chapter 7.

Many factors influence whether, and how much, an individual drives. Although studies have examined the impacts of various factors (see Chapter 3), there are few data sets available. We pursued the use of gasoline prices as an influential factor on VMT. This data set, from the Energy Information Administration, provided average gasoline prices, including taxes, by Census region. The first year for which this data set became available was 1984. Prices are also forecasted through 2020 (personal communication with Bruce Bawks, Bruce.Bawks@eia.doe.gov). The price of fuel was incorporated in the model (Section 7.2).

Another potential influence on the decision to drive is the availability of alternative transportation. This availability was based on a rural-urban designation from the U.S. Census Bureau. For more information, see Appendix B.2.7 and Table B.11.

Although studies have addressed self-regulation and the decision to stop driving (Chapter 3), there are no data sources that would help predict how this phenomenon will affect elderly decisions in the future. Therefore, this factor was not incorporated in the model.

4.2.4 Physical/Functional Limitations

There are speculations and projections about the health conditions of the elderly in the future; however, there is no data source that provides actual data-based health-related projections. Therefore, we attempted to locate historical data sets that provide data on health-related limitations to driving, driving cessation based on self-regulation, risks to elderly persons involved in crashes, and other factors that relate physical/functional limitations and driving.

The historical data sets examined included the Health and Retirement Study/Assets and Health Dynamics of the Oldest Old (HRS/AHEAD), NHIS, and the Social Security Administration (SSA)/U.S. Census Bureau.

The HRS and the AHEAD studies are conducted and maintained by the Institute for Social Research, University of Michigan. HRS is intended to provide data for researchers, policy analysts, and program planners who have to make policy decisions affecting retirement, health insurance, saving and economic well being. AHEAD is intended to provide detailed coverage of the joint dynamics among health (physical, cognitive, and functional), dementia, economic and family resources and care arrangements for the oldest old.

The HRS/AHEAD surveys have been conducted every two years since 1992 (AHEAD began in 1994). HRS and AHEAD data collection efforts merged in 1998. Sampling includes 9,473 households, 11,965 individuals, and 7,447 respondents. This combined data source is a well-documented, longitudinal survey of elderly individuals. (AHEAD studied adults born in 1923 or before; HRS studied adults born in 1931-1941; these dates were adjusted somewhat in 1998 for the joint survey.) Although this would appear to be an excellent source of data on the physical/functional limitations of the elderly, the very limited time ranges (since 1992) of these studies make it difficult to discern time trends from them. In addition, because the surveys did not exist in 1977, 1983, or 1990, the input would be difficult to use with NPTS data from those years. Finally, there were age group gaps in the data. For example, for 1994, the AHEAD data are restricted to adults 72 and older, and the HRS data are restricted to adults between 53 and 63; therefore, adults in the age group of 64-71 are omitted.

In addition to the HRS/AHEAD longitudinal studies, other longitudinal studies, including the Established Populations for Epidemiologic Studies of the Elderly (EPESE), Marin County Study, and the Sonoman studies, were considered. Because of a lack of data or incompatibility with other data sources, these studies were not used in our modeling.

NHIS is described in Section 4.2.2 above. Although not all data elements desired for matching NHIS data with NPTS data were available, we were able to use the Activity Limitation Status (ALS) variable from NHIS to compute an index which was used as a surrogate health indicator in the model. For more information on this calculation, see Section 5.4.

The SSA-Census data provides past and projected life expectancies. At one point in time, we considered using life expectancy as a factor; however, the ALS data appeared to be sufficient.

4.2.5 Crashes, Casualties, and Fatalities (National Center for Statistics and Analysis, National Highway Traffic Safety Administration, and the American Insurance Services Group)

NHTSA's National Center for Statistics and Analysis (NCSA) maintains data for the Fatality Analysis Reporting System (FARS). FARS data are available for 1975-1998 from ftp://www.nhtsa.dot.gov/FARS . This data set is complete for **all** fatal crashes in all fifty states, the District of Columbia, and Puerto Rico. Data are collected from police accident reports (PARs), death certificates, hospital records, etc. Data sets include (1) crashes, (2) vehicles/drivers, (3) persons. Because these data represent a census, FARS is especially useful for regional and national vehicle accident mortality tables. We used this data source for developing the crash risk portion of the model. Additional information on data usage is found in Chapter 8.

The NCSA also maintains the National Automotive Sampling System, General Estimation System (NASS GES). Data are collected from a nationally representative sample of police-reported motor vehicle crashes involving property damage, injury, or death. The system began collecting data in 1988. There is a bias due to restriction to police-reported crashes, but the data set **is** national and is collected by a three-stage weighted probability sample (Stage 1: sample geographic areas; Stage 2: sample police jurisdictions within geographic areas; and Stage 3: sample PARs within jurisdictions). A sample consisting of 55,000 PARs were collected in 1997. Sampling weights are calculated to adjust for variable sampling probability. In addition to estimates, valid standard errors can be calculated from these data and weights. Because of the national scope and scientific sampling, the NASS GES data seemed especially appropriate for making regional and national accident projections. The major limitations with GES data were that data were not available on a state/regional level and did not contain injury information prior to 1988. Because of these limitations of the GES data, this data source was not used in development of our model.

The Crash Outcome Data Evaluation System (CODES) is also maintained by NHTSA's NCSA. CODES uses *probabilistic data linkage* to merge data from eclectic sources such as motor vehicle crash reports, emergency medical services, hospitals, insurance claims, and death certificates. In probabilistic data linkage, data records are merged on the basis of estimates of the probability that they are from the same target (e.g., accident or driver), rather than conventional merging by ID variables that are definite links (and which do not necessarily exist). The probability estimates are computed from data variables such as time, location, and type of vehicle. Although probabilistic data linkage is automated by computers, some case-by-case tuning of the results is still necessary, and the process is laborious. CODES data are from seven states. The primary objective of CODES was to investigate effectiveness of seat belts and motorcycle helmets (and to prepare a report to Congress). Because CODES was limited in scope to only seven states, this data source was not used in development of our model.

The State Crash Data system, available from NCSA, includes a census of all crashes from 17 states. Data, which cover the time period of 1989-1994, are not uniform from state to state in either reporting criteria or coding schemes. Tabular reports are available on (1) trends, (2) crashes, (3) vehicles, and (4) people. Because the State Crash Data was limited in scope to 17 states, this data source was not used in the development of the model.

Index and the Property Insurance Loss Register (PILR) are maintained by the American Insurance Services Group (AISG). Index is a national clearinghouse for bodily-injury and automobile insurance claims data with a database of over 50 million claims. PILR collects first-party property claims data. PILR was merged with Index; together they contain over one billion records and receive 18 million claims each year. Index was used by CODES. Index and PILR are used primarily as tools for preventing insurance fraud. These data sources were not used in our model because of institutional barriers to obtaining the data.

Seat belt use has been shown to reduce the impact and casualty rate of a crash. The projected use of seat belts among the elderly was obtained from NHTSA and incorporated in the model (Section B.2.5 and Table B.10).

4.2.6 Other Data Categories Considered

In addition to the obvious impact of population, income, driver status and driving distance, health, and crash risk on projecting numbers of casualties of the elderly for the next 25 years, there were other potential factors that we examined. These included education level, transportation/highway budgets, and the impact of safer infrastructure/vehicles/drivers in the future [including possible impacts of the Intelligent Transportation System (ITS)]. Although these factors were not incorporated within our model, further research is needed to determine how they may impact future casualties (see also Chapters 3 and 10).

4.3 CONCLUSIONS

Table 4.1 lists the data sources examined and provides a cross-sectional reference to the data categories.

Table 4.1. Data Sources Examined

Data source	Type of data available	Notes
U.S. Census Bureau (http://www.census.gov)	• Population (historical data) • Population (projections) • Income • Education • Rural-urban percentages	We used population projections and rural-urban percentages in the model
BEA (http://www.bea.doc.gov/beahome.html)	• Income (historical) • Income (projections)	
NHIS	• Health • Income • Education • Employment status	We derived a health index from the ALS data
Standard & Poor's DRI	• Income (projections)	We used income projections
U.S. DOT, FHWA (http://www.fhwa.dot.gov/ohim/ohimstat.htm)	• Drivers licenses • Highway mileage • VMT • Transportation budgets	
NPTS (conducted by U.S. DOT) (see, for example, http://www-cta.ornl.gov/npts/1995/Doc/index.shtml)	• Driving status/availability of other driver in household • VMT/person • Income • Availability of transit	We used driving status, availability of other driver in household, VMT/person
EIA (Bruce.Bawks@eia.doe.gov)	• Fuel prices	We used fuel prices
HRS/AHEAD (http://www.umich.edu/~hrswww/)	• Physical/functional limitations	
SSA (http://www.ssa.gov/)	• Physical/functional limitations	
U.S. DOT, NHTSA's NCSA (http://www.nhtsa.dot.gov/people/ncsa/)	• Fatalities • Crashes	We used fatalities data
U.S. DOT, NHTSA (http://www.nhtsa.dot.gov/people/injury/airbags/presbelt/)	• Seat belt use	We used seat belt data
AISG (http://www.iso.com/AISG/indexsys/indexsys.html)	• Crashes	

Based on the requirements of the model and the appropriateness of the data set, the following public-use data sources were chosen for use:

- Population projections and rural-urban projections from the U.S. Census Bureau,
- Historical health data from the NHIS,
- Driver information (driving status, availability of other driver, and VMT/person) from NPTS,
- Fuel prices from EIA,
- Historical fatality data from FARS, and
- Income projections produced by DRI.